14. Calculus of Several Variables

Although the chapter title refers to calculus, the focus of the chapter is rather narrower, concentrating on various types of derivatives. Much of the material here is considered advanced calculus or introductory analysis.

14.1 The Ordinary Derivative

We begin by recalling the ordinary derivative. Let $U \subset \mathbb{R}$ be an open set. A function $f\colon U \to \mathbb{R}$ is *differentiable at* $x_0 \in U$ if the limit

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

exists. The function f is *differentiable on* U if the *derivative* $f'(x_0)$ exists for every $x_0 \in U$. The expression inside the limit, both here and in similar definitions, is the *difference quotient*. The derivative is also denoted df/dx.

14.2 Notations for Derivatives I

Notations for Derivatives. There is a wealth of notations for derivatives. The more common include

$$f', \quad \frac{df}{dx}, \quad \dot{f}, \quad Df, \quad df, \quad \nabla f, \quad \frac{\partial f}{\partial x}, \quad \text{and} \quad f_x,$$

all of which are used in these notes.

Euler used the notation f' as early as 1749.[1] It was popularized by Lagrange in connection with a different view of the meaning of derivatives.[2] This notation is often named after Lagrange. The notation df/dx was introduced by Leibniz in 1675, and reflects his infinitesimal approach to derivatives.

[1] The Swiss mathematician Leonhard Euler (1707–1783) is often ranked as the greatest mathematician of the 18^{th} century. He also is known for work in astronomy, fluid dynamics, mechanics, and optics. Euler turned power series into a powerful tool of analysis. Using them, he discovered Euler's Magic Formula, $e^{it} = \cos t + i \sin t$. Power series also led to his discovery of Euler's constant, and how it related to the harmonic series, gamma function, and values of the Riemann zeta function (before Riemann), and many other functions and problems. He was a pioneer in graph theory with his solution of the Seven Bridges of Königsberg problem (the former Prussian city of Königsberg is now Russian Kalingrad). Then there are the Euler equations of fluid dynamics and Euler-Lagrange equations of the calculus of variations and optimal control, which are usually called Euler equations in economics. There are just too many important contributions from his 866 papers to mention more than a fraction.

[2] Joseph-Louis Lagrange (1736–1813) was an Italian-French mathematician and astronomer. Today his name is most familiar from the Lagrangian and Lagrange multipliers. He contributed to analysis, number theory, and classical and celestial mechanics. Lagrange did major work in the calculus of variations, where the optimality equations are called the Euler-Lagrange equations. He revamped Newtonian mechanics, introducing the Lagrangian $L = T - V$ where T is kinetic energy of a system and V its potential energy. Then Lagrange's equations

$$\frac{d}{dt}\left(\frac{\partial L}{\partial q_i}\right) - \frac{\partial L}{\partial q_i} = 0$$

are the equations of motion. The Lagrangian approach has also proven fruitful in quantum physics, both at the atomic and subatomic levels. Lagrange also made a comprehensive analysis of the vibrating string, including beats, echoes, and compound sounds. Later, he helped develop the metric system.

14.3 Notations for Derivatives II

More Notations for Derivatives. Another common notation is Newton's fluxion or dot notation, $\dot{f}$, most often used when f is a function of time. The operator-style notation Df was introduced in 1800 by Arbogast.[3]

We can indicate higher derivatives by using numerical superscripts, thus D^2f and D^3f indicate the second and third derivatives. When f is a function of several variables, subscripts indicate which derivative(s) we want. Thus for $f(x, y, z)$, we can write $D_x f$ or $D_{(x,y)} f$. Finally, if we are evaluating the function at a specific point, we put it last, e.g., $Df(x_0, y_0)$. Earlier uses of D by Euler and Johann Bernoulli lacked the operator aspect.[4]

Here are some examples of the previous notations for higher derivatives:

$$f'', f''', f^{(k)}, \frac{d^k f}{dx^k}, \ddot{f}, \dddot{f}, f^{(iv)}, D^k f, d^k f, \frac{\partial^{j+k} f}{\partial x^j \, \partial y^k}, \text{ and } f_{xx}, f_{xxy}.$$

We will use forms of Df when f is a function of multiple variables. The notation df will be reserved for the exterior derivative. It appears in differential forms both separately and inside integrals.

There are other notations for derivatives besides these. Historically, derivative notation was the Wild West in the 17^{th} and 18^{th} centuries.

[3] Louis François Antoine Arbogast (1759–1803) was an Alsatian French mathematician. Besides being the first to treat derivatives as operators, he also generalized the chain rule to higher derivatives. It is now known as Faà di Bruno's formula, although Faà di Bruno was not even born when Arbogast first wrote down the formula. He recast earlier algebraic manipulation of power series as operator equations, introducing the concept of factorials in the process.

[4] Johann (aka John or Jean) Bernoulli (1667–1748) was a member of the famous Swiss family of mathematicians. He discovered what is today known as l'Hôpital's Rule, which l'Hôpital learned from Bernoulli in 1694. He also worked on the brachistochrone problem, of finding the curve of fastest descent. Although he obtained the correct solution, a cycloid, his proof was incorrect and he took his brother Jacob's proof, claiming it as his own.

14.4 Notations for Partial Derivatives

Notations for Partial Derivatives. Finally, there are partial derivatives. Although various fancy forms of d, including this one, were used as early as 1694 by Leibniz, the standard notation $\partial f/\partial x$ was developed by Legendre in 1786.[5] Legendre later abandoned the notation. It was reintroduced by Jacobi in 1841.[6] Tait seems to have been the first to use the nabla or del operator, ∇, for the gradient (1867).[7] Hamilton had earlier used a rotated form of del. Finally, the origin of f_x for $\partial f/\partial x$ is harder to track down (no luck so far).

[5] The French mathematician Adrien-Marie Legendre (1752–1833) is better-known for the Legendre transformation (useful for economic duality) and Legendre polynomials. Legendre also introduced a three-fold classification of elliptic functions (integrals) that simplified their study.

[6] Carl Gustav Jacobi (1804–1851) was a German mathematician who worked on elliptic functions and their relation to the theta function. These have a number of applications to physics, including spinning tops, pendula, periodic and non-periodic flows, and more. He is partly responsible for developing the Hamilton-Jacobi formulation of classical mechanics and its solution via the Hamilton-Jacobi equations. This form is particularly useful for quantum mechanics as it allows particles to represented as waves. He developed the Jacobian determinant, which appears when changing variables in integrals.

[7] P.G. Tait (1831–1901) was a Scottish mathematical physicist best known for his work on knot theory and his *Treatise on Natural Philosophy* with Lord Kelvin. The three Tait conjectures were part of an attempt by Tait to classify knots. last one was resolved in 1987 using the Jones polynomial. He's also known for the Tait equation, which relates the density of a liquid and hydrostatic pressure.

14.5 Notations for Integrals and Anti-Derivatives

I only use two notations for integrals in these notes, the definite integral and the indefinite integral or anti-derivative.

$$\int_a^b f(x)\,dx \quad \text{and} \quad \int f(x)\,dx,$$

all of which are used in these notes. We could potentially write anti-derivatives and $D^{-1}f$, but I haven't found a need for that notation.

Notations for Integrals and Anti-derivatives. The integral symbol was introduced by Leibniz in 1675 and was widely adopted. Newton's alternative of putting the integrand into a box was so clumsy that its inconvenience overcame British nationalism regarding Newton (vs. Leibniz). The definite integral symbol dates to Fourier in 1822.[8] The vertical bar notation, $|_a^b$, for evaluating integrals was introduced the following year by Pierre Sarrus. There is somewhat more diversity in notations for anti-derivatives, but we don't use them in this course, except as indefinite integrals.

[8] Joseph Fourier (1768–1830) was a French mathematician and physicist. He's best known for Fourier (trigonometric) series and integrals. He applied the former to analyze heat transfer (heat equation) and vibration. This later developed into the field of Fourier or harmonic analysis, which is one of the major branches of mathematical analysis. He also found that the Earth is substantially warmer than it would be if the surface temperature was only affected by incoming and outgoing solar radiation, and suggested that the atmosphere may act as an insulator. This was the first suggestion that the Earth's surface temperature is affected by what is now called the greenhouse effect.

