Lecture 4: Differential Calculus

1 Functions of one variable

1.1 Definition

Let $I \subseteq \mathbb{R}$ be an open interval and let $f : I \to \mathbb{R}$. If the input moves from x_0 to $x_0 + h$, then the input changes by h and the value of the function changes by

$$f(x_0 + h) - f(x_0).$$

For $h \neq 0$, the ratio

$$\frac{f(x_0 + h) - f(x_0)}{h} \tag{1}$$

is the *average rate of change* between the two points. Geometrically, it is the slope of the line that goes through $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$. The derivative is the limiting slope as $h \to 0$. Thus x_0 is the fixed point at which the derivative is evaluated, while $x_0 + h$ is the moving input.

Definition 1.1 (Derivative)**.** The function f is *differentiable at* x_0 if there is $a \in \mathbb{R}$ such that

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

The number a is the *derivative* of f at x_0, written $f'(x_0)$. If f is differentiable at every point of I, then f' denotes the *derivative function* $x \mapsto f'(x)$.

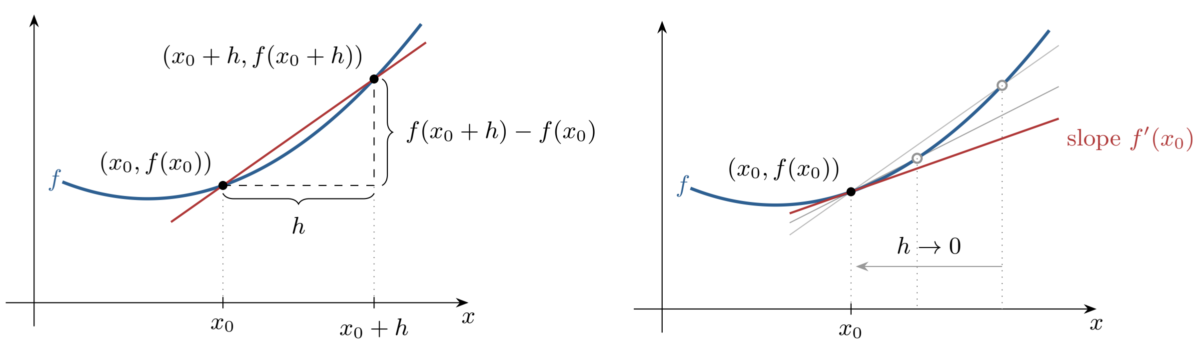

Figure 1.1: Left: for $h \neq 0$ the difference quotient (1) is the rise $f(x_0 + h) - f(x_0)$ over the run h, that is, the slope of the secant line through $(x_0, f(x_0))$ and $(x_0 + h, f(x_0 + h))$. Right: as $h \to 0$ the second point slides along the graph toward the first and the secants rotate toward a single limiting line. Its slope is the derivative $f'(x_0)$, and the line itself is the tangent to the graph at $(x_0, f(x_0))$.

We begin with three direct computations of the derivative using the definition.

Example 1.2 (Constant, affine, and quadratic functions)**.** Fix $x_0 \in \mathbb{R}$, and let $a, b, c \in \mathbb{R}$.

1. If $f(x) = c$, then $[f(x_0 + h) - f(x_0)]/h = 0$ for every $h \neq 0$, so $f'(x_0) = 0$.
2. If $f(x) = ax + b$, then
$$\frac{f(x_0 + h) - f(x_0)}{h} = \frac{a(x_0 + h) + b - (ax_0 + b)}{h} = a,$$
so $f'(x_0) = a$. The slope of an affine function is the same at every point.
3. If $f(x) = x^2$, then
$$\frac{f(x_0 + h) - f(x_0)}{h} = \frac{(x_0 + h)^2 - x_0^2}{h} = 2x_0 + h \longrightarrow 2x_0.$$
Thus $f'(x) = 2x$. In particular, $f'(3) = 6$.

1.2 Derivative as a linear approximation

Definition 1.1 has an equivalent form that will extend cleanly to several variables. Because I is open, $f(x_0 + h)$ is defined for every sufficiently small change h, so the limit below makes sense.

Theorem 1.3 (Derivative as a linear approximation)**.** *Let $I \subseteq \mathbb{R}$ be an open interval, let $f : I \to \mathbb{R}$, and let $x_0 \in I$. Then f is differentiable at x_0 if and only if there is $a \in \mathbb{R}$ such that*
$$\lim_{h \to 0} \frac{|f(x_0 + h) - f(x_0) - ah|}{|h|} = 0. \tag{2}$$
In that case a is unique, and $a = f'(x_0)$.

