Module 4: Functions on $\mathbb{R}^n$

This module concerns real-valued functions of one of several variable, i.e. functions $f : \mathbb{R}^n \to \mathbb{R}$. We begin with the univariate case, and then turn to multivariate owrk. As emphasized in a previous chapter, economics is inherently a multivariate affair; however, much of the analysis is conducted either by reducing a multivariate problem into a decoupled collection of univariate problems, or by extending univariate results to multivariate settings. The point is this: most of the main ideas and techniques are already present in the univariate world – a world which is attractive for many reasons, including its admission of graphical approaches.proposition

1 Continuity

A function $f : \mathbb{R} \to \mathbb{R}$ is *continuous at* $x \in \mathbb{R}$ provided that whenever $x_n \to x$ it follows that $f(x_n) \to f(x)$. The domain $\mathbb{R}$ in this definition could be replaced by any subset of $\mathbb{R}$, though usually the domain is restricted to be open or closed. Observe that this definition makes perfect sense also for functions between two metric spaces, and indeed this is what it means for a function between two metric spaces to be continuous at a point. It's a good exercise to show that this definition comports with the one you perhaps remember:

Exercise 1 *Show that $f : \mathbb{R} \to \mathbb{R}$ is continuous at $x \in \mathbb{R}$ provided that for any $\varepsilon > 0$ there is a $\delta > 0$ such that*

$$|y - x| < \delta \implies |f(y) - f(x)| < \varepsilon.$$

Importantly, continuity is a *local* notion: the definition speaks of continuity at a *point*. We say that the function is *continuous* on its domain (in the above case the domain is $\mathbb{R}$) provided that it is continuous at every point in the domain.

Continuity preserves some notion of nearness. Intuitively, if y is near x then $f(y)$ is near $f(x)$. However, notice, from exercise 1, that a more precise intuition is as follows: you can make $f(y)$ near $f(x)$ by choosing y sufficiently near x.[1] Continuity is useful in economic analysis for many reasons – principal among them the intermediate value theorem guaranteeing solutions to equations – but, at perhaps the broadest level, the benefit is this: If f is continuous at x then the behavior of f near x can't be too crazy.

The following propositionosition shows that continuity is preserved by many common point-wise transformations.

[1] How near you need to choose y may depend on the point x; the stronger notion of *uniform continuity* eliminates this dependence.

Proposition 1 *Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are continuous. Then the following functions are continuous:*

1. $f + g$
2. $f \cdot g$
3. $f \circ g$

Exercise 2 *Are polynomials continuous? Answer using the following steps.*

1. *Let $\alpha \in \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \alpha$. Show that f is continuous.*
2. *Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x$. Show that f is continuous.*
3. *Use parts 1 and 2 of this problem, together with the above proposition to show that polynomials are continuous.*

The most important result on continuous functions from $\mathbb{R}$ to $\mathbb{R}$ is the *intermediate value theorem*.

Theorem 1 (Intermediate value theorem) *Let $f : [a,b] \to \mathbb{R}$ be continuous. If $f(a) > 0 > f(b)$ or if $f(a) < 0 < f(b)$ then there exists $c \in (a,b)$ such that $f(c) = 0$.*

The proof of this theorem was sketched in a previous chapter. The importance of the theorem concerns solutions to equations. It is an existence theorem. It provides precise conditions under which an equation is guaranteed to have a solution.

Existence theorems are wonderfully powerful as they often require only weak assumptions; however there is a trade off: most important existence theorems do not guarantee uniqueness – there may be many zeros of f in (a,b) – and, more significantly, they are not typically *constructive*: they provide no method for constructing the object whose existence is establishedby the result.

Exercise 3 *Consider a competitive market for one good. Let $q = S(p)$ and $q = D(p)$ be supply and demand curves respectively. Assume S and D are continuous functions. Under what conditions can you guarantee the existence of a market equilibrium?*

2 Differentiability

A function $f : \mathbb{R} \to \mathbb{R}$ is *differentiable at x*, provided the following limit, denoted $f'(x)$, exists:

$$f'(x) = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}.$$

The domain $\mathbb{R}$ may be replaced by an open set. Observe that, like continuity, differentiability is a local notion: it is defined at a given point in the domain. A function is *differentiable* if it is differentiable at every point in the domain.

Whereas we may think of continuous functions as reasonably well-behaved – things don't get too crazy near a point of continuity – differentiable functions are *very* well-behaved; so well-behaved, in fact, that, locally, they act just like linear functions. This is the reason differentiability is so important and calculus is so useful.

To make this point more clear, notice that the definition of the derivative above means that for small values of Δx, the following approximation is pretty good:

$$f(x+\Delta x) - f(x) \approx f'(x)\Delta x. \tag{1}$$

Thus, letting $z = x+\Delta x$, we conclude that if z is near x (i.e. Δx is small) then

$$f(z) \approx f(x) + f'(x)(z-x),$$

i.e. the function f looks like the line $y = mz+b$ where $m = f'(x)$ and $b = f(x) - f'(x)x$. In particular, $f'(x)$ is the slope of the line tangent to the graph of f at x.

Another important way to intuit the local linearity of differentiable functions is to rewrite (1) as $\Delta f \approx f'(x)\Delta x$. Now recall that a linear function from $\mathbb{R}$ to $\mathbb{R}$ is exactly and only multiplication by a real number. Thus we may interpret the real number $f'(x)$ as defining a linear function $\mathbb{R} \to \mathbb{R}$. Which linear function? The linear function taking small changes in x to approximations of the corresponding small changes in f. Viewed this way – as a linear map – the derivative is often referred to as the *differential*, and written $df = f'(x)dx$, where here the dx and df are called *infinitesimals*, and loosely thought of as infinitely small: they should not be viewed as having meaning in isolation, but only when written in relation to other differentials. We think of df as the small change in f induced by the small change dx in x, and this small change in f is measured by their derivatives.

Exercise 4 *Show that a differentiable function is continuous.*

Like continuity, differentiability is preserved under many common point-wise transformations.

Proposition 2 *Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable. Then the following functions are differentiable:*

1. $f+g$
2. $f \cdot g$
3. $f \circ g$

While the definition of differentiability is somewhat daunting, it turns out that calculating derivatives is often mechanical, hence the name: the calculus.

Proposition 3 *Suppose $f, g : \mathbb{R} \to \mathbb{R}$ are differentiable. Then*

1. $h = f + g \implies h' = f' + g'$
2. $h = f \cdot g \implies h' = f' \cdot g + g' \cdot f$
3. $h = f \circ g \implies h' = (f' \circ g) \cdot g'$

Exercise 5 *How to differentiate polynomial? Answer using the following steps.*

1. *Let $\alpha \in \mathbb{R}$ and define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = \alpha$. Show that f is differentiable and $f'(x) = 0$.*
2. *Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = x$. Show that f is differentiable and $f'(x) = 1$.*
3. *Use parts 1 and 2 of this problem, together with the above propositions to show that polynomials are differentiable, and to compute the derivative.*

Exercise 6 *Let $f : (0, \infty) \to \mathbb{R}$ be defined by $f(x) = 1/x$. Show that $f'(x) = -1/x^2$.*

Exercise 7 *Show that if $f : \mathbb{R} \to \mathbb{R}$ is differentiable and $f' > 0$ everywhere then f is invertible.*

The most important result on differentiable functions is the *mean value theorem*.

Theorem 2 (Mean value theorem) *Let U be an open set containing the interval $[a, b]$, and let $f : U \to \mathbb{R}$ be differentiable. Then there exists $c \in (a, b)$ such that*

$$f'(c) = \frac{f(b) - f(a)}{b - a}.$$

Graphically, this theorem is straightforward: the RHS is the slope of the line connecting the endpoints of the graph of f, i.e. the secant line. The theorem says this slope is equal to the slope of the tangent line at some point inside the interval (a, b). A more important (and more generalizable) intuition is as follows: think of f as the position of something at a given time $t \in [a, b]$, so that f' is the instantaneous velocity. Then the RHS is the average velocity over the time period a to b. The theorem says that at some point in the interval of time (a, b) the instantaneous velocity must equal the average velocity.