14.6 Difference Quotients

The difference quotients used to define the derivative are the slopes of the chords running between $\big(x_0, f(x_0)\big)$ and $\big(x_0 + h, f(x_0 + h)\big)$. We take the limit of these slopes as $h \to 0$ to find the slope of the tangent to the graph of f at x_0. The derivative can be used to define a line, L by

$$y = f(x_0) + f'(x_0)(x - x_0).$$

The line L is tangent to the graph of f at the point $\big(x_0, f(x_0)\big)$.

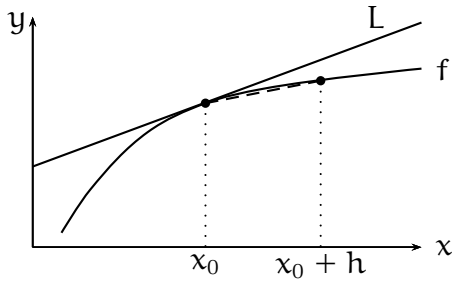

Figure 14.6.1: The dashed line segment is the chord between $\big(x_0, f(x_0)\big)$ and $\big(x_0 + h, f(x_0 + h)\big)$ for $h = 2$. The tangent line L at x_0 has the equation $y = f(x_0) + f'(x_0)(x - x_0)$.

Linear Approximation. The tangent line L is a linear approximation of $f(x)$ for x near x_0. Setting $\Delta x = x - x_0$, and $\Delta y = y - f(x_0)$, the equation becomes

$$\Delta y = y - f(x_0) = f'(x_0)\Delta x.$$

Now y approximates $f(x)$, so $f(x) \approx f(x_0) + f'(x_0)\Delta x$.

14.7 Partial Derivatives

In economics, we often deal with functions defined on a subset of $\mathbb{R}^m$. Production and utility functions, cost, expenditure, and profit functions are all examples. Let $U \subset \mathbb{R}^m$ be an open set and suppose $f\colon U \to \mathbb{R}$.

We define the i^{th} partial derivative of f at $\boldsymbol{x}_0 \in U$ by

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}_0) = \lim_{h\to 0} \frac{f(\boldsymbol{x}_0 + h\boldsymbol{e}_i) - f(\boldsymbol{x}_0)}{h}$$

provided the limit exists. Notice that only coordinate i is changed when computing the difference quotient. All other coordinates remain unchanged. This is why we can compute partial derivatives by treating variables other than x_i as constants.

14.8 Computing Partial Derivatives

In the partial derivative, all of the variables except for x_i are held constant. We then compute the partial derivative by taking the limit of difference quotients as $h \to 0$. It gives the slope of a tangent line in direction $\boldsymbol{e}_i$. It is also the rate of change of the function f in the direction $\boldsymbol{e}_i$.

Since all of the variables except x_i are held constant when computing $\partial f/\partial x_i$, we can compute $\partial f/\partial x_i$ by treating the other variables as constant and taking the ordinary derivative. Thus

$$\frac{\partial(x^2 + y^2)}{\partial x} = 2x,$$
$$\frac{\partial(x^2y + xyz + y^2z^3)}{\partial x} = 2xy + yz, \text{ and}$$
$$\frac{\partial(x^2y + xyz + y^2z^3)}{\partial y} = x^2 + xz + 2yz^3.$$

14.9 Higher Partial Derivatives

If a partial derivative of f exists, it too may have partial derivatives. Expressions such as

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial y\,\partial x} \quad \text{and} \quad \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = \frac{\partial^2 f}{\partial x^2}.$$

are used to write second (and higher) partial derivatives. Another common notation uses subscripts. Thus

$$\frac{\partial f}{\partial x} = f_x, \quad \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = (f_x)_y = f_{xy}, \quad \text{and} \quad \frac{\partial^2 f}{\partial x^2} = f_{xx}.$$

As you can see, the order of the variables is reversed when we use the notation f_{xy} compared with

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right).$$

14.10 How Smooth is a Function?

We usually use the term *smooth* to indicate that a function is at least continuous, and usually continuously differentiable. Generally speaking, what is meant by smooth depends on context. When we need to be precise, we can characterize the degree of smoothness by asking how many derivatives (or continuous derivatives the function has).

$\mathcal{C}^k$ Functions. The notation $\mathcal{C}^k$ is usually applied to functions defined on an open set U, thus $\mathcal{C}^k(U)$. It denotes the set of functions that are k-times continuously differentiable on U. A function $f \in \mathcal{C}^k(U)$ if and only if all of the first k partial derivatives exist and are continuous on U. When all of the partial derivatives of f exist and are continuous on U, we write $f \in \mathcal{C}^\infty$. A vector-valued function $\mathbf{f}\colon U \to \mathbb{R}^m$ is $\mathcal{C}^k(U)$ if and only if each component function is $\mathcal{C}^k(U)$. In such cases, we sometimes write $\mathcal{C}^k(U; \mathbb{R}^m)$.

14.11 Partial Derivatives and Utility

In consumer theory, partial derivatives can be used to compute marginal utilities and marginal rates of substitution.

► Example 14.11.1: Utility Functions. Suppose we have a Cobb-Douglas utility function[9]

$$u(x, y) = x^{\alpha} y^{1-\alpha}$$

defined on $\mathbb{R}^2_{++}$ where $0 < \alpha < 1$. The partial derivatives are the marginal utilities

$$MU_x = \frac{\partial u}{\partial x} = \alpha x^{\alpha-1} y^{1-\alpha} \quad \text{and} \quad MU_y = \frac{\partial u}{\partial y} = (1-\alpha) x^{\alpha} y^{-\alpha}.$$

We can use the marginal utilities to construct the *marginal rate of substitution*:

$$MRS_{xy} = \frac{MU_x}{MU_y} = \frac{\alpha}{1-\alpha} \frac{y}{x}.$$

◄

[9] The American economists Charles Wiggins Cobb (1875–1949) and Paul Douglas (1892–1976) introduced the Cobb-Douglas functions for production in 1928. Their first relevant paper was Charles W. Cobb and Paul H. Douglas (1928), A theory of production, *Amer. Econ. Rev.* **18**, no. 1. Supplement, Papers and Proceedings, 139–165. These functions are also often used for utility. Cobb worked in both mathematics and economics. Douglas later became a US Senator from Illinois from 1949–1967 where he supported both fiscal restraint and civil rights. Martin Luther King, Jr. called him the "greatest of all the Senators."

14.12 Partial Derivatives and Production

Partial derivatives are also used in producer theory. For one, they can be used to express marginal products and the marginal rate of technical substitution.

► Example 14.12.1: Production Functions. If

$$f(x_1, x_2) = \sqrt{x_1} + \sqrt{x_2}$$

is a production function on $\mathbb{R}^2_{++}$, the partial derivatives are the marginal products

$$MP_1 = \frac{\partial f}{\partial x_1} = \frac{1}{2\sqrt{x_1}} \quad \text{and} \quad MP_2 = \frac{\partial f}{\partial x_2} = \frac{1}{2\sqrt{x_2}}.$$

Then the *marginal rate of technical substitution* is the ratio of the marginal products:

$$MRTS_{12} = \frac{MP_1}{MP_2} = \sqrt{\frac{x_2}{x_1}}.$$

◄

14.13 Approximation via Partial Derivatives

We can also use partial derivatives to approximate functions. The key is that if we have a small change in x, Δx, the function will change by approximately

$$\Delta f \approx \frac{\partial f}{\partial x}\Delta x.$$

This type of approximation using the derivative is known as *Newton's Method*.

Here's an example to show how it works.

► **Example 14.13.1: Approximation.** Take the Cobb-Douglas production function $F(K, L) = 6K^{1/3}L^{2/3}$. Suppose $K = 8000$ and $L = 2744$, yielding $F(8000, 2744) = 23,520$. Now suppose K increases by $\Delta K = 10$. We estimate the effect on output using $\partial f/\partial K$.

$$\frac{\partial f}{\partial K}(8000, 2744) = 2K^{-2/3}L^{2/3} = 2(8000)^{-2/3}(2744)^{2/3} = .98$$

so we expect output to increase by about

$$\frac{\partial f}{\partial K}(8000, 2744) \times \Delta K = 0.98 \times 10 = 9.8.$$

to $23,520 + 9.8 = 23,529.8$. The actual value is $f(8010, 2744) \approx 23,529.796$, so you can see it is a pretty good approximation, at least for relatively small changes in K or L. ◄

14.14 Fréchet Differentiable Functions 10/18/22

New Homework: Problems 14.2, 14.4, 14.8, and 14.28 are due on **Tuesday, October 25**.

