Math 241
Calculus III

Thomas Honold

ZJU-UIUC Institute

Fall Semester 2020

Outline

Today's Lecture: Differentiation of Multivariable Functions

Introduction

The following examples, of which you should know the first one, illustrate the main purpose of differentiation:

Local approximation by a linear map

Example ($f(x) = x^3$)

Considering $x_0 \in \mathbb{R}$ as fixed, real numbers close to x_0 have the form $x = x_0 + h$ with $|h|$ small.

$$\begin{aligned} f(x) = f(x_0 + h) = (x_0 + h)^3 &= x_0^3 + 3x_0^2 h + 3x_0 h^2 + h^3 \\ &= f(x_0) + \text{something linear in } h + R(h) \\ &\approx f(x_0) + \text{something linear in } h \end{aligned}$$

with approximation error $R(h) = 3x_0 h^2 + h^3$.
The error satisfies $R(h)/h = 3x_0 h + h^2 \to 0$ for $h \to 0$.
This is exactly what we need to show that

$$\frac{f(x_0 + h) - f(x_0)}{h} = 3x_0^2 + \frac{R(h)}{h} \to 3x_0^2 \quad \text{for } h \to 0, \quad \text{i.e.,} \quad f'(x_0) = 3x_0^2.$$

Using little-o notation, $\lim_{h\to 0} R(h)/h = 0$ is expressed as $R(h) = \mathrm{o}(h)$, and hence the approximation in the previous example as

$$f(x_0 + h) = f(x_0) + \text{something linear in } h + \mathrm{o}(h).$$

Now comes the first multivariable example.

Example ($f(x, y) = x^3 - 3xy^2$)

Here the displacement is of the form $\mathbf{h} = (h_1, h_2)$ and we get (dropping the index '0' in the fixed point (x, y) considered)

$$\begin{aligned}
f(x + h_1, y + h_2) &= (x + h_1)^3 - 3(x + h_1)(y + h_2)^2 \\
&= x^3 + 3x^2h_1 + 3xh_1^2 + h_1^3 - 3xy^2 - 6xyh_2 - 3xh_2^2 - 3h_1y^2 - 6h_1yh_2 - 3h_1h_2^2 \\
&= x^3 - 3xy^2 + 3(x^2 - y^2)h_1 - 6xyh_2 + 3xh_1^2 - 6yh_1h_2 - 3xh_2^2 + h_1^3 - 3h_1h_2^2 \\
&= f(x, y) + \text{something linear in } \mathbf{h} + R(\mathbf{h})
\end{aligned}$$

Since every monomial appearing in $R(\mathbf{h})$ has degree ≥ 2 and

$$\frac{|h_i|}{|\mathbf{h}|} = \frac{|h_i|}{\sqrt{h_1^2 + h_2^2}} \leq 1 \quad \text{for } i = 1, 2,$$

we have $R(\mathbf{h})/|\mathbf{h}| \to 0$ for $\mathbf{h} \to \mathbf{0} \in \mathbb{R}^2$ (i.e., $h_1 \to 0$ and $h_2 \to 0$ in $\mathbb{R}$).

Using little-o notation, we can express $\lim_{\mathbf{h}\to\mathbf{0}} R(\mathbf{h})/|\mathbf{h}| = 0$ as $R(\mathbf{h}) = \mathrm{o}(\mathbf{h})$, and hence the approximation in the previous example as

$$f\big((x,y)+\mathbf{h}\big) = f(x,y) + \text{something linear in } \mathbf{h} + \mathrm{o}(\mathbf{h}).$$

Example ($V(x,y,z) = xyz$)

The function $V(x,y,z)$ returns the volume of a cuboid with side lengths x,y,z. We have

$$\begin{aligned} V(x+h_1,y+h_2,z+h_3) - V(x,y,z) &= (x+h_1)(y+h_2)(z+h_3) - xyz \\ &= yzh_1 + xzh_2 + xyh_3 + zh_1h_2 + yh_1h_3 + xh_2h_3 + h_1h_2h_3 \\ &\approx yzh_1 + xzh_2 + xyh_3 \end{aligned}$$

with an error of order $\mathrm{o}(\mathbf{h})$, and thus substantially smaller than the maximum of $|h_1|, |h_2|, |h_3|$.

This says that a small change/error in the input of V, represented by $\mathbf{h} = (h_1,h_2,h_3)$, “propagates” to a change/error of approximately $yzh_1 + xzh_2 + xyh_3$ in the output of V, i.e., in the computed volume.

Example (squaring map)

The squaring map (real representation) was defined as $s(x,y) = (x^2 - y^2, 2xy)$ for $(x,y) \in \mathbb{R}^2$.

Using column vectors, its linear approximation in (x,y) is

$$\begin{aligned}
s(x+h_1, y+h_2) &= \begin{pmatrix} (x+h_1)^2 - (y+h_2)^2 \\ 2(x+h_1)(y+h_2) \end{pmatrix} \\
&= \begin{pmatrix} x^2 - y^2 + 2xh_1 - 2yh_2 + h_1^2 - h_2^2 \\ 2xy + 2yh_1 + 2xh_2 + 2h_1 h_2 \end{pmatrix} \\
&= \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix} + \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \begin{pmatrix} h_1^2 - h_2^2 \\ 2h_1 h_2 \end{pmatrix} \\
&= s(x,y) + \text{something linear in } \mathbf{h} + R(\mathbf{h})
\end{aligned}$$

Here $R(\mathbf{h})$ is vector-valued, but from

$$|R(\mathbf{h})| \leq \sqrt{2} \max\{\left|h_1^2 - h_2^2\right|, |2h_1 h_2|\}$$

we still get $|R(\mathbf{h})| / |\mathbf{h}| \to \mathbf{0}$ for $\mathbf{h} \to \mathbf{0}$ in the same way as before, i.e., $R(\mathbf{h}) = \mathrm{o}(\mathbf{h})$.