Proof. The proof relies on two facts. First, for any real-valued function u of h, $|u(h)| \to 0$ if and only if $u(h) \to 0$ as $h \to 0$ (see Lecture 2). Second, $|b/c| = |b|/|c|$ for all $b \in \mathbb{R}$ and $c \in \mathbb{R} \setminus \{0\}$. We have:
$$\begin{aligned} f \text{ is differentiable at } x_0 &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} = a. \\ &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h \to 0} \left(\frac{f(x_0 + h) - f(x_0) - ah}{h} \right) = 0. \\ &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h \to 0} \left| \frac{f(x_0 + h) - f(x_0) - ah}{h} \right| = 0. \\ &\iff \exists a \in \mathbb{R} \text{ such that } \lim_{h \to 0} \frac{|f(x_0 + h) - f(x_0) - ah|}{|h|} = 0, \end{aligned}$$
where the first equivalence is by the definition of differentiability, the second equivalence is by subtracting a from both sides, the third equivalence is the first fact, and the fourth equivalence is the second fact. Furthermore, the a in Definition 1.1 is unique and equals $f'(x_0)$, so the a in (2) is also unique and equals $f'(x_0)$.

□

Setting $x = x_0 + h$ gives an equivalent moving-point form of (2):

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0) - a(x - x_0)|}{|x - x_0|} = 0. \tag{3}$$

In both forms x_0 is fixed. In (2), $h \to 0$; in (3), the input $x = x_0 + h$ tends to x_0.

The reason why eq. (2) is called derivative as a linear approximation is as follows. Let $r(h) = f(x_0 + h) - f(x_0) - ah$ in the numerator of (2). Then, $f(x_0 + h) = f(x_0) + ah + r(h)$, where $f(x_0)+ah$ is the *linear approximation* of f evaluated at x_0+h, while $r(h)$ is the *remainder* term. The condition in (2), now written as $|r(h)|/|h| \to 0$ as $h \to 0$ says that the remainder term is negligible compared to h as $h \to 0$.

Example 1.4 (Reading the derivative from an expansion)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = x^2$, and let $x_0 = 3$. The function evaluated at $x_0 + h$ is

$$f(3 + h) = (3 + h)^2 = \underbrace{9}_{f(3)} + \underbrace{6h}_{\text{linear term}} + \underbrace{h^2}_{r(h)}.$$

The remainder satisfies

$$\lim_{h \to 0} \frac{|r(h)|}{|h|} = \lim_{h \to 0} \frac{|h^2|}{|h|} = \lim_{h \to 0} |h| = 0.$$

Therefore the linear coefficient is the derivative: $f'(3) = 6$.

We now have two ways of calculating derivatives for functions from $\mathbb{R}$ to $\mathbb{R}$: the difference-quotient Definition 1.1 and the linear-expansion method of Theorem 1.3. The latter method is generalizable to functions of several variables, so we will use it in the next section.

1.3 Properties of differentiability

Theorem 1.5 (Differentiability implies continuity)**.** *Let $I \subseteq \mathbb{R}$ be an open interval, let $f : I \to \mathbb{R}$, and let $x_0 \in I$. If f is differentiable at x_0, then f is continuous at x_0.*

Proof (optional). For $h \neq 0$ small enough that $x_0 + h \in I$,

$$\begin{aligned}
\lim_{h \to 0} \bigl[f(x_0 + h) - f(x_0)\bigr] &= \lim_{h \to 0} \left(\frac{f(x_0 + h) - f(x_0)}{h} \cdot h \right) \\
&= \left(\lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h} \right) \left(\lim_{h \to 0} h \right) \\
&= f'(x_0) \cdot 0 = 0,
\end{aligned}$$

where the last line uses Definition 1.1. Hence

$$\lim_{h \to 0} f(x_0 + h) = f(x_0) + \lim_{h \to 0} \bigl[f(x_0 + h) - f(x_0)\bigr] = f(x_0),$$

which is continuity of f at x_0. $\square$

The converse fails: a continuous function need not be differentiable.

Example 1.6 (Continuity without differentiability)**.** Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = |x|$. At the origin, $|f(x) - f(0)| = |x| \to 0$, so f is continuous. But for $h \neq 0$, the difference quotient $[f(h) - f(0)]/h = |h|/h$ equals 1 when $h > 0$ and -1 when $h < 0$. It has no limit as $h \to 0$, so f is not differentiable at 0.