The principal importance for economics of the mean value theorem is that it is a special case of Taylor's theorem, so I will put off discussion of applications until the next chapter.

Exercise 8 *Suppose $f, g : [a,b] \to \mathbb{R}$. Show that if $f' = g'$ then there is a constant α such that $f(x) = g(x) + \alpha$.*

A nod to higher-order derivatives is worth making here. Note that if $f : \mathbb{R} \to \mathbb{R}$ is differentiable then f' may be viewed as a function $\mathbb{R} \to \mathbb{R}$, and it may be differentiable. If it is, then f'' is called the *second-derivative* of f. This process may be repeated so long as the derivative exists. The notation $f^{(k)}$ is used to denote the kth derivative of f. In general we say that f is C^k (pronounced "see kay") if its derivative may be computed k-times *and* as a function $f^{(k)}$ is continuous.

Exercise 9 *What is the geometric interpretation of f''?*

3 Integration

Whereas differentiation allows for the local analysis of a given function, integration speaks to a measure of its global behavior. Intuitively, given a function $F : [a,b] \to \mathbb{R}$, we may use differential calculus to think about small changes in f given small changes in $x \in [a,b]$, i.e. $dF = f(x)dx$, where here, for rhetorical reasons, I am using f to suggest the derivative of F. Conversely, suppose that, instead of knowing the function F, we know how it changes at each point in an interval. What can we say about the total change $F(b) - F(a)$? It is perhaps not unreasonable to expect that we may compute this total change by adding up all the little changes:

$$F(b) - F(a) = \sum dF = \sum f(x)dx. \tag{2}$$

Riemann integration provides the machinery needed to make these intuitive, if meaningless statements meaningful.

We need some terminology. A *partition* of an interval $[a,b]$ is

$$\mathscr{P}_n = \{a = x_0, x_1, \ldots, x_n = b\}$$

with $x_i < x_{i+1}$. Note that a partition as defined here partitions the interval $[a,b]$ into a collection of intervals with endpoints x_i, x_{i+1} (these intervals may be open or closed depending on how you choose to construct the partition). Given $\mathscr{P}_n$, we say that $\hat{x} = \{\hat{x}_1, \ldots, \hat{x}_n\}$ is *admissible* if $\hat{x}_i \in [x_{i-1}, x_i]$. The partition/admissible-point pair $(\mathscr{Q}_m, \hat{y})$ is a *refinement* of the pair $(\mathscr{P}_n, \hat{x})$ provided that three properties hold:

1. $m \geq n$
2. as sets $\mathscr{P}_n \subset \mathscr{Q}_m$
3. for each $i \in \{1, \ldots, n\}$ there is a $j \in \{1, \ldots, m\}$ so that $\hat{x}_i = \hat{y}_j$.

Intuitively, $(\mathscr{Q}_m, \hat{y})$ is a refinement if it dissects the interval more finely that $\mathscr{P}_n$, while at the same time respecting the pair $(\mathscr{P}_n, \hat{x})$'s cutoffs and admissible points.

Given f, $\mathscr{P}_n$, and admissible $\hat{x}$ we may define the associated *Riemann sum* as

$$\mathscr{R}(f, \mathscr{P}_n, \hat{x}) = \sum_{i=1}^{n} f(\hat{x}_i)\,\Delta x_i, \tag{3}$$

where $\Delta x_i = x_i - x_{i-1}$. Think of (3) as providing a formal approximation to the far right expression in (2), i.e.

$$\sum f(x)dx \approx \sum_{i=1}^{n} f(\hat{x}_i)\,\Delta x_i.$$

Now we just want to make this approximation precise, i.e. take a limit.

Informally, the *Riemann integral of f over $[a,b]$*, written $\int_a^b f(x)dx$, is the limit, when it is exists, of the Riemann sums (3), where the limit is taken over refinements. More formally, f is *Riemann integrable* on $[a,b]$ provided there a real number $\int_a^b f(x)dx$ such that for any $\varepsilon > 0$ there is a pair $(\mathscr{P}_n, \hat{x})$ with the property that whenever $(\mathscr{Q}_m, \hat{y})$ is a refinement of $(\mathscr{P}_n, \hat{x})$ it follows that

$$\left| \mathscr{R}(f, \mathscr{Q}_m, \hat{y}) - \int_a^b f(x)dx \right| < \varepsilon.$$

Importantly, *continuous functions are Riemann integrable.*

We conclude the following: if $F'(x) = f(x)$ and f is Riemann integrable then $F(b) - F(a) = \int_a^b f(x)dx$. This, of course, is not a proof: we have relied on the intuitive connection $dF = f'(x)dx$. The careful justification of this step is at the heart of the following theorem:

Theorem 3 (Fundamental theorem of calculus) *Let $g : [a,b] \to \mathbb{R}$ be continuous. Define $G : [a,b] \to \mathbb{R}$ by*

$$G(x) = \int_a^x g(s)ds.$$

Then $G'(x) = g(x)$.

Exercise 10 *Let $f : [a,b] \to \mathbb{R}$ be continuous. Is there a function $F : [a,b] \to \mathbb{R}$ whose derivative is given by f?*

To see how this theorem provides the answer to our original quest of finding the total change in F given all the little changes, we need the following result:

Exercise 11 *If $F, G : [a,b] \to \mathbb{R}$ are continuous and differentiable on (a,b), and if $F'(x) = G'(x)$ then $F(b) - F(a) = G(b) - G(a)$.*

Now assume that we know $F' \equiv f$ is continuous on $[a,b]$. Suppose we can find $G : [a,b] \to \mathbb{R}$ such that $G'(x) = f(x)$. Then, by the FTC, together with exercise 8 we know that $G(x) = \int_a^x f(x)dx + \alpha$. Observing that

$$G(a) = \int_a^a f(x)dx + \alpha = \alpha,$$

it follows that

$$G(b) - G(a) = \int_a^b f(x)dx.$$

Finally, since $F' = f = G'$, it follows from exercise 11 that $F(b) - F(a) = G(b) - G(a)$, whence

$$F(b) - F(a) = \int_a^b f(x)dx.$$

as desired. Thus, if we are given the little changes and want to compute the total change, instead of adding up all the little changes which is a difficult limit to even contemplate, all we need to do is to find a function whose derivative gives us the little changes, and then evaluate this function at the endpoints.

The fundamental theorem is used in many ways in economics. A common application is to solve differential equations, the idea being that you may know how a variable changes over time and you want to compute its value at each point in time. The next exercise provides an example.

Exercise 12 *Suppose that you know the change in the price level at time t is given by βt, where $\beta > 0$. What is the price level in time $t = 10$? Here you should assume that the price level is capture by a differentiable function $p : \mathbb{R}_+ \to \mathbb{R}_+$.*

More generally, it is useful to think of an integral as capturing an aggregrate or an average associated to a given function. For example, if a model has a collection of consumers indexed by the unit interval then aggregate consumption is the integral of individual consumption levels over the interval.

4 Exponents and logs

The derivative of $1/n \cdot x^n$ is x^{n-1} whenever $n \neq 0$. So what about $n = 0$? Or, perhaps said differently, what function is it whose derivative is x^{-1}? That's easy to answer: use the fundamental theorem. For $x \geq 1$ define

$$\log(x) = \int_1^x \frac{1}{t} dt, \tag{4}$$

and for $x \in (0,1)$ let $\log(x) \equiv -\log(1/x)$.

Exercise 13 *Show that* $\frac{d}{dx}\log x = 1/x$.

Note that (4) is the definition of the logarithm: it's a limit of Riemann sums, and you know what its derivative is because you know the FTC.