The next step is to consider whether a vector-valued function is differentiable. For this, we use the Fréchet derivative.

The definition of the Fréchet derivative may look a little strange if you have never seen it before. However, we will soon see that it really is a generalization of the ordinary derivative. The definition focuses on the linear approximation aspect of the derivative.

Fréchet Derivative. Suppose $\mathbf{f}\colon U \to \mathbb{R}^m$ where U is an open subset of $\mathbb{R}^k$ The function $\mathbf{f}$ is *Fréchet differentiable* at $\mathbf{x}_0 \in U$ if there is a linear function $L\colon \mathbb{R}^k \to \mathbb{R}^m$ such that

$$\lim_{\|\mathbf{h}\|\to 0} \frac{\mathbf{f}(\mathbf{x}_0 + \mathbf{h}) - \mathbf{f}(\mathbf{x}_0) - L(\mathbf{h})}{\|\mathbf{h}\|} = 0.$$

In that case L is the *Fréchet derivative* at $\mathbf{x}_0$. As a linear operator from $\mathbb{R}^k$ to $\mathbb{R}^m$, L can be written as a $m \times k$ matrix.

If $\mathbf{f}$ is Fréchet differentiable at every $\mathbf{x}_0 \in U$, we say that $\mathbf{f}$ is Fréchet differentiable on U. The Fréchet derivative at $\mathbf{x}_0$ is denoted $D\mathbf{f}(\mathbf{x}_0)$, or $D\mathbf{f}|_{\mathbf{x}_0}$. If we need to to specify which variables we are using to differentiate, we will use subscripts on the D operator. Thus $D_x f(x_0, y_0)$ or $D_y f(x_0, y_0)$.

14.15 Derivative of the Dot Product

To see how the Fréchet derivative works, we take the derivative of the dot product $\mathbf{p}\cdot\mathbf{x}$ with respect to $\mathbf{x}$. We obtain the covector $\mathbf{p}^{\mathbf{T}}$. This makes sense as the derivative is a linear functional of the vector $\mathbf{x}$, which means it must be a covector, and should be represented as a $1 \times k$ matrix, $\mathbf{p}^{\mathbf{T}}$.

▶ Example 14.15.1: Dot Product. An easy example of a Fréchet derivative is $f(\mathbf{x}) = \mathbf{p}\cdot\mathbf{x} = \sum_{i=1}^{k} p_i x_i$. Here $f\colon \mathbb{R}^k \to \mathbb{R}$. Since it is a linear function, it is its own linear approximation.

If we set $L(\mathbf{h}) = f(\mathbf{h}) = \mathbf{p}\cdot\mathbf{h}$, we find that

$$f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - L(\mathbf{h}) = \mathbf{p}\cdot(\mathbf{x}_0 + \mathbf{h}) - \mathbf{p}\cdot\mathbf{x}_0 - \mathbf{p}\cdot\mathbf{h} = 0.$$

We divide by $\|\mathbf{h}\|$ and take the limit. The limit is of course zero, so the linear function $L\colon \mathbb{R}^k \to \mathbb{R}$ is the Fréchet derivative. As a linear functional, it is represented by the **row vector** $\mathbf{p}^{\mathbf{T}}$. In other words,

$$D_{\mathbf{x}}(\mathbf{p}\cdot\mathbf{x}) = \mathbf{p}^{\mathbf{T}} = (p_1, \ldots, p_k)$$

at every $\mathbf{x}_0 \in \mathbb{R}^k$. ◀

It's may be easier to understand this derivative if you recall that the dot product can be written as a matrix product, $\mathbf{p}\cdot\mathbf{x} = \mathbf{p}^{\mathbf{T}}\mathbf{x}$, where having $\mathbf{p}^{\mathbf{T}}$ as derivative makes perfect sense.

Due to symmetry of the dot product, it is also the case that

$$D_{\mathbf{p}}(\mathbf{p}\cdot\mathbf{x}) = \mathbf{x}^{\mathbf{T}} = (x_1, \ldots, x_k)$$

at every $\mathbf{p}_0 \in \mathbb{R}^k$.

14.16 Fréchet and Ordinary Derivatives

Suppose $k = m = 1$, so $f\colon \mathbb{R} \to \mathbb{R}$. In this case the Fréchet derivative is the same as the ordinary derivative. One way to see that is to rewrite the definition of the ordinary derivative.

$$\begin{aligned} f'(x_0) &= \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h} \\ 0 &= \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0)}{h} - f'(x_0) \\ &= \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0) - hf'(x_0)}{h} \\ &= \lim_{h\to 0} \frac{f(x_0 + h) - f(x_0) - L(h)}{|h|} \end{aligned}$$

where the $|h|$ is ok since flipping the sign of zero leaves it as zero. Now $L(h) = f'(x_0)h$ is the required linear function of h.

14.17 Matrix Form of the Derivative — Column Vectors

We're about ready to write down the general form of the Fréchet derivative for a $\mathcal{C}^1$ function.

When dealing with derivatives of vector functions from $\mathbb{R}^k$ to $\mathbb{R}^m$ where $m > 1$, we will have to be somewhat pedantic about how the vectors are written. Writing them properly allows us to express certain relations as matrix products. Write them wrongly, and you get nonsense. Although we sometimes write a vector $\mathbf{x} \in \mathbb{R}^m$ in casual fashion as $(x_1, \ldots, x_m)$, here it needs to be written as a column

$$\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}.$$

When dealing with derivatives, we must use the formal column form. If $\mathbf{f}\colon \mathbb{R}^k \to \mathbb{R}^m$ is a vector function, we must write it as

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}$$

The derivative of $\mathbf{f}$ must be a linear function from $\mathbb{R}^k \to \mathbb{R}^m$. That means it can be represented by an $m \times k$ matrix with terms

$$\big[D\mathbf{f}(\mathbf{x}_0)\big]_{ij} = \frac{\partial f_i}{\partial x_j}(\mathbf{x}_0).$$

14.18 Matrix Form of the Derivative — The Jacobian

Then we write the derivative of

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{pmatrix}$$

as

$$D\mathbf{f}(\mathbf{x}_0) = \begin{pmatrix} \partial f_1/\partial x_1 & \partial f_1/\partial x_2 & \cdots & \partial f_1/\partial x_k \\ \partial f_2/\partial x_1 & \partial f_2/\partial x_2 & \cdots & \partial f_2/\partial x_k \\ \vdots & \vdots & \ddots & \vdots \\ \partial f_m/\partial x_1 & \partial f_m/\partial x_2 & \cdots & \partial f_m/\partial x_k \end{pmatrix}$$

where all of the partial derivatives are evaluated at $\mathbf{x}_0$. The derivatives expand the original function vector to the right to make a matrix. The value of the linear mapping at $\mathbf{x}_0$ is then

$$[L(\mathbf{h})]_i = \sum_{j=1}^{k} \left[\frac{\partial f_i}{\partial x_j}(\mathbf{x}_0) \right] h_j.$$

The derivative matrix $D\mathbf{f}$ is sometimes called the *Jacobian derivative*, *Jacobian matrix*, or just plain *Jacobian*.[10] The linear function $L(\mathbf{h})$ that the Jacobian represents is formed by multiplying the Jacobian matrix by a vector $\mathbf{h} \in \mathbb{R}^k$, obtaining a vector in $\mathbb{R}^m$.

[10] The Jacobian is named after Carl Gustav Jacobi. There's also a Jacobian determinant which is used when changing coordinates inside integrals. It too is sometimes just called the Jacobian.