Theorem 1.7 (Rules of differentiation)**.** *Let f and g be differentiable at x_0, and let $\alpha, \beta \in \mathbb{R}$. Then*

1. $(\alpha f + \beta g)'(x_0) = \alpha f'(x_0) + \beta g'(x_0)$;
2. $(fg)'(x_0) = f'(x_0)g(x_0) + f(x_0)g'(x_0)$;
3. *if* $g(x_0) \neq 0$, *then*
$$(f/g)'(x_0) = \frac{f'(x_0)g(x_0) - f(x_0)g'(x_0)}{g(x_0)^2}.$$

Its proof uses Definition 1.1 and is left as an exercise.

Theorem 1.8 (Chain rule)**.** *Let $I, J \subseteq \mathbb{R}$ be open intervals, let $g : I \to \mathbb{R}$ satisfy $g(I) \subseteq J$, and let $f : J \to \mathbb{R}$. If g is differentiable at $x_0 \in I$ and f is differentiable at $g(x_0)$, then $f \circ g$ is differentiable at x_0 and*
$$(f \circ g)'(x_0) = f'(g(x_0))\, g'(x_0).$$

Proof (optional). Write $y_0 = g(x_0)$. Because J is open, $y_0 + t \in J$ for every sufficiently small t, and for those t define
$$\varphi(t) = \begin{cases} \dfrac{f(y_0 + t) - f(y_0)}{t}, & t \neq 0, \\[2ex] f'(y_0), & t = 0. \end{cases}$$
By Definition 1.1, $\varphi(t) \to f'(y_0) = \varphi(0)$ as $t \to 0$, so φ is continuous at 0. Multiplying by t,
$$f(y_0 + t) - f(y_0) = t\,\varphi(t), \tag{4}$$
which holds for every t in the domain of φ, including $t = 0$.

Next, let $h \neq 0$ tend to zero with $x_0 + h \in I$, and set $t = g(x_0 + h) - g(x_0)$. Since g is continuous at x_0 by Theorem 1.5, $t \to 0$, so for every sufficiently small h, (4) reads
$$f(g(x_0 + h)) - f(g(x_0)) = \big[g(x_0 + h) - g(x_0)\big]\, \varphi\big(g(x_0 + h) - g(x_0)\big),$$
Dividing by h,
$$\frac{f(g(x_0 + h)) - f(g(x_0))}{h} = \frac{g(x_0 + h) - g(x_0)}{h} \cdot \varphi\big(g(x_0 + h) - g(x_0)\big).$$
As $h \to 0$, the first factor tends to $g'(x_0)$ by Definition 1.1. For the second, $t \to 0$ as noted above, and φ is continuous at 0, so $\varphi(g(x_0 + h) - g(x_0)) \to \varphi(0) = f'(y_0)$. Hence
$$\begin{aligned} \lim_{h \to 0} \frac{f(g(x_0 + h)) - f(g(x_0))}{h} &= \left(\lim_{h \to 0} \frac{g(x_0 + h) - g(x_0)}{h}\right) \left(\lim_{h \to 0} \varphi\big(g(x_0 + h) - g(x_0)\big)\right) \\ &= g'(x_0)\, f'(y_0). \end{aligned}$$
$\square$

1.4 Useful theorems for functions of one variable

Theorem 1.9 (Interior extremum)**.** *Let $a < b$, let $f : [a, b] \to \mathbb{R}$, and let $x^* \in (a, b)$. Suppose f is differentiable at x^* and that for some $\varepsilon > 0$,*

$$f(x^*) \geq f(x) \qquad \text{for every } x \in [a, b] \text{ with } |x - x^*| < \varepsilon.$$

Then $f'(x^*) = 0$. The same conclusion holds if instead $f(x^*) \leq f(x)$ for every such x.

Proof (optional). Because $x^* \in (a, b)$, we have $x^* + h \in [a, b]$ for every sufficiently small $|h|$. For such h with $0 < |h| < \varepsilon$ the hypothesis gives $f(x^* + h) - f(x^*) \leq 0$, and dividing by h preserves the inequality when $h > 0$ and reverses it when $h < 0$:

$$\frac{f(x^* + h) - f(x^*)}{h} \leq 0 \quad \text{if } h > 0, \qquad \frac{f(x^* + h) - f(x^*)}{h} \geq 0 \quad \text{if } h < 0.$$

Since f is differentiable at x^*, the two-sided limit of the difference quotient exists, so both one-sided limits exist and equal $f'(x^*)$. Letting $h \downarrow 0$ gives $f'(x^*) \leq 0$, and letting $h \uparrow 0$ gives $f'(x^*) \geq 0$, so $f'(x^*) = 0$. For the reversed inequality, apply this to $-f$. □

Theorem 1.10 (Rolle's theorem)**.** *Let $a < b$ and let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). If $f(a) = f(b)$, then there is $c \in (a, b)$ such that $f'(c) = 0$.*

Proof (optional). If f is constant, then $f'(c) = 0$ for every $c \in (a, b)$.