It can be shown that the range of log is $\mathbb{R}$. By the fundamental theorem, we know that log is a differentiable function; also, since its derivative is positive, we know that it is invertible. Let $\exp : \mathbb{R} \to (0,\infty)$ be its inverse. Let's compute the derivative of exp. We find

$$\begin{aligned} \log(\exp(x)) = x \quad &\Longrightarrow \quad \frac{d}{dx}\log(\exp(x)) = \frac{d}{dx}x \\ &\Longrightarrow \quad \frac{1}{\exp(x)}\frac{d}{dx}\exp(x) = 1 \\ &\Longrightarrow \quad \frac{d}{dx}\exp(x) = \exp(x), \end{aligned}$$

where the second implication uses the chain rule.

Exercise 14 *Show that if* $f : \mathbb{R} \to \mathbb{R}$ *is such that* $f'(x) = f(x)$ *then* $f(x) = \alpha\exp(x)$ *for some* $\alpha \in \mathbb{R}$.

Exercise 15 *Let* $f(x) = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. *Here* $0! \equiv 1$. *Assume this sum converges for all x and that you are free to differentiate across an infinite sum. Compute* $f'(x)$.

The exponential function, and its inverse, the logarithm, are the most important functions in mathematics. They are very likely the only functions you know of that are not polynomials (trig functions are built via exponentials). They have the following dual properties:

1. $\log(xy) = \log(x) + \log(y)$ and $\exp(x)\exp(y) = \exp(x+y)$.
2. $\log\left(x^{-1}\right) = -\log(x)$ and $\exp(-x) = \exp(x)^{-1}$.

These properties allow us to define the more general exponential of a positive real number:

$$x^{\alpha} \equiv \exp(\alpha\log(x)) \text{ for } x > 0, \alpha \in \mathbb{R}.$$

The following exercise has you verify that this function behaves like you think it should.

Exercise 16 *Show the following:*

1. $x^{\alpha}x^{\beta} = x^{\alpha+\beta}$
2. $x^{-\alpha} = \frac{1}{x^{\alpha}}$
3. $\frac{d}{dx}x^{\alpha} = \alpha x^{\alpha-1}$

4. $(x^\alpha)^\beta = x^{\alpha\beta}$
5. $\log(x^r) = r\log(x)$

The uses of logs and exponents in economics are myriad. Here are two examples.

Exercise 17 *Suppose gdp y grows at a constant rate $r > 0$, i.e. y satisfies $\dot{y} = ry$. Solve for y as a function of t. More generally, the exponential function can be defined as the solution to this differential equation.*

Exercise 18 *If $f : \mathbb{R} \to \mathbb{R}$ and $y = f(x)$ define the* elasticity *of y with respect to x as $\varepsilon_{yx} = \frac{dy}{dx}\frac{x}{y}$. Intuitively the elasticity is the percent change in y given a percent change in x; the advantage elasticity is that it's a unit-free measure of change. Let $f(x) = x^\alpha$ for some $\alpha \in \mathbb{R}$.*

1. *Compute ε_{yx}.*
2. *Let $\hat{y} = \log y$ and $\hat{x} = \log x$. Compute $\frac{d\hat{y}}{d\hat{x}}$.*

Thus by taking log transforms of the data and then performing a regression, the estimated coefficients are the elasticities.

In chapter we extend to a multivariate setting many of the notions and results obtained for univariate functions. In particular, we are interested in functions $f : \mathbb{R}^n \to \mathbb{R}^m$, and we exploit heavily the fact that $\mathbb{R}^n$ may be viewed as a *Hilbert* space, i.e. is a complete normed vector space with a norm – the usual Euclidean norm – that may be derived from an inner product: $\|x\|_2 \equiv \sqrt{\langle x, x\rangle}$. Unless otherwise stated, the norm used here will always be the Euclidean norm, and so to eliminate clutter, we will drop the subscript.

5 Continuity

The definition of continuity for functions from $\mathbb{R}^n$ to $\mathbb{R}^m$, or functions between any two metric spaces for that matter, is exactly the same as in the univariate case: continuous functions take convergent sequences to convergent sequences. Similarly, the results that continuity is preserved under point-wise addition and composition generalize to the multivariate case.

Recall that the main result for univariate continuous functions was the implicit function theorem. The most direct analog of this theorem in the multivariate setting is that continuous functions respective *connectivity*, that is, they take connected sets to connected sets. In the univariate case, this notion of connectedness is simple: a set is connected exactly when it is an interval. In the more general setting however, the notion becomes quite subtle, and is perhaps outside the reach of this review.

Instead, we state two important fixed point theorems, which provide the kinds of existence results needed for the establishment of equilibria. Recall that a compact subset of $\mathbb{R}^n$ is exactly a closed and bounded set. A subset K of $\mathbb{R}^n$ is *convex* provided that whenever $x, y \in K$ it follows that $\alpha x + (1-\alpha)y \in K$ for all $\alpha \in (0,1)$.

Theorem 4 (Brouwer) *Let $K \subset \mathbb{R}^n$ be compact and convex, and let $f : K \to K$ be continuous. Then there is a point $x \in K$ such that $f(x) = x$.*

This existence theorem is quite powerful because its premise is weak and easy to check; however, like many existence theorems it suffers in the applicability of its conclusion: there is no mention of construction or uniqueness of the fixed point.

The other fixed point theorem has a much stronger premise, but also a much stronger conclusion. We say that $f : \mathbb{R}^n \to \mathbb{R}^n$ is a *contraction* if there is a $\beta \in (0,1)$ such that for any $x, y \in \mathbb{R}^n$ we have that

$$\|f(x) - f(y)\| \leq \beta \|x - y\|.$$

Intuitively, f shrinks by at least a factor β the distance between points.

Exercise 19 *Show that if f is a contraction then it is continuous.*

Theorem 5 (Contraction mapping theorem) *Let $X \subset \mathbb{R}^n$ be closed. Suppose $f : X \to X$ is a contraction. Then there exists a unique $x^* \in X$ such that $f(x^*) = x^*$. Furthermore, if x_0 is any point in X then*

$$f^n(x_0) \equiv f(f^{n-1}(x_0)) \to x^*.$$

Perhaps it is not hard to imagine that this theorem holds on any Banach space. The power of this theorem is easy to see: not only do we get existence of a fixed point, we get uniqueness and we are provided a mechanism to approximate the solution: just iterate on f. This theorem is the reason dynamic programming works in practice as well as in theory.

One more theorem will be of great use for us.

Theorem 6 *Let $K \subset \mathbb{R}^n$ be compact and $f : K \to \mathbb{R}$ be continuous. Then there exists $x \in K$ such that*

$$f(x^*) = \sup_{x \in K} f(x) < \infty.$$

Intuitively, if f is a continuous real-valued function on a compact set then its supremum is finite and is attained by the function at some point in that set.

Exercise 20 *Two quick counterexamples:*

1. *Let $f:(0,1)\to\mathbb{R}$ be given by $f(x)=(1-x)^{-1}$. What is the* $\sup f$*?*
2. *Let $f:(0,1)\to\mathbb{R}$ be given by $f(x)=x$. What is the* $\sup f$*? Is it attained at a point in the set?*
3. *What happens when $(0,1)$ is replaced by $[0,1]$?*

Exercise 21 *Let $f:[a,b]\to[a,b]$ be continuously differentiable. Suppose that $f'(x)\in(0,1)$. Show that f is a contraction.*

6 Derivatives

To generalize the notion of derivative to that of a function from $\mathbb{R}^n$ to $\mathbb{R}^m$, two steps are needed: first, we must define a *partial derivative* from which ceteris paribus considerations arise; and then we must define the derivative at a point as a linear map. First things first.

Let $f:\mathbb{R}^n\to\mathbb{R}$. Then the *partial derivative of f with respect to x_i* at $x\in\mathbb{R}^n$, when it it exists, is computed as

$$f_i(x)=\lim_{\Delta x_i\to 0}\frac{f(x_1,\ldots,x_{i-1},x_i+\Delta x_i,x_{i+1},\ldots,x_n)-f(x)}{\Delta x_i}.$$

Intuitively, it is the derivative of f with respect to x_i, holding all other coordinates constant; or, it is the change in f that obtains when x changes only along the ith coordinate axis.