14.19 The Derivative: An Example

Consider the function $\mathbf{f}\colon \mathbb{R}^3 \to \mathbb{R}^2$.

$$\mathbf{f}(\mathbf{x}) = \begin{pmatrix} x_1x_2^2 + x_1x_3 \\ x_2^2 + x_1^2x_3 + x_2x_3 \end{pmatrix}.$$

The derivative is a linear mapping from $\mathbb{R}^3 \to \mathbb{R}^2$, and is represented by a 2×3 matrix.

$$D\mathbf{f}(\mathbf{x}) = \begin{pmatrix} x_2^2 + x_3 & 2x_1x_2 & x_1 \\ 2x_1x_3 & 2x_2 + x_3 & x_1^2 + x_2 \end{pmatrix}$$

We evaluate the derivative at $\mathbf{x}_0 = (1, 1, 2)^{\mathsf{T}}$, obtaining

$$D\mathbf{f}\begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 3 & 1 & 1 \\ 4 & 4 & 2 \end{pmatrix}.$$

Finally, the linear function $L\colon \mathbb{R}^3 \to \mathbb{R}^2$ is

$$L(\mathbf{h}) = \left[D\mathbf{f}|_{x_0}\right]\mathbf{h} = \begin{pmatrix} 3 & 1 & 1 \\ 4 & 4 & 2 \end{pmatrix}\begin{pmatrix} h_1 \\ h_2 \\ h_3 \end{pmatrix} = \begin{pmatrix} 3h_1 + h_2 + h_3 \\ 4h_1 + 4h_2 + 3h_3 \end{pmatrix}.$$

14.20 Derivatives of Row Vectors

Should we have a function $\mathbf{f}\colon \mathbb{R}^k \to \mathbb{R}^m$ that starts as a row vector, $\big(f_1(\mathbf{x}), \dots, f(\mathbf{x}_m)\big)$ we will take the derivative to be $k \times m$ matrix $D\mathbf{f} = (D(\mathbf{f}^T))^T$. In other words,

$$D\mathbf{f} = \begin{pmatrix} \partial f_1/\partial x_1 & \cdots & \partial f_m/\partial x_1 \\ \partial f_1/\partial x_2 & \cdots & \partial f_m/\partial x_2 \\ \vdots & \ddots & \vdots \\ \partial f_1/\partial x_k & \cdots & \partial f_m/\partial x_k \end{pmatrix}$$

where the derivatives expand the matrix downwards.

In such a case, the linear mapping $L(\mathbf{h})$ is not formed by post-multiplying $D\mathbf{f}$ by $\mathbf{h}$, but by **pre-multiplying** by $\mathbf{h}^T$. To see that it makes sense, we do the math!

$$L(\mathbf{h}) = \mathbf{h}^T D\mathbf{f} = \left(\sum_{j=1}^{k} \frac{\partial f_1}{\partial x_j} h_j, \sum_{j=1}^{k} \frac{\partial f_2}{\partial x_j} h_j, \dots, \sum_{j=1}^{k} \frac{\partial f_m}{\partial x_j} h_j \right),$$

so that $L(\mathbf{h})$ is a covector, as desired.

When we consider second derivatives of functions from $\mathbb{R}^m \to \mathbb{R}$, the first derivative will be a row vector, and we can use the same method as above to take its derivative that will then be represented by an $m \times m$ matrix. It will define a bilinear functional. One of the vectors will be used to post-multiply $D^2 f$, the other will be transposed so it can pre-multiply $D^2 f$. In combination, this gives us a bilinear form.

14.21 Fréchet Derivatives and Approximation

Examining the definition, we see that the Fréchet derivative $Df(\mathbf{x}_0)$ defines a linear approximation to the function $\mathbf{f}$ by

$$\mathbf{f}(\mathbf{x}_0 + \mathbf{h}) \approx \mathbf{f}(\mathbf{x}_0) + L(\mathbf{h}) \tag{14.21.1}$$

where $L = D\mathbf{f}(\mathbf{x}_0)$.

To understand the approximation a bit better, we define the *remainder* R by

$$R(\mathbf{x}_0, \mathbf{h}) = f(\mathbf{x}_0 + \mathbf{h}) - f(\mathbf{x}_0) - L(\mathbf{h}).$$

The remainder is the difference between the function f and its linear approximation $f(\mathbf{x}_0) + L(\mathbf{h})$. By the definition of the derivative,

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{R(\mathbf{x}_0, \mathbf{h})}{\|\mathbf{h}\|} = 0.$$

This can also be written

$$\frac{R(\mathbf{x}_0, \mathbf{h})}{\|\mathbf{h}\|} = o\big(\|\mathbf{h}\|\big) \text{ at } \mathbf{0}.$$

meaning that it converges to zero as $\|\mathbf{h}\| \to 0$. We will discuss the $o(\cdot)$ notation later, in section 30.18.

We conclude that the linear approximation is fairly accurate near $\mathbf{x}_0$, and gets more accurate the closer we are to $\mathbf{x}_0$. In fact, the error term, the remainder, converges to zero enough faster than $\|\mathbf{h}\|$ that their ratio also converges to zero.

14.22 Approximation with Deltas

To see how the the approximation works, let's rewrite in terms of changes in each x_i. Because the derivative of f is a linear approximation to f, if we feed $Df(\mathbf{x}_0)$ a vector such as

$$\Delta\mathbf{x} = \begin{pmatrix} \Delta x_1 \\ \vdots \\ \Delta x_m \end{pmatrix} = \sum_i \Delta x_i \, \mathbf{e}_i,$$

we obtain

$$\begin{aligned} \Delta f &\approx Df_{\mathbf{x}_0} \Delta\mathbf{x} \\ &= \sum_{i=1}^{m} \left[\frac{\partial f}{\partial x_i}(\mathbf{x}_0) \, dx_i \left(\sum_{j=1}^{m} \Delta x_j \, \mathbf{e}_j \right) \right] \\ &= \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \, \delta_{ij} \, \Delta x_j \\ &= \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(\mathbf{x}_0) \, \Delta x_i \end{aligned}$$

which is a linear approximation to the change in f as $\mathbf{x}_0$ is replaced by $\mathbf{x}_0 + \Delta x$ (see also equation (14.21.1)).

14.23 Marginal Rate of Substitution: Interpretation

We can use our knowledge of approximation to interpret the marginal rate of substitution. Suppose we start at $\mathbf{x}_0$. Take a utility function u and consider the indifference curve through $\mathbf{x}_0$, $\{\mathbf{x} : u(\mathbf{x}) = u(\mathbf{x}_0)\}$. Let's make a small change in x_i, holding everything else constant. This moves us off the $u(\mathbf{x})$ indifference curve. Now change x_j to return us the the indifference curve. Since indifference curves slope downward (assuming preferences are monotonic), there is a trade-off between goods i and j. They must have opposite signs.

We can now use the linear approximation. Define $\mathbf{h}$ by $h_i = \Delta x_i$, $h_j = \Delta x_j$, and $h_k = 0$ for $k \neq i, j$, so $\mathbf{h} = \Delta x_i \boldsymbol{e}_i + \Delta x_j \boldsymbol{e}_j$. Then

$$\Delta u \approx Df(\mathbf{x}_0)\mathbf{h} = \frac{\partial u}{\partial x_i}\Delta x_i + \frac{\partial u}{\partial x_j}\Delta x_j = MU_i\,\Delta x_i + MU_j\,\Delta x_j$$

It follows that $MU_i\,\Delta x_i + MU_j\,\Delta x_j \approx 0$, so

$$MRS_{ij} = \frac{MU_i}{MU_j} \approx -\frac{\Delta x_j}{\Delta x_i},$$

regardless of the number of dimensions we started with.