Otherwise $f(t) \neq f(a)$ for some $t \in [a, b]$. Suppose $f(t) > f(a)$. Since $[a, b]$ is compact and f is continuous, Theorem 3.5 from Lecture 2 gives a maximizer $c \in [a, b]$ of f on $[a, b]$. We thus have

$$f(c) \geq f(t) > f(a) = f(b),$$

in particular, $c \in (a, b)$. Then f is differentiable at c, and $f(c) \geq f(x)$ holds for every $x \in [a, b]$, including those x with $|x - c| < \varepsilon$ for some $\varepsilon > 0$. By Theorem 1.9, we conclude that $f'(c) = 0$.

If instead $f(t) < f(a)$, run the same argument with c being a minimizer. □

Theorem 1.11 (Mean value theorem)**.** *Let $a < b$ and let $f : [a, b] \to \mathbb{R}$ be continuous on $[a, b]$ and differentiable on (a, b). Then there is $c \in (a, b)$ such that*

$$f(b) - f(a) = (b - a) f'(c).$$

Proof (optional). Let

$$g(t) = f(t) - f(a) - \frac{f(b) - f(a)}{b - a}(t - a), \qquad t \in [a, b].$$

Then g is continuous on $[a, b]$, differentiable on (a, b), and $g(a) = g(b) = 0$. By Theorem 1.10 there is $c \in (a, b)$ with

$$0 = g'(c) = f'(c) - \frac{f(b) - f(a)}{b - a},$$

yielding the result. □

2 Real-valued functions of several variables

In this section, we consider functions from $\mathbb{R}^n$ to $\mathbb{R}$. Input vectors are column vectors. We use x for a generic or moving input and x_0 for a fixed point at which a derivative is evaluated. Thus $h = x - x_0$ is the change from x_0 to x.

2.1 Partial derivatives

Recall from Lecture 3 that e_i is the ith standard basis vector of $\mathbb{R}^n$, with 1 in coordinate i and 0 in every other coordinate. For a real number t, the vector te_i has t in coordinate i and 0 elsewhere, and $x_0 + te_i$ adds t to coordinate i of x_0 while leaving every other coordinate of x_0 unchanged.

Definition 2.1 (Partial derivative)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$, and let $x_0 \in U$. The *ith partial derivative* of f at x_0, where $i \in \{1, \ldots, n\}$, is

$$\frac{\partial f}{\partial x_i}(x_0) = \lim_{t \to 0} \frac{f(x_0 + te_i) - f(x_0)}{t},$$

provided the limit exists.

This is an ordinary one-variable derivative: hold all variables except x_i fixed and differentiate with respect to x_i.

Example 2.2 (A function of two variables)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by

$$f(x_1, x_2) = x_1^2 x_2.$$

Holding x_2 fixed and differentiating in x_1, and then reversing their roles, gives

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = 2x_1x_2, \qquad \frac{\partial f}{\partial x_2}(x_1, x_2) = x_1^2.$$

2.2 Differentiability and C^1 functions

Partial derivatives describe a function only along the coordinate lines through a point. Differentiability requires one linear approximation to work for small changes in all directions at once. A linear map from $\mathbb{R}^n$ to $\mathbb{R}$ has the form $h \mapsto a \cdot h$ for some $a \in \mathbb{R}^n$, so a and h, which were scalars in the one-variable case, are now both vectors.

Definition 2.3 (Differentiability)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$, and let $x_0 \in U$. The function f is *differentiable at* x_0 if there is $a \in \mathbb{R}^n$ such that

$$\lim_{h \to 0} \frac{|f(x_0 + h) - f(x_0) - a \cdot h|}{\|h\|} = 0,$$

where the limit is over $h \in \mathbb{R}^n$ with $h \neq 0$. It is *differentiable on* U if it is differentiable at every point of U.

When $n = 1$, the vector h and the coefficient a are scalars, $a \cdot h = ah$, and $\|h\| = |h|$. Thus Definition 2.3 is exactly the linear-approximation form of the one-variable derivative in Theorem 1.3. Only the form of the linear term has changed.

Equivalently,

$$f(x_0 + h) = f(x_0) + a \cdot h + r(h), \qquad \frac{|r(h)|}{\|h\|} \longrightarrow 0 \quad \text{as } h \to 0. \tag{5}$$

The same rule $h \mapsto a \cdot h$ must work for every way in which h can approach zero.