Higher-order partials are defined analogously, and hopefully the notation f_{ij}, i.e. the partial with respect to the jth component of the partial with respect to the ith component, is suggestive. It is a somewhat remarkable fact that if f_{ij} and f_{ji} are continuous then $f_{ij}=f_{ji}$. This is not intuitive to me, but extremely useful.

Exercise 22 *Let $y=f(k,l)=k^\alpha l^{1-\alpha}$ for $\alpha\in(0,1)$. We may interpret y as output contingent on the inputs capital k and labor l. The functional form provided is called "Cobb-Douglass". Compute $\partial y/\partial l$ and interpret. Compute the elasticities of output with respect to capital and labor. Compute the second partials of y and interpret.*

Now recall that a linear map from $\mathbb{R}$ to $\mathbb{R}$ is exactly and only multiplication. Let $f:\mathbb{R}\to\mathbb{R}$ be differentiable. Then $df=f'(x)dx$. Thus $f'(x)$ is the linear map that takes dx to df. To make this formal, we may work as follows: Let $T:\mathbb{R}\to\mathbb{R}$ be given by $T(y)=f'(x)y$.

Then

$$\begin{aligned} f'(x) &= \lim_{\Delta x \to 0} \frac{f(x+\Delta x)-f(x)}{\Delta x} \\ &\Longrightarrow \lim_{\Delta x \to 0} \left| \frac{f(x+\Delta x)-f(x)}{\Delta x} - f'(x) \right| = 0 \\ &\Longrightarrow \lim_{\Delta x \to 0} \left| \frac{f(x+\Delta x)-f(x)-f'(x)\Delta x}{\Delta x} \right| = 0 \\ &\Longrightarrow \lim_{\Delta x \to 0} \left| \frac{f(x+\Delta x)-f(x)-T(\Delta x)}{\Delta x} \right| = 0 \\ &\Longrightarrow \lim_{\Delta x \to 0} \frac{|f(x+\Delta x)-f(x)-T(\Delta x)|}{|\Delta x|} = 0. \end{aligned} \tag{5}$$

Thus the derivative of f at x is the linear map T satisfying (5). It further follows that the map $f' : \mathbb{R} \to \mathbb{R}$ is really a function from $\mathbb{R}$ to the linear maps from $\mathbb{R}$ to $\mathbb{R}$, i.e. $f : \mathbb{R} \to \mathbb{R}^*$. Of course $\mathbb{R}^* \approx \mathbb{R}$ which is why we usually think $f' : \mathbb{R} \to \mathbb{R}$. But this will become an issue for functions from $\mathbb{R}^n \to \mathbb{R}$. And to address this issue, further language is needed.

6.1 Total differentials

To improve intuition before turning to the full definition of a derivative, we first consider the notion of a *total differential*. Let $f : \mathbb{R}^2 \to \mathbb{R}$, and write $y = f(x_1, x_2)$. What happens to y when x_1 changes by dx_1 and x_2 is held fixed? From our univariate work we know that, in differential form, $dy = f_1 dx_1$. Now what happens if x_2 *also* changes, by, say, dx_2? It turns out that the corresponding differential is $dy = f_1 dx_1 + f_2 dx_2$. Intuitively, the total change in f is the sum of the changes in f resulting from changes in x along each coordinate axis. More generally, if $f : \mathbb{R}^n \to \mathbb{R}$ then

$$df = \sum_{i=1}^{n} f_i dx_i = f_i dx_i.$$

Written this way, df is the total differential of f. Notice that while I have ignored the presence of arguments, df is a local notion: the total differential of f depends on x since all of the f_i depend on x.

The total differential is an *extremely* useful tool for gaining intuition and deriving results concerning optimization and comparative statics. Here is an exercise to justify this claim:

Exercise 23 *Consider the Keynesian cross model of output determination, as given by*

$$Y = C(Y - T) + I + G,$$

where Y is real gdp, T is lump-sum taxes, I is exogenous investment and G is government spending. Compute the government-spending multiplier, that is, let G change by dG and compute the corresponding change dY in Y. Hint: let

$$f(Y,G) = Y - C(Y-T) - I - G,$$

and compute the total differential df. Then, noting that for the equation to continue to hold after the change in G it must be that $df = 0$. So set $df = 0$ and "solve" for dY in terms of dG. Observe that this process is replicated by taking the total differential of each side of the equation with respect to Y and G and solving for dY in terms of dG. We will make this process formal using the implicit function theorem below.

6.2 Derivatives

Again let $f : \mathbb{R}^n \to \mathbb{R}$. Let dx be a column vector of differentials, i.e.

$$dx = (dx_1, \ldots, dx_n)^T.$$

Now let Df be the *row* vector of partial derivatives (and you should be thinking linear functional here), i.e. $Df = (f_1, \ldots, f_n)$. This is the *derivative* of f, and again I should emphasize the point that is made formal below: this is a local notion, all of the partials are evaluated at a point in $\mathbb{R}^n$. In physics this row vector would be written ∇f, but this is not physics.

Observe that

$$df = \sum_{i=1}^{n} f_i dx_i = Df\,dx \tag{6}$$

Thus by writing Df as a row vector, and interpreting Df as a linear functional acting on dx, we recover the previous intuition that the derivative is a linear map Df taking changes dx in x to changes df in f.

To make this notion of a derivative formal, we need define with care the vector space of linear maps. If W and V are two normed vector spaces (we use $\|\cdot\|$ to identify the norm on each space), then we may let $L(V,W)$ be the collection of all linear maps from V to W. Noting that sums and scalar multiples of linear maps are again linear, we conclude that $L(V,W)$ is a vector space.

Exercise 24 *Show that if* $\dim W = m$ *and* $\dim V = n$ *then* $\dim L(V,W) = nm$. *Hint:* $L(V,W) \approx \mathbb{R}^{m\times n}$.

We have maintained an attachment to abstract vector spaces, that is, vector spaces without preferred coordinates, to aid the transition to infinite dimensions; however the notion of a

derivative is greatly aided by interpreting a normed finite-dimensional vector space as $\mathbb{R}^n$.[2] Thus let $V = \mathbb{R}^n$, which we endow with the usual L^2 (or Cartesian) norm: $\|v\|^2 = \sum v_i^2$.

Let $f : \mathbb{R}^n \to \mathbb{R}^m$, and let $x \in \mathbb{R}^n$. Then f is *differentiable at x* provided that there exists a linear map $Df(x) \in L(\mathbb{R}^n, \mathbb{R}^m)$ so that the following limit exists and equals zero:

$$\lim_{\|\Delta x\| \to 0} \frac{\|f(x+\Delta x) - f(x) - Df(x)(\Delta x)\|}{\|\Delta x\|} = 0. \tag{7}$$

If f is differentiable at x then we call $Df(x)$ the derivative of f at x. Notice that if f is differentiable for all points x in some open set $U \subset \mathbb{R}^n$ then $Df : U \to L(\mathbb{R}^n, \mathbb{R}^m)$. Notice that (7) is the higher-dimensional analog of (5).

Just as before, Df provides a linear approximation to f. More carefully, think of $Df(x)(\Delta x)$ as approximating how the image of f at x changes when x changes by Δx, noting that Δx is now a vector of changes. In differential form we write $df = Df(x)dx$.

The discussion of derivatives so far is quite abstract, but the following theorem will help anchor ideas, and greatly aid in the computation of derivatives. Here and in the sequel we will use superscripts to index the range. Thus $f : \mathbb{R}^n \to \mathbb{R}^m$ then

$$f = (f^1, \ldots, f^m)^T,$$

i.e. f is comprised of a column vector of functions $f^i : \mathbb{R}^n \to \mathbb{R}$.