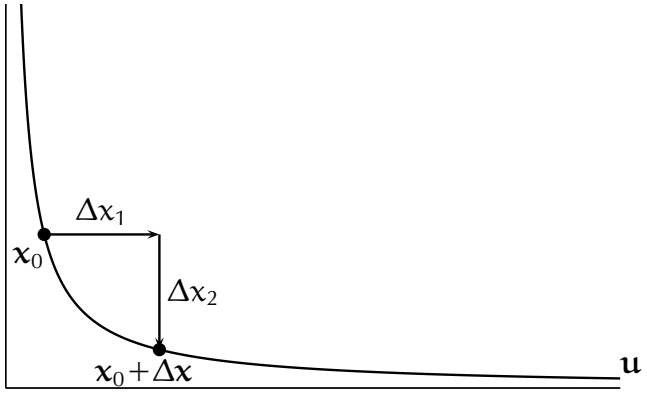

Figure 14.23.1: In other words, if we plot a slice of the indifference curve in x_i-x_j space, with x_i on the horizontal axis, the marginal rate of substitution approximates the negative slope of the secant through the points $\big(\mathbf{x}_0 + \Delta x, u(\mathbf{x}_0 + \Delta x)\big)$ and $\big(\mathbf{x}_0, u(\mathbf{x}_0)\big)$. Of course, if we let $\Delta x_i, \Delta x_j \to 0$, we obtain the negative slope of the tangent at $u(\mathbf{x}_0)$.

14.24 Total Derivatives and Linear Functionals

When $f\colon \mathbb{R}^k \to \mathbb{R}$, the derivative takes the simpler form

$$Df = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_k}\right) \tag{14.24.2}$$

Sometimes the derivative is written in what appears to be a quite different manner and called the *total derivative*:

$$df = \frac{\partial f}{\partial x_1}\,dx_1 + \frac{\partial f}{\partial x_2}\,dx_2 + \cdots + \frac{\partial f}{\partial x_k}\,dx_k. \tag{14.24.3}$$

They're supposed to both be the derivative. What is going on here?

Recall that the derivative of f is a linear function from $\mathbb{R}^k$ to $\mathbb{R}$. As such, it can properly be written as a covector, a row vector. That is what we have done in equation (14.24.2). But what about equation (14.24.3)?

The obvious interpretation is that we mean that dx_i must the i^{th} basis element $\boldsymbol{e}_i^* = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in the i^{th} position.

But why is $\boldsymbol{e}_i^*$ written dx_i? Consider the i^{th} coordinate function $x_i(\boldsymbol{x}) = x_i$, which maps $\mathbb{R}^k$ to $\mathbb{R}$. Its derivative, here written as dx_i instead of Dx_i, must be the linear functional $\boldsymbol{e}_i^*$.

The linear functional $\boldsymbol{e}_i^*$ maps $\mathbb{R}^k$ to $\mathbb{R}$ such that $dx_i(\boldsymbol{e}_j) = \boldsymbol{e}_i^*(\boldsymbol{e}_j)$ is given by the Kronecker delta δ_{ij}. Either way we write it, the differentials of the such functions form a basis for row vectors.

14.25 One-Forms

The total differential in equation (14.24.3) can be regarded as a simple type of *differential form*. A functional linear combination of the basis vectors dx_i is a called a *differential* 1*-form*. It is important that there are no products of the dx_i terms and each term has a single dx_i. Real-valued functions are sometimes referred to a 0-forms.

One-forms are a linear combination of the basis vectors of $(\mathbb{R}^m)^*$ where the coefficients are functions. The functions are the derivatives $\partial f/\partial x_i$. One-forms produce a linear functional on $\mathbb{R}^m$ for every $\mathbf{x}$ In other words, a *differential* 1*-form* is a function whose values are linear functionals.

14.26 Derivatives and Differential Forms

Both the derivative Df and the one-form df version seem to give the same results. So why have two versions?

Among other things, differential forms are used in integration. Products there appear in higher derivatives (k-forms) and correspond to area or volume elements, etc. The appropriate product here is something called the **exterior** or **wedge product**, which is both **alternating** and **multilinear**, like determinants.[11]

We also use products for higher Fréchet derivatives, $D^k f$. These are k-multilinear functions, k-tensors. They are **not** alternating, and are not to be confused with the k-forms used in multi-dimensional integration. The k-forms are alternating, the k^{th}-order derivatives are not. Quite the contrary, terms such as $\partial^2 f / \partial x_i^2$ appear and at times are important. They would be zero if tensors were alternating.

[11] The differential forms and their wedge products form a Grassmann (exterior) algebra. Grassmann algebras were developed in 1844 by Hermann Günther Grassmann. The exterior products were intended to extend line elements into areas, area elements into volumes, etc.

During his life, Grassmann (1809–1877) was better known as a linguist than a mathematician. Mathematicians gradually noticed his work. By the time the concepts of vector spaces and linear functions formalized by Giuseppe Peano (1858–1932) in 1888, Grassmann algebras started to make more sense to other mathematicians, and were eventually turned into a powerful tool of analysis around 1900. Henri Poincaré (1854–1912), Élie Cartan (1869–1951), and Gaston Darboux (1842–1917) played important roles in this.

14.27 Curves in $\mathbb{R}^m$

A *curve* in $\mathbb{R}^m$ is an m-tuple of continuous functions

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_m(t) \end{pmatrix}$$

where each $x_i : I \to \mathbb{R}$ where I is a open subset of $\mathbb{R}$. The functions are the *coordinate functions* of the curve $\mathbf{x}$ and t is a parameter describing the curve. Curves are allowed to cross or repeat themselves.

One of the simplest curves is a straight line. Recall that straight lines can be written in parametric form as

$$\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{x}_1 \tag{10.56.5.}$$

The curve $\mathbf{x}(t) = (\cos t, \sin t)^T$ repeatedly traces out a circle of radius one centered at the origin. By adding a coordinate, we obtain $\mathbf{y}(t) = (\cos t, \sin t, t)^T$, which is a right-handed helix in $\mathbb{R}^3$.

14.28 Tangent Vectors

If $\mathbf{x}(t)$ is a differentiable curve, we can define the *tangent vector* as

$$\mathbf{x}'(t) = \begin{pmatrix} x_1'(t) \\ x_2'(t) \\ \vdots \\ x_m'(t) \end{pmatrix}.$$

The tangent vector describes the instantaneous rate of change of the curve, both direction and magnitude.

When $\mathbf{x}(t)$ is the path of an actual object in motion and t is time, $\mathbf{x}'(t)$ is the velocity at any time t and $\|\mathbf{x}'(t)\|$ is the speed. The second derivative

$$\mathbf{x}''(t) = \begin{pmatrix} x_1''(t) \\ x_2''(t) \\ \vdots \\ x_m''(t) \end{pmatrix}.$$

is the acceleration at time t. Here $x_i''(t)$ denotes the ordinary second derivative

$$x_i''(t) = \frac{d^2 x_i}{dt^2}(t).$$

14.29 Examples of Curves

Our first example is a straight line.

► **Example 14.29.1: Straight Lines.** The straight line $\mathbf{x}(t) = \mathbf{x}_0 + t\mathbf{x}_1$ has tangent $\mathbf{x}'(t) = \mathbf{x}_1$, the direction of the line. In this form, there is no acceleration along the line, $\mathbf{x}''(t) = \mathbf{0}$. The same line can be traced out in other ways, for example the curve $\mathbf{x}(t) = \mathbf{x}_0 + t^3\mathbf{x}_1$ visits all of the points on that straight line, but does it in a different fashion, with tangent $3t^2\mathbf{x}_1$. Except at $t = 0$, it points in the same direction as $\mathbf{x}_1$, but has a different magnitude, reflecting the variously slower and faster motion along the line. Moreover the acceleration is not zero, but $\mathbf{x}''(t) = 6t\mathbf{x}_1$. ◄

The second example circle around and around the unit circle about zero.

► **Example 14.29.2: Perpetual Circle.** The perpetual circle is defined by $\mathbf{x}(t) = (\cos t, \sin t)$. It continually retraces the circle of radius 1 about the origin. Its tangent vector is $\mathbf{x}'(t) = (-\sin t, \cos t)$. You'll notice that for circular motion, the tangent is orthogonal to the curve, $\mathbf{x}'(t)\cdot\mathbf{x}(t) = 0$. The acceleration $\mathbf{x}''(t) = (-\cos t, -\sin t) = -\mathbf{x}(t)$ always points toward the origin. ◄

Our third example is a right-handed helix, meaning the it winds counter-clockwise.