Alternatively, set $x = x_0 + h$. Then f is differentiable at x_0 if and only if there is $a \in \mathbb{R}^n$ such that

$$\lim_{x \to x_0} \frac{|f(x) - f(x_0) - a \cdot (x - x_0)|}{\|x - x_0\|} = 0. \tag{6}$$

The base point x_0 is fixed in this limit; x is the moving input.

Definition 2.4 (Gradient)**.** If all partial derivatives of f exist at x_0, the *gradient* of f at x_0 is the column vector

$$\nabla f(x_0) = \begin{pmatrix} \partial f/\partial x_1(x_0) \\ \vdots \\ \partial f/\partial x_n(x_0) \end{pmatrix} \in \mathbb{R}^n.$$

When f is differentiable, the vector in its linear approximation is exactly the gradient.

Theorem 2.5 (Consequences of differentiability)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable at $x_0 \in U$, with vector a as in Definition 2.3. Then every partial derivative of f exists at x_0 and*

$$a = \nabla f(x_0).$$

In particular, the vector a is unique. Moreover, f is continuous at x_0.

Proof (optional). Fix i and set $h = te_i$. Then $\|h\| = |t|$ and $a \cdot h = ta_i$, so Definition 2.3 gives

$$\frac{|f(x_0 + te_i) - f(x_0) - ta_i|}{|t|} \longrightarrow 0.$$

Equivalently,

$$\left| \frac{f(x_0 + te_i) - f(x_0)}{t} - a_i \right| \longrightarrow 0,$$

so the difference quotient tends to a_i. Thus the ith partial derivative exists and equals a_i. Repeating the argument for every i gives $a = \nabla f(x_0)$, and because the partial derivatives are defined independently of a, this also proves uniqueness.

For continuity, the Cauchy–Schwarz inequality gives

$$|a \cdot h| \leq \|a\| \|h\| \longrightarrow 0,$$

and

$$|r(h)| = \frac{|r(h)|}{\|h\|} \|h\| \longrightarrow 0.$$

Thus (5) gives $f(x_0 + h) \to f(x_0)$. □

Example 2.6 (A linear approximation in two variables)**.** Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x_1, x_2) = x_1^2 x_2$, and let $x_0 = (1, 2)^\top$. Expanding around x_0 gives

$$\begin{aligned} f(x_0 + h) &= (1 + h_1)^2 (2 + h_2) \\ &= 2 + 4h_1 + h_2 + \left(2h_1^2 + 2h_1 h_2 + h_1^2 h_2\right). \end{aligned}$$

The constant is $f(x_0)$ and the linear term is $(4, 1)^\top \cdot h$. If $\|h\| \leq 1$, the absolute value of the remainder is at most

$$2|h_1|^2 + 2|h_1 h_2| + |h_1|^2 |h_2| \leq 5\|h\|^2.$$

After division by $\|h\|$, this bound tends to zero. Thus f is differentiable at $x_0 = (1, 2)^\top$ and

$$\nabla f(x_0) = \begin{pmatrix} 4 \\ 1 \end{pmatrix}.$$

The converse of the first claim in Theorem 2.5 is false: partial derivatives can exist without a valid linear approximation. A convenient condition rules out this problem.

Definition 2.7 (C^1 function)**.** Let $U \subseteq \mathbb{R}^n$ be open. A function $f : U \to \mathbb{R}$ is C^1 *on* U, or *continuously differentiable on* U, if every partial derivative $\partial f/\partial x_i$ exists on U and is continuous as a function from U to $\mathbb{R}$.

Theorem 2.8 (C^1 implies differentiable)**.** *If $f : U \to \mathbb{R}$ is C^1 on the open set $U \subseteq \mathbb{R}^n$, then f is differentiable on U and the gradient $\nabla f : U \to \mathbb{R}^n$ is continuous.*

We use this theorem without proof. It provides the usual way to verify differentiability: compute the partial derivatives and check that they are continuous. In particular, polynomials are C^1 on $\mathbb{R}^n$, and rational functions are C^1 wherever their denominators are nonzero.

The implications established in this section are

$$\begin{gathered} C^1 \quad \Longrightarrow \quad \text{differentiable} \quad \Longrightarrow \quad \text{continuous}, \\ \text{differentiable} \quad \Longrightarrow \quad \text{all partial derivatives exist.} \end{gathered}$$

None of the converses holds in general.

The linear approximation can also be written using differential notation.