Theorem 7 *Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$ is differentiable (on some open set) and that f^i_j is continuous. Then the derivative Df of f is the matrix of first partials: $Df_{ij} = f^i_j$.*

This theorem corresponds in a very precise way to the observation (6). Specifically, let $f : \mathbb{R}^n \to \mathbb{R}^m$. Then $f = (f^i)$ where $f^i : \mathbb{R}^n \to \mathbb{R}$. It follows that $df = (df^i)$, i.e. the change in f is the column of changes in the f^i. By (6) we have that $df^i = Df^i dx$. But now notice that

$$Df^i = (f^1_i, \ldots, f^n_i),$$

i.e. Df^i is the ith row of Df. Thus $df = Dfdx$ as desired.

The chain rule continues to hold, but remember that, since we are dealing with matrices, order matters.

Theorem 8 (Chain rule) *If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^k$ are differentiable then $h : \mathbb{R}^n \to \mathbb{R}^k$ is differentiable and $Dh(v) = Dg(f(v)) \circ Df(v)$.*

[2] This is not required: the Frechet derivative generalizes the usual notion of differentiation to functions between normed linear spaces; but, to interpret the Frechet derivative as a "matrix of first partials," a coordinate system is required: after all, how can a partial derivative be computed without reference to a specific direction?

6.3 Taylor's theorem part I: the easy way

Suppose $K \subset \mathbb{R}^n$ is compact, and let $f : K \to \mathbb{R}$ be continuous. It would be great to know how to approximate f by some function that we understand. It can be shown that f can be approximated arbitrarily well by a polynomial: this is the Stone-Weierstrass theorem. More formally, recall the definition of the normed vector space $(C(K), \|\cdot\|_\infty)$ above, and notice that $f \in C(K)$. The theorem says that for any $\varepsilon > 0$ there is a polynomial $p \in C(K)$ such that $\|f - p\|_\infty < \varepsilon$.

As wonderful as this result it – after all, polynomials can be evaluated by computers – the theorem doesn't determine the approximation – it's an existence theorem. If we allow for the additional restriction that our functions are differentiable, and if we only require local approximations, then Taylor's theorem provides the construction we are looking for. In this subsection, we'll work somewhat loosely so that the practical aspects of the theorem are laid bare. Tthe details are pinned down in the supplemental material. We start with the univariate case.

Let $U \subset \mathbb{R}$ be an open set and let $f : U \to \mathbb{R}$ be C^2, i.e. twice continuously differentiable. Taylor's theorem says that for a given $x^* \in U$,

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \mathscr{O}(|x - x^*|^2).$$

The function $\mathscr{O}$, pronounced "Big Oh", is a special terminology for error functions, and is used when the only interesting property of the function is how it behaviors near a contextually determined point, in this case the origin. Here, think of the $\mathscr{O}(y)$ as a function that is the same "size" as y, thus if y is small then $\mathscr{O}(y)$ is small. It follows that if x is near x^* then $|x - x^*|^2$, and thus $\mathscr{O}(|x - x^*|^2)$, is very nearly zero, so

$$f(x) \approx f(x^*) + f'(x^*)(x - x^*). \tag{8}$$

We have used Taylor's theorem to obtain the *first-order approximation* of f around x^*. The RHS of (8) is called the first-order (or linear) approximation, and is a polynomial of degree one. Note that it is exactly Taylor's theorem that makes formal the sense in which a differentiable function is locally linear.

Now let's up the ante. Suppose f is C^3. Then Taylor's theorem says

$$f(x) = f(x^*) + f'(x^*)(x - x^*) + \frac{1}{2} f''(x^*)(x - x^*)^2 + \mathscr{O}(|x - x^*|^3). \tag{9}$$

Without the error term, the RHS of (9) is, naturally, the *second-order approximation* of f around x^*, and is a polynomial of degree 2. If x is sufficiently near x^* then the second-order approximation will necessarily be better than the first-order one. Intuitively, the second-order term is partially correcting for the fact that f' varies with x.

The general pattern can now be recognized: If f is C^{N+1} then

$$f(x) = f(x^*) + \sum_{n=1}^{N} \frac{f^{(n)}(x^*)}{n!}(x-x^*)^n + \mathscr{O}(|x-x^*|^{N+1}).$$

Thus, provided a function is sufficiently smooth, Taylor's theorem provides a polynomial approximation to arbitrary accuracy.

Exercise 25 *Compute the Taylor expanse of e^x around the origin.*

Now let's turn to the multivariate case. Let $U \subset \mathbb{R}^n$, and assume that f is C^2, which means all the partials are twice continuously differentiable. The first order approximation of f near x^* is a the direct analog to (8):

$$f(x) = f(x^*) + Df(x^*)(x-x^*) + \mathscr{O}(\|x-x^*\|^2). \tag{10}$$

Note that $Df(x^*)$ may be viewed as a row-vector, and as such, it is the linear functional taking changes in x to changes in f; thus (10) comports with our intuition that $df = Df(x^*)dx$. Indeed, (10) is a precise statement of this intuition.

Assume now that f is C^3. To obtain the second-order (and higher order) approximates, we need to differentiate Df, and this takes considerable machinery. In particular, as noted above, $Df : U \to L(\mathbb{R}^n, \mathbb{R}^m)$: how do we differentiate that beast? In the supplemental section, I provide an intro to the needed development – tensors. Here, I'll just cover mechanics and intuition. The *Hessian* D^2f of f at x^* is the $n \times n$ symmetric (why?) "matrix of second partials", i.e. $D^2f(x^*) = (f^i_j(x^*))$. Taylor's theorem says

$$f(x) = f(x^*) + Df(x^*)(x-x^*) + \frac{1}{2}(x-x^*)^T D^2f(x^*)(x-x^*) + \mathscr{O}(\|x-x^*\|^3). \tag{11}$$

The second-order term is captured by the Hessian acting a quadratic form. As in the univariate case, it corrects for the fact that Df is not constant. It may be helpful to write this out in more detail:

$$(x-x^*)^T D^2f(x^*)(x-x^*) = \sum_{ij} f_{ij}(x^*)(x_i - x^*_i)(x_j - x^*_j).$$

A representative summand is $f_{ij}(x^*)(x_i - x^*_i)(x_j - x^*_j)$. Intuitively, we think this way: if x moves a little along the ith coordinate then f moves by $df = f_i(x^*)dx_i$; but if x also moves in the jth direction then $f_i(x^*)dx_i$ changes by $f_{ij}(x^*)dx_i dx_j$, so we should adjust for this.

7 The implicit function theorem

The implicit function theorem is a tool used almost daily during the first year, principally when invoked to assess comparative statics experiments. To motivate its application, some terminology specific to economics is warranted.

For this section an economic model is a triple (x,y,f) with $x \in \mathbb{R}^m$ representing the *exogenous* variables, $y \in \mathbb{R}^n$ representing the *endogenous* variables, and

$$f : \mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{R}^n$$

capturing *equilibrium restrictions*.[3] The idea is this: we take the values of x as given and then use the n equations $f(x,y) = 0$ to determine the values of y. In this way, we may thing of f defining the y's as implicit functions of the x's. This is important: the equilibrium restrictions of the model provide the mechanism through which we may write the endogenous variables as functions of the exogenous variables, at least implicitly.

A simple example might be useful to keep in mind. Thus let $D(p,I)$ be the quantity demanded (of some good) given price p and income I, and let $S(p)$ be the quantity supplied given the price. Define $Z(p,I) = D(p,I) - S(p)$ to be the excess demand function. This model of a simple market has I as the exogenous variable, p as the endogenous variable, and $Z(p,I) = 0$ as the equilibrium restriction.

In practice, while we may be able to conclude that an equilibrium exists and is even unique, it is rare that we can explicitly solve for the endogenous variables a functions of the exogenous variables: f is nonlinear after all. On the other hand, by computing total differentials, we can solve for the *changes* in the endogenous variables in terms of the changes in the exogenous variables: that's because the total differential is *linear* in the differential changes!