► **Example 14.29.3: Right-handed Helix.** The helix $\mathbf{y}(t) = (\cos t, \sin t, t)$ has tangent $\mathbf{y}'(t) = (-\sin t, \cos t, 1)$. It is not perpendicular to $\mathbf{x}(t)$ due to the x_3 component. The acceleration is $\mathbf{x}''(t) = (-\sin t, -\cos t, 0)$ and points toward the x_3-axis. The fact that the third component of acceleration is zero is due to the constant motion along the x_3-axis. The other components reflect the circular motion about it. ◄

14.30 Regular Curves

We will often require that curves be sufficiently smooth. Such curves are called *regular*.

Regular Curve. A curve $\mathbf{x}(t) = \big(x_1(t), \ldots, x_m(t)\big)^T$ in $\mathbb{R}^m$ is *regular* if each x_i is continuously differentiable in t and $\mathbf{x}'(t) \neq \mathbf{0}$ for all t.

Examples (14.29.1)–(14.29.3) are all regular.

14.31 Where Regularity Fails

Regularity rules out sudden changes of direction.

► **Example 14.31.1: Curve with a Cusp.** Let

$$\mathbf{x}(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix} = \begin{pmatrix} t^3 \\ t^2 \end{pmatrix} \quad \text{so} \quad \mathbf{x}'(t) = \begin{pmatrix} 3t^2 \\ 2t \end{pmatrix}.$$

This curve is not regular at 0 because $\mathbf{x}(0) = (0, 0)$. It makes a sudden change of direction there.

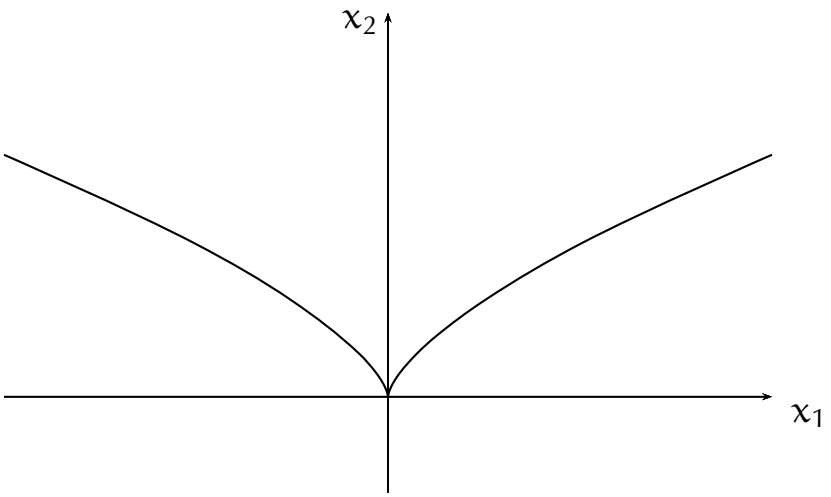

Figure 14.31.2: The point where the curve makes a sudden change of direction is called a *cusp*.

It's not possible to re-parameterize this curve to make it regular. The tangent changes from going straight down to straight up. Reversing the tangent requires that it be zero at the reversal point, regardless of how the parameter is applied. One way to see this is that if $\mathbf{x}'(t)$ is $\mathcal{C}^1$, so is $x_2'(t)$. For $t < 0$, $x_2'(t)$ is negative, and for $t > 0$, $x_2'(t)$ is positive. Continuity requires that $x_2'(0) = 0$. The only way $\mathbf{x}$ can be regular at zero is if $x_1'(0) \neq 0$. But in that case the tangent could not be straight up or down at $\mathbf{0}$. ◄

14.32 Functions, Curves, and Derivatives

It is sometimes useful to evaluate the derivative of a function f defined along a curve. If $\mathbf{x}(t)$ is a regular curve in $\mathbb{R}^m$ and $f\colon \mathbb{R}^m \to \mathbb{R}$, we can define $g(t) = f \circ \mathbf{x}(t) = f\big(\mathbf{x}(t)\big)$. We can take the derivative of g in the following fashion, based on the chain rule:

$$\begin{aligned} g'(t) &= \frac{\partial f}{\partial x_1}\big(\mathbf{x}(t)\big)\, x_1'(t) + \frac{\partial f}{\partial x_2}\big(\mathbf{x}(t)\big)\, x_2'(t) \\ &\qquad + \cdots + \frac{\partial f}{\partial x_m}\big(\mathbf{x}(t)\big)\, x_m'(t) \\ &= Df|_{\mathbf{x}(t)}\, \mathbf{x}'(t). \end{aligned}$$

where $Df|_{\mathbf{x}(t)}$ indicates that Df is evaluated at $\mathbf{x}(t)$.

In the final formula, keep in mind that Df is a $1 \times m$ row vector and $\mathbf{x}'$ is an $m \times 1$ column vector, so the matrix product is a number. Writing the vectors ($\mathbf{x}$) and covectors (Df) in the right way ensures everything lines up properly in the product.

As for the formula above, we state the relevant theorem, without proof.

Chain Rule I. *Let $\mathbf{x}(t) = \big(x_1(t), \ldots, x_m(t)\big)$ be a $\mathcal{C}^1$ curve on an interval about t_0 and f a $\mathcal{C}^1$ function defined on a ball in $\mathbb{R}^m$ about $\mathbf{x}(t_0)$. Then $g(t) = f\big(\mathbf{x}(t)\big)$ is a $\mathcal{C}^1$ function on an interval about t_0 and*

$$\begin{aligned} \frac{dg}{dt}(t_0) &= \frac{\partial f}{\partial x_1}\big(\mathbf{x}(t_0)\big)\, x_1'(t_0) + \cdots + \frac{\partial f}{\partial x_m}\big(\mathbf{x}(t_0)\big)\, x_m'(t_0) \\ &= \big[Df|_{\mathbf{x}(t_0)}\big]\, \mathbf{x}'(t_0). \end{aligned}$$

14.33 Directional Derivatives

One way to think about derivatives in a particular direction is to consider a curve that goes in that direction. We want a curve with $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ and $\boldsymbol{x}'(t_0) = \boldsymbol{v}$. There are many such curves, one is the line through $\boldsymbol{x}_0$ in the direction $\boldsymbol{v}$, given by $\boldsymbol{x}(t) = \boldsymbol{x}_0 + (t - t_0)\boldsymbol{v}$.

If $f\colon \mathbb{R}^m \to \mathbb{R}$, we can consider the composite function $g(t) = f\big(\boldsymbol{x}(t)\big)$ and take its derivative

$$\frac{df}{dt}(t_0) = \Big[Df|_{\boldsymbol{x}_0}\Big]\,\boldsymbol{x}'(t_0) = \Big[Df|_{\boldsymbol{x}_0}\Big]\,\boldsymbol{v}.$$

In fact, Chain Rule I ensures that **any curve** with $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ and $\boldsymbol{x}'(t_0) = \boldsymbol{v}$ will have the **same derivative**. We refer to this as the *directional derivative of* f *in the direction* $\boldsymbol{v}$. Two notations sometimes used for the directional derivative are

$$\frac{\partial f}{\partial \boldsymbol{v}}(\boldsymbol{x}_0) \quad \text{and} \quad D_{\boldsymbol{v}} f(\boldsymbol{x}_0)$$

We prefer the former notation as $D_{\boldsymbol{v}} f$ conflicts with our notation for the Fréchet derivative.

14.34 The Gradient Vector

It is sometimes useful to regard the derivative of a real-valued function f on $U \subset \mathbb{R}^m$ as a vector rather than a covector. That vector is called the *gradient*, and is written

$$\nabla f(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}_0) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}_0) \\ \vdots \\ \frac{\partial f}{\partial x_m}(\mathbf{x}_0) \end{pmatrix}$$

where $\mathbf{x}_0 \in U$. Of course, $\nabla f(\mathbf{x}_0) = Df(\mathbf{x}_0)^T$. We can then write the directional derivative as

$$\frac{\partial f}{\partial \boldsymbol{v}}(\mathbf{x}_0) = \nabla f(\mathbf{x}_0) \cdot \boldsymbol{v}.$$

14.35 The Chain Rule

The following theorem is called "Chain Rule IV" in Simon and Blume. We'll just call it *the* Chain Rule.