Remark 2.9 (Differential notation)*.* If f is differentiable at x_0, its *differential at* x_0 is the linear map

$$\mathrm{d}f_{x_0}(h) := \nabla f(x_0) \cdot h = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) h_i.$$

Thus $\mathrm{d}f_{x_0}(h) = \nabla f(x_0) \cdot h$ is exactly the linear term in

$$f(x_0 + h) = f(x_0) + \nabla f(x_0) \cdot h + r(h).$$

Economics texts often write the same expression as

$$\mathrm{d}f = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0)\,\mathrm{d}x_i.$$

Here $\mathrm{d}x_i$ labels a coordinate change; it is not the product of a number d and x_i. We use the actual changes h_i when applying the differential.

2.3 Directional derivatives

A partial derivative permits only one coordinate to change. A directional derivative allows the coordinates to change together in fixed proportions: for $v = (v_1, \ldots, v_n)^\top$, varying t traces the line $x_0 + tv$ through x_0.

Definition 2.10 (Directional derivative)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$, and let $x_0 \in U$. The *directional derivative* of f at x_0 in the direction $v \in \mathbb{R}^n$ is

$$\partial_v f(x_0) = \lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t},$$

provided the limit exists.

Taking $v = e_i$ gives the ith partial derivative. We do not require $\|v\| = 1$, so v specifies both the relative changes in the coordinates and their scale. When f is differentiable, its gradient gives every directional derivative.

Theorem 2.11 (Directional derivatives from the gradient)**.** *Let $U \subseteq \mathbb{R}^n$ be open and let $f : U \to \mathbb{R}$ be differentiable at $x_0 \in U$. Then, for every $v \in \mathbb{R}^n$, the directional derivative exists and*

$$\partial_v f(x_0) = \nabla f(x_0) \cdot v = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x_0) v_i.$$

Proof (optional). If $v = 0$, the difference quotient is zero for every $t \neq 0$, so the result holds. If $v \neq 0$, substitute $h = tv$ into (5) and use $a = \nabla f(x_0)$:

$$\frac{f(x_0 + tv) - f(x_0)}{t} = \nabla f(x_0) \cdot v + \frac{r(tv)}{t}.$$

Moreover,

$$\left| \frac{r(tv)}{t} \right| = \|v\| \frac{|r(tv)|}{\|tv\|} \longrightarrow 0.$$

Taking the limit proves the first equality; the second writes out the inner product coordinate by coordinate. □

2.4 Steepest increase

To compare directions independently of scale, we restrict v to unit vectors and ask which one gives the largest directional derivative.

Theorem 2.12 (Steepest increase)**.** *Let $U \subseteq \mathbb{R}^n$ be open, let $f : U \to \mathbb{R}$ be differentiable at $x_0 \in U$, and suppose $\nabla f(x_0) \neq 0$. Define the unit vector*

$$v^* = \frac{\nabla f(x_0)}{\|\nabla f(x_0)\|}.$$

Among all unit vectors $v \in \mathbb{R}^n$, the directional derivative $\partial_v f(x_0)$ has its unique maximum at $v = v^*$ and its unique minimum at $v = -v^*$. The corresponding values are

$$\partial_{v^*} f(x_0) = \|\nabla f(x_0)\|, \qquad \partial_{-v^*} f(x_0) = -\|\nabla f(x_0)\|.$$

Proof (optional). Write $g = \nabla f(x_0)$. For every unit vector v, Theorem 2.11 and the Cauchy–Schwarz inequality give

$$\partial_v f(x_0) = g \cdot v \leq |g \cdot v| \leq \|g\| \|v\| = \|g\|.$$

The unit vector $v^* = g/\|g\|$ satisfies $g \cdot v^* = \|g\|$, so it attains the bound. If a unit vector v also attains it, then $g \cdot v = \|g\|$ and

$$\|g - \|g\|v\|^2 = 2\|g\|\big(\|g\| - g \cdot v\big) = 0.$$

Hence $v = g/\|g\| = v^*$. Applying the maximum result to $-f$ gives the minimum and its unique minimizer. $\square$

Thus the gradient points in the direction of steepest increase, and its norm is the largest rate of increase per unit change. By Theorem 2.11, $\partial_v f(x_0) = 0$ exactly when $\nabla f(x_0) \cdot v = 0$. In this case, v and $\nabla f(x_0)$ are *orthogonal*. If $\nabla f(x_0) = 0$, every directional derivative is zero, so first-order information does not select a direction of increase or decrease.

This result underlies *gradient descent*, an algorithm used to train many machine-learning and AI models. If f is the loss function and x_k is the current parameter vector, gradient descent chooses a step size $\alpha_k > 0$ and updates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

The negative gradient gives the direction of steepest local decrease, while α_k determines the size of the step.