Return now to the example. The comparative statics question is this: what happens to p if I increases? Intuitively, writing p^* as the equilibrium value of p, we know that $p^* = p^*(I)$ and we would like to compute $\partial p^*/\partial I$. To do this, take the total differential of dZ of Z:

$$dZ = Z_p dp + Z_I dI.$$

We need the system to remain in equilibrium, i.e. after the change in I it must be that Z still equals zero, thus $dZ = 0$. We conclude

$$Z_p dp + Z_I dI = 0 \text{ or } \frac{\partial p^*}{\partial I} = -\frac{Z_I}{Z_p}.$$

Noting that $Z_I = D_I > 0$ and $Z_p = D_p - S_p < 0$, it follows that

$$\frac{\partial p^*}{\partial I} = \frac{D_I}{S_p - D_p} > 0.$$

Thus an increase in income leads to an increase in equilibrium price.

The more general case follows the same pattern. The comparative statics question is this: if x changes by dx, how must y change so that the model remains in equilibrium? Intuitively, we may simply compute total differentials again, and set the changes to zero:

$$df = D_x f dx + D_y f dy = 0$$

[3] Recall the exogenous variables are taken as given by the model (determined by nature or fiat) and the endogenous variables are pinned down by the model.

and solve for dy in terms of dx. The difference now is that $D_x f$ is the $n \times m$ matrix of partials of the f^i with respect to the variables $x_1, \ldots, x_m$ and $D_y f$ is the $n \times n$ matrix of partials of the f^i with respect to the variables $y_1, \ldots, y_n$. It is perhaps useful to be explicit in this case:

$$D_x f(x,y) = \begin{pmatrix} f^1_{x_1}(x,y) & \cdots & f^1_{x_m}(x,y) \\ \vdots & \ddots & \vdots \\ f^n_{x_1}(x,y) & \cdots & f^n_{x_m}(x,y) \end{pmatrix}, D_y f(x,y) = \begin{pmatrix} f^1_{y_1}(x,y) & \cdots & f^1_{y_n}(x,y) \\ \vdots & \ddots & \vdots \\ f^n_{y_1}(x,y) & \cdots & f^n_{y_n}(x,y) \end{pmatrix}.$$

We can "solve" for dy provided $\det D_y f(x,y) \neq 0$. This is made formal in the following theorem.

Theorem 9 (Implicit function theorem) *Suppose $f : \mathbb{R}^m \oplus \mathbb{R}^n \to \mathbb{R}^n$ has continuous first partials. If $f(x^*, y^*) = 0$ and $\det D_y f(x^*, y^*) \neq 0$ then there exists open set $U \subset \mathbb{R}^m$, with $x^* \in U$, and a continuously differentiable function $g : U \to \mathbb{R}^n$ so that*

1. $y^* = g(x^*)$
2. $x \in U \implies f(x, g(x)) = 0$
3. $Dg(x^*) = -D_y f(x^*, y^*)^{-1} \circ D_x f(x^*, y^*)$

Like most theorems, the statement of the implicit function theorem is somewhat daunting. However, in practice, the application of the implicit function theorem is *very* easy. It goes like this:

1. Check premise that functions are nice.
2. Compute the total differential of each side of each equation with respect to *all* the endogenous variables and *only* the exogenous variables under examination.
3. Solve for the endogenous differentials in terms of the exogenous differentials.

Item 2 is important: you don't need to write your list of equations identifying your equilibrium as a function which, when set equal to zero, reproduces the equilibrium restrictions. Just start differentiating. And, you don't need to worry about exogenous variables that aren't changing in your particular comparative statics experiment – and most often, only one variable changes! Let's do some examples.

Exercise 26 *Consider the following IS-LM Model*

$$\begin{aligned} Y &= C(Y - T, r) + I(r, Y) + G. \\ \frac{M}{P} &= L(Y, r). \end{aligned}$$

The endogenous variables are Y and r, and the exogenous variables are T, G, M, and P. Assume $0 < MPC + I_Y < 1, C_r < 0, I_r < 0, L_r < 0, L_Y > 0$. Use the implicit function theorem to compute the signs of $\partial r/\partial M$ and $\partial Y/\partial T$.

Exercise 27 *Consider the following consumer choice problem:*

$$\begin{aligned} \max \quad & u(x,y) \\ & p_x x + p_y y = I \end{aligned}$$

where $u(x,y) = x + h(y)$, with $h' > 0$ and $h'' < 0$. Apply the implicit function theorem to the first order conditions to compute $\partial y/\partial p_x$.

Exercise 28 *Consider the following consumer choice problem:*

$$\begin{aligned} \max \quad & u(c_1) + u(c_2) \\ & c_1 + s = y_1 \\ & c_2 = (1+r)s + y_2 \end{aligned}$$

where $u(c) = \frac{1}{1-\sigma}\left(c^{1-\sigma} - 1\right)$, and $\sigma > 0$. Here, think of c_1 as consumption in period 1, and c_2 as consumption in period 2; also, y_i is income in period i and r is the interest rate on savings s. Use the implicit function theorem to determine the range of values for σ for which $\partial c_1/\partial r > 0$.

8 Supplementary material

8.1 Taylor's theorem part II: the right way

We begin with a precise definition of the "order" language used to indicating degree of local approximation. Let $f, g : \mathbb{R}^n \to \mathbb{R}^m$. We say $f = g + \mathscr{O}(n, x_0)$ provided that

$$\lim_{x \to x_0} \frac{\|f(x) - g(x)\|}{\|x - x_0\|^n} < \infty,$$

And we call g and $n-1$-order approximation to f. Often, the point x_0 is implicit and we will write $f = g + \mathscr{O}(n)$. Intuitively, if $f = g + \mathscr{O}(n, x_0)$ then near x_0, $\|f - g\|$ behaves like a polynomial of degree n in $\|x - x_0\|$. Of course, if n is large and $\|x - x_0\|$, a degree n polynomial in $\|x - x_0\|$ is vanishingly small, so that f looks very much like g.

Exercise 29 *If $f = g + \mathscr{O}(n, x_0)$ and $g = h + \mathscr{O}(n, x_0)$ then $f = h + \mathscr{O}(n, x_0)$ and $g - h = \mathscr{O}(n, x_0)$.*

Let's begin with the univariate case. Let $x_0 \in \mathbb{R}$, and let $f : \mathbb{R} \to \mathbb{R}$ and assume f is C^{N+1}, that is, $D^{N+1}f$ is continuous, at least in a neighborhood of some x_0. Define $p_N(f)(\cdot, x_0) : \mathbb{R} \to \mathbb{R}$ as follows:

$$p_N(f)(x, x_0) = f(x_0) + \sum_{n=1}^{N} \frac{D^n f(x_0)}{n!}(x - x_0)^n.$$

The function $p_N(f)(\cdot, x_0)$ is called the N-th order Taylor polynomial of f at x_0.

Theorem 10 *Let $x_0 \in \mathbb{R}$, and let $f : \mathbb{R} \to \mathbb{R}$ and assume f is C^{N+1}. Then*

$$f = p_N(f)(\cdot, x_0) + \mathscr{O}(N+1). \tag{12}$$

In fact,

$$|f(x) - p_N(f)(x, x_0)| \leq \left(\frac{|x - x_0|^{N+1}}{(N+1)!} \right) \sup \left\{ \left| D^{N+1} f(x_0 + \xi) \right| : |\xi| < |x - x_0| \right\}. \tag{13}$$

Equation (12) together with Exercise 1 shows that $p_N(f)(\cdot, x_0)$ is the unique N-th order approximation to f near x_0. Equation (15) provides a simple error bound to this approximation in terms of the size of the $N+1$-st derivative. The following exercise establishes a precise sense in which $p_N(f)(\cdot, x_0)$ provides a good approximation to f.