The Chain Rule. *Let U and V be open subsets of $\mathbb{R}^k$ and $\mathbb{R}^\ell$, respectively. Suppose $\mathbf{f}\colon U \to \mathbb{R}^\ell$ and $\mathbf{g}\colon V \to \mathbb{R}^m$ are $\mathcal{C}^1$ functions, and that $\mathbf{x}_0 \in U$ and $\mathbf{y}_0 = f(\mathbf{x}_0) \in V$. We will write $\mathbf{y} = \mathbf{f}(\mathbf{x})$ and $\mathbf{g}(\mathbf{y})$.*

Then $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ is also $\mathcal{C}^1$ on an open ball about $\mathbf{x}_0$ with

$$D_{\mathbf{x}}\mathbf{h}(\mathbf{x}_0) = D_{\mathbf{y}}\mathbf{g}(\mathbf{y}_0) \times D_{\mathbf{x}}\mathbf{f}(\mathbf{x}_0)$$

In other words, the Jacobian derivative of $\mathbf{h}$ is the product of the Jacobian derivatives of $\mathbf{g}$ and $\mathbf{f}$.

The Chain Rule both asserts that the composite function $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ is differentiable and gives a formula for its derivative.

We know that $D\mathbf{h}$ is an $k \times \ell$ matrix, and $D\mathbf{f}$ is an $\ell \times m$ matrix, so $D\mathbf{h}$ is an $k \times m$ matrix. Unpacking the matrix product shows that

$$\frac{\partial h_i}{\partial x_j} = \sum_{k=1}^{m} \left(\frac{\partial g_i}{\partial y_k}\right)\left(\frac{\partial f_k}{\partial x_j}\right) \tag{14.35.4}$$

for all $i = 1, \ldots, k$ and $j = 1, \ldots, m$. Equation (14.35.4) can be written more fully as

$$\frac{\partial h_i}{\partial x_j}(\mathbf{x}_0) = \sum_{k=1}^{m} \left(\frac{\partial g_i}{\partial y_k}(\mathbf{y}_0)\right)\left(\frac{\partial f_k}{\partial x_j}(\mathbf{x}_0)\right),$$

again for all $i = 1, \ldots, k$ and $j = 1, \ldots, m$. Keep in mind that $\mathbf{y}_0 = \mathbf{f}(\mathbf{x}_0)$.

14.36 Chain Rule for Direct and Indirect Variables

The Chain Rule has many applications beyond the obvious ones. Some functions will use the same variables both directly and indirectly via another function.

Consider the function

$$\Phi(\mathbf{x}) = \mathbf{f}\big(\mathbf{x}, g(\mathbf{x})\big).$$

Here $\mathbf{x}$ appears both directly in Φ, and indirectly via the real-valued function g. If $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{f}$ takes $m+1$ arguments. Partition those arguments as $(\mathbf{x}, y)$. Define a function $\mathbf{h}\colon \mathbb{R}^m \to \mathbb{R}^{m+1}$ by

$$\mathbf{h}(\mathbf{x}) = \begin{pmatrix} \mathbf{x} \\ g(\mathbf{x}) \end{pmatrix} \quad \text{with} \quad D\mathbf{h}(\mathbf{x}) = \begin{pmatrix} \mathbf{I}_m \\ D_x g \end{pmatrix}$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix.

Then $\Phi(\mathbf{x}) = \mathbf{f}\big(\mathbf{h}(\mathbf{x})\big)$, so the Chain Rule tells us that

$$\begin{aligned} D_x\Phi &= D_{(x,y)}\mathbf{f} \times D_x\mathbf{h} \\ &= D_{(x,y)}\mathbf{f} \times \begin{pmatrix} \mathbf{I}_m \\ D_x g \end{pmatrix} \\ &= D_x\mathbf{f} + D_y\mathbf{f} \times D_x g. \end{aligned}$$

14.37 An Application of the Chain Rule to Integrals

One useful application of the chain rule is to integrals with variable limits.

► **Example 14.37.1: Chain Rule and Integrals.** Now consider let $g\colon \mathbb{R} \to \mathbb{R}$ and $h\colon \mathbb{R}^2 \to \mathbb{R}$ be $\mathcal{C}^1$ and define

$$\Phi(x) = \int_0^{g(x)} h(x,t)\,dt.$$

Now define

$$f(x,y) = \int_0^{y} h(x,t)\,dt$$

and apply the previous result to $\Phi(x) = f\big(x, g(x)\big)$. This calculation yields

$$\begin{aligned}\Phi'(x) &= \frac{d}{dx}\left(\int_0^{g(x)} h(x,t)\,dt\right)\\ &= h\big(x,g(x)\big)\,g'(x) + \int_0^{g(x)} \frac{\partial h}{\partial x}(x,t)\,dt.\end{aligned}$$

When $g(x) = x$, this formula reduces to

$$\frac{d}{dx}\left(\int_0^{x} h(x,t)\,dt\right) = h(x,x) + \int_0^{x} \frac{\partial h}{\partial x}(x,t)\,dt.$$

A similar method can be used on the case where both limits of integration are defined by functions. ◄

14.38 Second Derivatives: The Hessian

Second derivatives are a little more complex than first derivatives and we will start with the case where $f\colon \mathbb{R}^m \to \mathbb{R}$ because it is easier to see what is going on. In that case, the first-derivative is a row vector

$$\left(\frac{\partial f}{\partial x_1}(\mathbf{x}) \quad \dots \quad \frac{\partial f}{\partial x_m}(\mathbf{x}) \right).$$

This was discussed in section 14.20, where we used a modified matrix representation of such derivatives. We apply that method here.

For functions that are twice continuously differentiable, we write the matrix of second partial derivatives, the *Hessian matrix*, or *Hessian*, as[12]

$$D^2 f(\mathbf{x}) = \begin{pmatrix} \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_m} \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_m} \\ \vdots & \vdots & & \vdots \\ \dfrac{\partial^2 f(\mathbf{x})}{\partial x_m \partial x_1} & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_m \partial x_2} & \cdots & \dfrac{\partial^2 f(\mathbf{x})}{\partial x_m^2} \end{pmatrix}.$$

Where

$$\frac{\partial^2 f}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left(\frac{\partial f}{\partial x_\ell} \right).$$

[12] The Hessian matrix is due to Ludwig Hesse (1811–1874). He was a German mathematician who mainly worked on geometry and algebraic invariants. The hyperplane equation $\mathbf{p} \cdot \mathbf{x} = c$ that we use to describe hyperplanes has been called the Hessian normal form.

14.39 Another Notation for Hessians

This is one of those times when it is useful to employ a space-saving notation for partial derivatives. Define

$$f_i = \frac{\partial f}{\partial x_i}, \quad f_{ij} = (f_i)_j = \frac{\partial^2 f}{\partial x_j \, \partial x_i} = \frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right), \text{ etc.}$$

Notice the reversal of order between the two notations, ij versus ji. The alternate notation allows us to write the Hessian in the more compact form

$$D^2 f(\mathbf{x}) = [f_{ji}] = \begin{pmatrix} f_{11}(\mathbf{x}) & f_{21}(\mathbf{x}) & \cdots & f_{m1}(\mathbf{x}) \\ f_{12}(\mathbf{x}) & f_{22}(\mathbf{x}) & \cdots & f_{m2}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ f_{1m}(\mathbf{x}) & f_{2m}(\mathbf{x}) & \cdots & f_{mm}(\mathbf{x}) \end{pmatrix}.$$

When f is twice continuously differentiable, it does not matter which order is used for the second partial derivatives because

$$\frac{\partial^2 f}{\partial x_i \, \partial x_j} = \frac{\partial^2 f}{\partial x_j \, \partial x_i}.$$

The derivative is the same either way.[13] If f is not $\mathcal{C}^2$, the order in which we take partial derivatives may affect the result.

[13] In the economics literature, this result is often called Young's Theorem, although a host of mathematicians have worked on the problem, including Cauchy and Lagrange. My experience is that mathematics books usually don't name the theorem after anyone, although I have seen it called the Clairaut-Schwartz Theorem. Alexis Clairaut (1713–1765) published a proof in 1740 that does not meet modern standards of rigor. The first rigorous proof was due to Herman Schwartz in 1873. The most commonly used proof is that of Camille Jordan (1838–1922) published in 1883. E.W. Hobson (1856–1933) and W.H. Young (1863–1942) later proved the theorem under weaker conditions. The name Young's Theorem may derive from the economist R.G.D. Allen (1906–1983), who used that name for the theorem in his 1938 book, "Mathematical Analysis for Economists".