3 Vector-valued functions

3.1 Jacobians

A map $F : U \subseteq \mathbb{R}^n \to \mathbb{R}^m$ has m outputs. Its linear approximation is therefore a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, represented by an $m \times n$ matrix.

Definition 3.1 (Jacobian)**.** Let $U \subseteq \mathbb{R}^n$ be open, let $F : U \to \mathbb{R}^m$, and let $x_0 \in U$. The map F is *differentiable at* x_0 if there is an $m \times n$ matrix A such that

$$\lim_{h \to 0} \frac{\|F(x_0 + h) - F(x_0) - Ah\|}{\|h\|} = 0,$$

where the limit is over $h \in \mathbb{R}^n$ with $h \neq 0$. The matrix A is the *derivative*, or *Jacobian*, of F at x_0, written $\mathrm{D}F(x_0)$. If $F = (F_1, \dots, F_m)^\top$, the scalar-valued maps $F_i : U \to \mathbb{R}$ are the *component functions* of F. The map F is C^1 *on* U if every component F_i is C^1 on U.

Equivalently,

$$F(x_0 + h) = F(x_0) + \mathrm{D}F(x_0)h + r(h), \qquad \frac{\|r(h)\|}{\|h\|} \longrightarrow 0 \quad \text{as } h \to 0. \tag{7}$$

The linear term has dimensions

$$\mathrm{D}F(x_0)h : \qquad (m \times n)(n \times 1) = m \times 1,$$

so each row corresponds to an output and each column to an input.

With the moving input $x = x_0 + h$, the same definition is

$$\lim_{x \to x_0} \frac{\|F(x) - F(x_0) - \mathrm{D}F(x_0)(x - x_0)\|}{\|x - x_0\|} = 0. \tag{8}$$

Theorem 3.2 (Entries of the Jacobian)**.** *The map $F : U \to \mathbb{R}^m$ is differentiable at x_0 if and only if every component F_i is differentiable at x_0. In that case the Jacobian is unique and*

$$\mathrm{D}F(x_0) = \begin{pmatrix} \frac{\partial F_1}{\partial x_1}(x_0) & \cdots & \frac{\partial F_1}{\partial x_n}(x_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1}(x_0) & \cdots & \frac{\partial F_m}{\partial x_n}(x_0) \end{pmatrix}.$$

Thus the ith row of $\mathrm{D}F(x_0)$ is $\nabla F_i(x_0)^\top$.

Proof (optional). Suppose first that F is differentiable with derivative A, and let $r(h) = F(x_0 + h) - F(x_0) - Ah$. Since the absolute value of each coordinate of $r(h)$ is at most $\|r(h)\|$,

$$\frac{|F_i(x_0 + h) - F_i(x_0) - \sum_{j=1}^{n} A_{ij} h_j|}{\|h\|} \longrightarrow 0.$$

Thus F_i is differentiable, and Theorem 2.5 gives

$$(A_{i1}, \ldots, A_{in})^\top = \nabla F_i(x_0).$$

Conversely, suppose each F_i is differentiable, and define $A = (A_{ij})$ by

$$A_{ij} = \frac{\partial F_i}{\partial x_j}(x_0).$$

If $r_i(h)$ is the approximation error for component i, then $|r_i(h)|/\|h\| \to 0$. Because there are only finitely many components,

$$\frac{\|F(x_0 + h) - F(x_0) - Ah\|}{\|h\|} = \left(\sum_{i=1}^{m} \left(\frac{r_i(h)}{\|h\|} \right)^2 \right)^{1/2} \longrightarrow 0.$$

Thus F is differentiable with derivative A. The first part identifies every entry of any possible derivative, so the Jacobian is unique. $\square$

It follows from Theorems 2.8 and 3.2 that every C^1 map is differentiable.

Example 3.3 (A 2×3 Jacobian)**.** Let $F : \mathbb{R}^3 \to \mathbb{R}^2$ be given by

$$F(x, y, z) = \begin{pmatrix} x^2 y \\ y + 3z \end{pmatrix}.$$

Then

$$\mathrm{D}F(x, y, z) = \begin{pmatrix} 2xy & x^2 & 0 \\ 0 & 1 & 3 \end{pmatrix}, \qquad \mathrm{D}F(x_0) = \begin{pmatrix} 4 & 1 & 0 \\ 0 & 1 & 3 \end{pmatrix} \quad \text{at } x_0 = (1, 2, 0)^\top.$$

The two rows correspond to the two outputs, and the three columns correspond to the inputs x, y, z.