Exercise 30 *If $f = g + \mathscr{O}(n)$ and $f \neq h + \mathscr{O}(n)$ then there exists $\delta > 0$ so that $|x - x_0| < \delta$ implies $|f(x) - g(x)| < |f(x) - h(x)|$. So g is a better local approximation than h.*

Most functional equations relevant to economic models have solutions in $C^k(\mathbb{R}^n)$, and so local approximation requires the multivariate version of Talyor's theorem. Conceptually, this version is not more difficult: we approximate our solution locally using a multivariate polynomial of finite degree; and the polynomial's coefficients coincide with the higher-order derivatives of the associated function. In practice the situation is complicated by the fact that $D^n f(x)$ is an n-th order tensor, which can be somewhat tedious to work with.

Recall from Chapter 2 that if $f : \mathbb{R}^n \to \mathbb{R}$ then $D^k f(x)$ may be interpreted as a multi-indexed list with elements $f_{i_1 \cdots i_k}$ for $i_j \in \{1, \ldots, n\}$. Using Einstein's notation, we may write, for example,

$$\begin{aligned} f_i(x^*)(x_i - x_i^*) &= f_1(x^*)(x_1 - x_2^*) + f_2(x^*)(x_2 - x_2^*) \\ f_{ij}(x^*)(x_i - x_i^*)(x_j - x_j^*) &= f_{11}(x^*)(x_1 - x_1^*)^2 + 2f_{12}(x^*)(x_1 - x_1^*)(x_2 - x_2^*) + f_{22}(x^*)(x_2 - x_2^*)^2 \end{aligned}$$

With this notation, we may write the Taylor polynomial as

$$p_N(f)(x, x^*) = f(x^*) + \sum_{k=1}^{N} \frac{1}{k!} f_{i_1 \cdots i_k}(x^*) \prod_{j=1}^{k} (x_{i_j} - x_{i_j}^*).$$

We have the following result:

Theorem 11 *Let $x^* \in \mathbb{R}$, and let $f : \mathbb{R}^n \to \mathbb{R}$ and assume f is C^{N+1}. Then*

$$f = p_N(f)(\cdot, x^*) + \mathscr{O}(N+1). \tag{14}$$

In fact,

$$|f(x) - p_N(f)(x,x^*)| \leq \left(\frac{\|x-x^*\|^{N+1}}{(N+1)!} \right) \sup \left\{ \left\| D^{N+1} f(x^* + \xi) \right\| : |\xi| < \|x - x^*\| \right\}. \tag{15}$$

8.2 Direction and projection

A norm provides a vector space with a notion of distance; an inner product imparts geometry. If V is a (real) vector space then an inner product is a function $\langle \cdot, \cdot \rangle : V \times V \to R$ which is

1. Symmetric: $\langle v, w \rangle = \langle w, v \rangle$,
2. Bilinear, i.e. linear in both variables,
3. Positive definite: $\langle v, v \rangle \geq 0$ and $\langle v, v \rangle = 0$ if and only if $v = 0$.

The associated notion of geometry is provided by the concept of orthogonality: the vectors v and w are orthogonal if $\langle v, w \rangle = 0$.

If $V = \mathbb{R}^n$ then the canonical inner product is the usual dot product: $\langle v, w \rangle = v'w$; however, if Q is any $n \times n$ positive definite matrix then $v'Qw$ also identifies an inner-product (in fact, these quadratic forms characterize the collection of inner products on $\mathbb{R}^n$). The norm on V induced by an inner product (which will always be the assumed norm, given an inner product) is $\|v\|^2 = \langle v, v \rangle$. Notice that the norm induced by the canonical inner product on $\mathbb{R}^n$ is the usual Euclidian norm.[4]

Exercise 31 *Show that finite set of orthogonal vectors is linearly independent.*

Exercise 32 *Show that if $v, w \in V$ then there exists $\alpha \in \mathbb{R}$ so that $v - \alpha w$ is orthogonal to w.*

How, then, to think about the general meaning of $\langle v, w \rangle$? Imagine w has unit length, and let $\xi = v - \alpha w$ so that $\langle w, \xi \rangle = 0$. Then

$$\langle v, w \rangle = \alpha \langle w, w \rangle + \beta \langle w, \xi \rangle = \alpha,$$

where the last equality follows since w has unit norm. Thus $\langle v, w \rangle$ identifies how far v extends in the direction determined by w, that is, $\langle v, w \rangle$ is the projection of v onto w.

[4] Nothing in the definition of an inner product or the induced norm requires finite dimensions.

This notion of projection can be generalized. Suppose $B = \{b_1, \ldots, b_m\} \subset V$, and let $v \in V$. Our goal is to find a vector $w \in \text{Span}(B)$ so that $v = w + \xi$ where ξ is orthogonal to b_i. Thus we seek w so that $\langle b_i, v - w \rangle = 0$ for each i. Writing $w = \sum \alpha_i b_i$, we have

$$\langle b_i, v - w \rangle = 0 \implies \langle b_i, v \rangle = \sum_j \alpha_j \langle b_i, b_j \rangle .$$

Letting M be the $n \times m$ matrix having columns as b_i, we obtain $\alpha = (M'M)^{-1}M'v$, where the invertibility comes from the fact that we may assume B is linearly independent. It is no coincide that this formula resembles the usual linear regression formula: indeed, it is a special case.

The relevance of orthogonality is also evident here. Given a set of linearly independent vectors B, we may always construct a new set, $\hat{B}$ so that B and $\hat{B}$ span the same space and the elements of $\hat{B}$ are orthogonal. In this case, the columns of the matrix M are orthogonal, so that $M'M$ is diagonal. A diagonal matrix is very easy to invert, and has minimal conditioning number. These benefits are analogous to those arising from conditioning on independent regressors in a linear regression model. Note that the Schur decomposition above generates orthogonal matrices Q and Z. In fact, these matrices are unitary, so that $Q^{-1} = Q'$ and $Z^{-1} = Z'$; thus they are very easy to invert as well.

A final concept relating to direction, that of directional derivative, will also be useful later. Let $f : \mathbb{R}^n \to \mathbb{R}^m$, and recall the interpretation that if $x \in \mathbb{R}^n$ changes by $dx \in \mathbb{R}^n$ then $f \in \mathbb{R}^m$ changes by $df = Dfdx$. We consider the same intuition here, but now thing of x as changing in a given direction, as specified by the unit vector $v \in \mathbb{R}^n$. The directional derivative, then, is intended to provide a measure of how fast f is changing when x changes in the direction of v. More formally, fix $x_0 \in \mathbb{R}^n$ and let $g : \mathbb{R} \to \mathbb{R}^n$ be given by $g(t) = x_0 + tv$. Let $h : \mathbb{R} \to \mathbb{R}^m$ be given by $h(t) = f(g(t))$. Then $h'(0) = Df(x_0) \cdot v$ is defined as the derivative of f in the direction of v. Notice that if $m = 1$ then $Df(x_0) \cdot v$ may be interpreted as the rate of change of f projected onto the vector v. For this rate of change to be maximal, v should "point" in the same direction of the derivative itself: this is why we often say that the derivative of f points in the direction of the maximal rate of change.

8.3 Tensors and higher-order derivatives

Deep breath. Let V be an n-dimensional vector space. A linear functional on V is a linear map from V to $\mathbb{R}$. The set of all linear functionals on V is a vector space, called the *dual space* of V, and is denoted V^*. Some notation: of B is any basis for V, and $v \in V$, let $\alpha^B(v)$ the be coordinates of v with respect to B. Now let B be any basis for V, and define $T^B : V \to V^*$ by $T^B(v)(w) = (\alpha^B(v))'(\alpha^B(w))$. The map T^B is a vector space isomorphism: $\dim(V^*) = n$, and given a basis B, any vector $v \in V$ may be naturally interpreted as a linear functional $v^* \in V^*$.