14.40 Interpreting The Hessian 10/20/22

So what exactly is the Hessian? We're representing it with a matrix, but how do we use that matrix?

Let $f\colon \mathbb{R}^m \to \mathbb{R}$. The derivative of f, Df maps $\mathbf{x}$ to the linear functional $Df(\mathbf{x})$. As a linear functional, the latter takes a vector $\mathbf{y} \in \mathbb{R}^m$ and produces a real number. We can write the values of this linear functional for $\mathbf{y} \in \mathbb{R}^m$ as the matrix product $[Df(\mathbf{x})]\mathbf{y}$. In this case Df is a covector (row vector)

The Hessian is a way of writing the derivative of the mapping $\mathbf{x} \mapsto D^2f(\mathbf{x})$. The second derivative at $\mathbf{x}$ is a linear function from $\mathbb{R}^m \to (\mathbb{R}^m)^*$. Our convention is to create the Hessian matrix by extending the derivatives downward.

We must feed the Hessian two vectors, the first will give us a linear functional (row vector), the second will use that linear functional to produce a number.

The Hessian matrix makes it possible to do this. The Hessian is an $m \times m$ matrix. We first feed it a vector $\mathbf{z}$ by multiplying on the left by $\mathbf{z}^T$, which is $1 \times m$. That gives us a $1 \times m$ matrix, $\mathbf{z}^T\left[D^2f(\mathbf{x})\right]$. It defines a linear functional on $\mathbb{R}^m$, a linear mapping from $\mathbb{R}^m$ to $\mathbb{R}$. We implement that by mutiplying on the right by $\mathbf{y} \in \mathbb{R}^m$. The result is a real number.

In other words, the Hessian defines a bilinear map $B\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ by the formula

$$B(\mathbf{z}, \mathbf{y}) = \mathbf{z}^T\left[D^2f(\mathbf{x})\right]\mathbf{y} \tag{14.40.5}$$

It takes pairs of vectors $(\mathbf{y}, \mathbf{z})$ and gives us a real number. We will expand this on the next slide.

14.41 The Hessian is a Symmetric Matrix

Writing equation (14.40.5) out, we obtain

$$\begin{aligned} B(z, y) &= z^T D^2 f(x)\, y \\ &= (z_1, \dots, z_m) \begin{pmatrix} f_{11}(x) & f_{21}(x) & \cdots & f_{m1}(x) \\ f_{12}(x) & f_{22}(x) & \cdots & f_{m2}(x) \\ \vdots & \vdots & & \vdots \\ f_{1m}(x) & f_{2m}(x) & \cdots & f_{mm}(x) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} \\ &= \sum_{i,j=1}^{m} f_{ji}(x)\, y_j z_i \\ &= \sum_{i,j=1}^{m} \frac{\partial}{\partial x_i}\left(\frac{\partial f}{\partial x_j}\right) y_j z_i. \end{aligned}$$

Since f is $\mathcal{C}^2$, $D^2 f(x)$ is a symmetric matrix. Then it doesn't really matter which vector is y and which is z. The sum is unchanged.

14.42 The Hessian is a Bilinear Form — A 2-Tensor

The important thing about the Hessian is that it defines a bilinear form. The mapping from $\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ given by

$$(\mathbf{y}, \mathbf{z}) \mapsto \mathbf{z}^T \left[D^2 f(\mathbf{x}) \right] \mathbf{y}$$

is bilinear, and is best thought of as a 2-tensor, a linear mapping from $\mathbb{R}^m \otimes \mathbb{R}^m$ to $\mathbb{R}$. As such we can write

$$D^2 f(\mathbf{x}) = \sum_{ij=1}^{m} f_{ji}(\mathbf{x})\, dx_i \otimes dx_j$$

when we can use the tensor product to write

$$B(\mathbf{z}, \mathbf{y}) = \left[D^2 f(\mathbf{x}) \right] \cdot (\mathbf{z} \otimes \mathbf{y}).$$

The dot product of the two matrices indicates we are multiplying the corresponding terms and adding, just like the ordinary dot product, but for matrices. This makes sense because linear functionals on $\mathbb{R}^m$ involve dot products. It shouldn't be surprising to find one here. We use the outer product to represent $\mathbf{z} \otimes \mathbf{y}$ as the matrix with $z_i y_j$ in row i, column j.

You may run across constructions where you vectorize the matrices. This is sometimes done in econometrics. Then taking the dot product of the resulting vectors would give you the same result.

14.43 Taylor's Formula: Preview

One important use of the Hessian is Taylor's formula, which we will prove shortly.[14]

First, a couple of definitions. In a vector space, the *line segment between* $\mathbf{x}$ *and* $\mathbf{y}$ is $\ell(\mathbf{x},\mathbf{y}) = \{(1-t)\mathbf{x} + t\mathbf{y} : 0 \leq t \leq 1\}$. A set S is *convex* if it contains $\ell(\mathbf{x},\mathbf{y})$ whenever $\mathbf{x},\mathbf{y} \in S$.

First Order Taylor's Formula. *Let* $f\colon U \to \mathbb{R}$ *be* $\mathcal{C}^2$ *on a convex open set* $U \subset \mathbb{R}^m$. *Then for every* $\mathbf{x},\mathbf{x}' \in U$, *there is a* $\mathbf{y}$ *on the line segment connecting* $\mathbf{x}$ *and* $\mathbf{x}'$ *such that*

$$f(\mathbf{x}') = f(\mathbf{x}) + Df(\mathbf{x})(\mathbf{x}'-\mathbf{x}) + \frac{1}{2}(\mathbf{x}'-\mathbf{x})^T\left[D^2f(\mathbf{y})\right](\mathbf{x}'-\mathbf{x}). \quad (14.43.6)$$

In Taylor's formula, not only is the Hessian symmetric, but we feed it the same vector on each side, so there is even more symmetry. If $B(\cdot,\cdot)$ is a bilinear form, then $Q(z) = B(z,z)$ is called a *quadratic form*. To see why, write it out. When B is the Hessian, we get

$$B(z,z) = \sum_{ij=1}^{m} f_{ij}(\mathbf{x})\, z_i z_j.$$

It is a purely second degree polynomial in the z_i's, hence the term quadratic. The Taylor approximation in equation (30.14.2) consists of a constant term, a linear term, and a quadratic term. For small changes in $\mathbf{x}$ (and sometimes large changes), it will better approximate f than the first derivative alone does.

There are higher Taylor expansions that involve third and higher degree terms, terms which are 3-tensors, 4-tensors, etc.

[14] Brook Taylor (1685–1731) was an English mathematician, best known for Taylor's Formula. He also studied optical phenomena including refraction and perspective.

14.44 Higher Derivatives as Tensors

Let $f\colon \mathbb{R}^m \to \mathbb{R}$ be k times continuously differentiable with $k > 2$. What do its higher derivatives, those past the Hessian, look like. Well, we know that $D^k f(\mathbf{x})$ must be a k-linear form. That means it can be written as linear map from $(\mathbb{R}^m)^{\otimes k} \to \mathbb{R}$, a k-tensor.

Although perhaps a bit tedious, it's not hard to figure out what the $D^k f(\mathbf{x})$ look like, even though they do not have convenient matrix representations.

$$\left[D^k f(\mathbf{x})\right](\mathbf{y}^1, \ldots, \mathbf{y}^k) = \sum_{i_1,\ldots,i_k=1}^{m} f_{i_1\cdots i_k}(\mathbf{x})\, y^k_{i_k} \cdots y^1_{i_1}$$

where y^i_j is the j^{th} component of $\mathbf{y}^i$. As written here, $y^j_{i_k}$ is the i_k component of $\mathbf{y}^j$. This means that $D^k f(\mathbf{x})$ can be written as the k-tensor

$$D^k f(\mathbf{x}) = \sum_{i_1,\ldots,i_k=1}^{m} f_{i_1\cdots i_k}(\mathbf{x})\, dx_{i_k} \otimes \cdots \otimes dx_{i_1}.$$

November 10, 2022