For a scalar-valued function $f : U \to \mathbb{R}$, the Jacobian is a $1 \times n$ row, whereas the gradient is an $n \times 1$ column:

$$\mathrm{D}f(x_0) = \nabla f(x_0)^\top. \tag{9}$$

Consequently, $\mathrm{D}f(x_0)h = \nabla f(x_0) \cdot h$. The gradient is convenient for geometric statements, while the Jacobian is convenient for composition. When $n = 1$, both contain the ordinary derivative $f'(x_0)$.

3.2 The multivariable chain rule

The one-variable chain rule says that derivatives multiply. The same rule holds for functions between Euclidean spaces, with matrix multiplication in place of scalar multiplication.

Theorem 3.4 (Chain rule)**.** *Let $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^m$ be open. Let $G : U \to \mathbb{R}^m$ satisfy $G(U) \subseteq V$, and let $F : V \to \mathbb{R}^p$. If G is differentiable at $x_0 \in U$ and F is differentiable at $G(x_0)$, then $F \circ G$ is differentiable at x_0 and*

$$\mathrm{D}(F \circ G)(x_0) = \mathrm{D}F(G(x_0))\,\mathrm{D}G(x_0).$$

We use this theorem without proof. Using the row description from Theorem 3.2, the product is

$$\mathrm{D}(F \circ G)(x_0) = \underbrace{\begin{pmatrix} \nabla F_1(G(x_0))^\top \\ \vdots \\ \nabla F_p(G(x_0))^\top \end{pmatrix}}_{p \times m} \underbrace{\begin{pmatrix} \nabla G_1(x_0)^\top \\ \vdots \\ \nabla G_m(x_0)^\top \end{pmatrix}}_{m \times n}.$$

The product is $p \times n$, the required size for the derivative of $F \circ G : U \to \mathbb{R}^p$. The outer rows are evaluated at $G(x_0)$, while the inner rows are evaluated at x_0. When $n = m = p = 1$, the formula reduces to the one-variable chain rule.

A useful special case follows a curve through the domain of a real-valued function.

Corollary 3.5 (Derivative along a curve)**.** *Let $V \subseteq \mathbb{R}^m$ be open, let $\gamma : I \to V$ be differentiable at $t_0 \in I$, where $I \subseteq \mathbb{R}$ is an open interval, and let $f : V \to \mathbb{R}$ be differentiable at $\gamma(t_0)$. Then*

$$(f \circ \gamma)'(t_0) = \nabla f(\gamma(t_0)) \cdot \gamma'(t_0) = \sum_{i=1}^{m} \frac{\partial f}{\partial x_i}(\gamma(t_0))\gamma_i'(t_0).$$

Indeed, (9) and the chain rule give $\mathrm{D}f(\gamma(t_0))\mathrm{D}\gamma(t_0) = \nabla f(\gamma(t_0))^\top \gamma'(t_0)$.

3.3 A worked chain rule

Example 3.6 (A vector-valued composition)**.** Define

$$G : \mathbb{R} \to \mathbb{R}^3, \qquad G(t) = \begin{pmatrix} 1+t \\ 2 \\ 3+2t \end{pmatrix}, \qquad F : \mathbb{R}^3 \to \mathbb{R}^2, \qquad F(u,v,w) = \begin{pmatrix} uw \\ v + w^2 \end{pmatrix}.$$

To find the derivative of $F \circ G$ at $t = 0$, compute

$$\mathrm{D}G(t) = G'(t) = \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix}, \qquad \mathrm{D}F(u,v,w) = \begin{pmatrix} w & 0 & u \\ 0 & 1 & 2w \end{pmatrix}.$$

Because $G(0) = (1,2,3)^\top$, the chain rule gives

$$\begin{aligned} \mathrm{D}(F \circ G)(0) &= \mathrm{D}F(G(0))\,\mathrm{D}G(0) \\ &= \begin{pmatrix} 3 & 0 & 1 \\ 0 & 1 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 12 \end{pmatrix}. \end{aligned}$$

The dimensions are $(2 \times 3)(3 \times 1) = 2 \times 1$, as required for a map from $\mathbb{R}$ to $\mathbb{R}^2$.

For a direct check,

$$(F \circ G)(t) = \begin{pmatrix} (1+t)(3+2t) \\ 2 + (3+2t)^2 \end{pmatrix} = \begin{pmatrix} 3 + 5t + 2t^2 \\ 11 + 12t + 4t^2 \end{pmatrix}.$$

Differentiating the components at 0 again gives $(5, 12)^\top$.