Let $\{V_i\}_{i=1}^m$ be a finite collection of n-dimensional real vector spaces. A function $f : \prod_i V_i \to \mathbb{R}$ is k-linear if given any $v = (v_1, \ldots, v_k)$ and any $i \in \{1, \ldots, k\}$, the function $f_i : V_i \to \mathbb{R}$ given by

$$f_i(w) = f(v_1, \ldots, v_{i-1}, w, v_{i+1}, \ldots, v_k)$$

is linear. Let $\bigotimes_i V_i$ be the set of all k-linear maps on $\prod_i V_i$. It can be shown that $\bigotimes_i V_i$ is a real vector space. This space is called the *tensor product* of the spaces V_i, and and its elements are k^{th}-order tensors. Notice that V^* may be thought of as a trivial tensor product, so that a vector may be interpreted as a first order tensor.

To examine higher order tensors, let's focus on the case in which all the V_i are the same. Let V be an n-dimensional vector space and B be any basis for V. Let $A = (a_{ij})$ be any $n \times n$ matrix implicitly written with respect to the basis B; note that we are not interpreting A a linear map now. Define $T^A : V \times V \to \mathbb{R}$ by $T(v, w) = v'Aw$. Then $T \in V \bigotimes V$, and further, every element of $V \bigotimes V$ has this form. Thus an $n \times n$ matrix may be interpreted as a second-order tensor.

Let V be an n-dimensional vector space and B be any basis for V. The $n \times n$ matrix $A = (a_{ij})$ may be interpreted as a doubly-indexed list, so that

$$T(v, w) = \sum_i \sum_j a_{ij} v_i w_j.$$

This suggests the following generalization: if $A = (a_{ijk})$ is a triply-indexed list then we define $T^A : V \times V \times V \to \mathbb{R}$ by

$$T^A(v, w, u) = \sum_i \sum_j \sum_k a_{ijk} v_i w_j u_k.$$

It is straightforward to show that $T^A \in V \bigotimes V \bigotimes V$, and further, every element of $V \bigotimes V \bigotimes V$ looks like A.

In general, written against a given basis, a k^{th}-order tensor is a list with k-indices. As you can imagine, the notation gets a bit tricky with higher order tensors, which is why Einstein invented the following "sum" notation for multi-indexed lists. Assume that the underlying vector space is n-dimensional so that each index "runs" from 1 to n. The idea is to replace an indexed list by a representative element in such a way as to efficiently manipulate sums (and derivatives!). The only rule is that an index appearing twice indicates a sum, and is thus "integrated out;" otherwise, the index is assume to run over all possible values. To see how this works, let's look at a few examples. Assume that we have chosen a basis, so we may represent any vector $v \in V$ in terms of its coordinates.

1. The vector $v = (\alpha_1, \ldots, \alpha_n) \in V$ is represented α_i. Note that the index i only appears once, so the index runs over all possible values.

2. The matrix A given by
$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix}$$
is represented as a_{ij}.
3. If $v = \alpha_i$ and $w = \beta_i$ then $v'w = \alpha_i\beta_i$.
4. If $A = a_{ij}$ and $v = \alpha_i$ then $Av = a_{ij}\alpha_j$.
5. To compute the action of the 3rd-order tensor T associated to the list a_{ijk} on $v = v_i$, $w = w_i$, $u = u_i$, we write $T(v,w,u) = a_{ijk}v_iw_ju_k$.
6. The notation can be used to easily change coordinates. Let $B = \{b_1,\ldots,b_n\}$ and $D = \{d_1,\ldots,d_n\}$ be bases for V, and write $B = \beta_{ij}d_i$. If $v \in V$ then $v = \alpha_j^B b_j = \beta_{ij}\alpha_j^B d_i$.

A tensor is a multilinear map, and vectors and matrices may be viewed as special cases. Our interest in tensors lies in their connection to derivatives. Recall that if $f : \mathbb{R}^n \to \mathbb{R}$ then $Df : \mathbb{R}^n \to L(\mathbb{R}^n,\mathbb{R})$: intuitively, $Df(x)(dx)$ tells us how f moves if x moves to $x+dx$. So similarly, $D^2f : \mathbb{R}^n \to L(\mathbb{R}^n, L(\mathbb{R}^n,\mathbb{R}))$. Intuitively, $D^2f(x)(dx)$ tells us how Df moves if x moves to $x+dx$. This is perhaps not so intuitive, but there is a better way to think about it. If $g \in L(\mathbb{R}^n, L(\mathbb{R}^n,\mathbb{R}))$, we may define $\hat{g} \in \mathbb{R}^n \otimes \mathbb{R}^n$ by $\hat{g}(x,y) = g(x)(y)$. In this way $L(\mathbb{R}^n, L(\mathbb{R}^n,\mathbb{R})) \simeq \mathbb{R}^n \otimes \mathbb{R}^n$. We may then interpret $D^2f(x) \in \mathbb{R}^n \otimes \mathbb{R}^n$ as follows: $D^2f(x)(dx,dy)$ tells us how the rate of change of f in the direction of dx changes when x changes in the direction of dy. More generally $D^kf(x)$ is a k^{th}-order tensor – a multilinear map.

That's all fine in theory, but how about in practice? If $f : \mathbb{R}^n \to \mathbb{R}$, denote by f_i the partial derivative of f with respect to the i-th variable. Note that f_i is a function from $\mathbb{R}^n$ to $\mathbb{R}$, and so f_{ij} is its partial with respect to the j-th variable. Coupling this notation with Einstein's we may write $Df = f_i$, $D^2f = f_{ij}$, $D^3f = f_{ijk}$, etc.

It gets better. If $f : \mathbb{R}^n \to \mathbb{R}^m$ then we will use superscripts to indicate the coordinates in the range. The $f = (f^1,\ldots,f^m)'$, or, following Einstein, $f = f^i$. In this case, $Df = f^i_j$, $D^2f = f^i_{jk}$ and $D^3f = f^i_{jkl}$. To be more concrete, consider D^2f. For fixed i, $f^i : \mathbb{R}^n \to \mathbb{R}$, so that $f^i_{jk} \in \mathbb{R}^n \otimes \mathbb{R}^n$. Thus
$$D^2f \in \bigoplus_{i=1}^{m} (\mathbb{R}^n \otimes \mathbb{R}^n).$$

Now fix $x \in V$. Then $D^2f(x)(dx,dy)$ yields an m-dimensional vector, telling us how the rate of change of each f^i in the direction of dx changes when x changes in the direction of dy.

As emphasized, $D^k f(x)$ is an m-dimensional vector of k-linear maps on $\mathbb{R}^n$. When applying Talyor's theorem in Chapter 5, it will be important to explicitly evaluate $D^k f(x)$ at an element $(v_1,\ldots,v_k) \in V \times \ldots \times V$. Using Einstein notation, we have that

$$D^k f(x)(v_1,\ldots,v_k) = f^i_{j_1,\ldots,j_k}(x) v_{j_1} \cdot \ldots \cdot v_{j_k} \in \mathbb{R}^m.$$

We note that if $m = 1$ then $D^2 f(x)(v,w)) = f_{ij} v_i w_j = v'(f_{ij})w$, where (f_{ij}) is the usual matrix formulation of the second derivative, sometimes called the Hessian.

The power of Einstein's notation is evidenced in the Chain rule. Let $f : \mathbb{R}^n \to \mathbb{R}^N$ and $g : \mathbb{R}^N \to \mathbb{R}^m$. Let $h = g \circ f$. Then $Dh = g^i_r f^r_j$. Not bad. Next, we have to also use the product rule:

$$D^2 h = g^i_{rs} f^s_k f^r_j + g^i_r f^r_{jk}.$$

One more:

$$D^3 h = g^i_{rst} f^t_l f^s_k f^r_j + g^i_{rs} f^s_{kl} f^r_j + g^i_{rs} f^s_k f^r_{jl} + g^i_{rt} f^t_l f^r_{jk} + g^i_r f^j_{jkl}.$$

This may look a little messy, but imagine keeping track of all of the sums. Finally, notice that the only indices that don't integrate out are i, j, k, l which is consistent with $D^3 h$ being an m-dimensional vector of third order tensors.