In the complex world, using $z = (x, y) = x + y\mathrm{i}$, $h = (h_1, h_2) = h_1 + h_2\mathrm{i}$, the approximation just obtained reads

$$(z + h)^2 = z^2 + 2zh + h^2 = z^2 + 2zh + \mathrm{o}(h),$$

so that the approximating linear map is multiplication by $2z = (z^2)'$. This is no coincidence.

Example ($f(x, y) = \mathrm{e}^{xy}$)

This example has been included, because it is genuinly non-polynomial. Here we can argue as follows:

$$\begin{aligned}
\mathrm{e}^{(x+h_1)(y+h_2)} - \mathrm{e}^{xy} &= \mathrm{e}^{xy+yh_1+xh_2+h_1h_2} - \mathrm{e}^{xy} = \mathrm{e}^{xy}\left(\mathrm{e}^{yh_1+xh_2+h_1h_2} - 1\right) \\
&= \mathrm{e}^{xy}\left(yh_1 + xh_2 + \text{terms of degree} \geq 2 \text{ in } \mathbf{h}\right) \\
&= (y\mathrm{e}^{xy})h_1 + (x\mathrm{e}^{xy})h_2 + \mathrm{o}(\mathbf{h}).
\end{aligned}$$

Without going into details, this follows by expanding

$$\mathrm{e}^{yh_1+xh_2+h_1h_2} = \sum_{k=0}^{\infty} \frac{(yh_1 + xh_2 + h_1h_2)^k}{k!}$$

into a double series and suitably rearranging terms.

Notes on the preceding examples

- According to the subsequent definition of differentiable maps, all five functions are differentiable everywhere (the case of $V(x, y, z)$ requires further justification!), and the differential of the function in a particular point is the linear map which sends $\mathbf{h}$ (a vector in $\mathbb{R}$, $\mathbb{R}^2$, resp., $\mathbb{R}^3$) to the blue expression stated in the approximation formula.
- Be sure to understand that a given function f is associated with many linear maps in this way, one for each point $\mathbf{x}$ (denoted by x_0, (x, y), or (x, y, z) in the examples) of its domain. The differential $\mathrm{d}f$ of f is the map (non-linear in general) that sends $\mathbf{x}$ to the linear map used at $\mathbf{x}$, viz. $\mathrm{d}f(\mathbf{x})$.
- For the explicit computation of the linear maps involved we need a formula for obtaining their “coefficients”, i.e., the entries of their representing matrices. In the examples the underlying pattern can be already seen, e.g., think how the coefficients of $(h_1, h_2) \mapsto (y\mathrm{e}^{xy})h_1 + (x\mathrm{e}^{xy})h_2$ arise from e^{xy}. In this regard also note that the coefficient of h_1, say, can be obtained by setting $h_2 = 0$ and considering the resulting one-dimensional approximation problem.

Differentiable Maps

Definition

Suppose $D \subseteq \mathbb{R}^n$ and $\mathbf{x}_0 \in D$. The point $\mathbf{x}_0$ is said to be an *inner point* of D if D contains a ball of positive radius around $\mathbf{x}_0$, i.e., there exists $r > 0$ such that $|\mathbf{x} - \mathbf{x_0}| < r$ implies $\mathbf{x} \in D$. The set of inner points of D is denoted by D°.

The remaining points of D are boundary points (but the boundary ∂D may contain points in $\mathbb{R}^n \setminus D$).

Now comes the most important definition of Calculus III.

Definition

Suppose $f\colon D \to \mathbb{R}^m$ is a map with domain $D \subseteq \mathbb{R}^n$ and $\mathbf{x}_0$ is an inner point of D. The map f is said to be *differentiable* at $\mathbf{x}_0$ if there exists a linear map $L\colon \mathbb{R}^n \to \mathbb{R}^m$ such that

$$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + L(\mathbf{h}) + \mathrm{o}(\mathbf{h}) \quad \text{for} \quad \mathbf{h} \to \mathbf{0}. \tag{TD}$$

If this is the case then the linear map L, which is uniquely determined, is called the *differential* of f at $\mathbf{x}_0$ and denoted by $\mathrm{d}f(\mathbf{x}_0)$.

Notes

- If you are uncomfortable with the little-o notation, take the following equivalent formulation of (TD):

  $$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{x}_0+\mathbf{h})-f(\mathbf{x}_0)-L(\mathbf{h})}{|\mathbf{h}|} = \mathbf{0} \in \mathbb{R}^m$$

  Both formulations say in particular that for $\mathbf{h}\to\mathbf{0}$ the error term of the linear approximation $f(\mathbf{x}_0+\mathbf{h}) \approx f(\mathbf{x}_0)+L(\mathbf{h})$ is substantially smaller than $\mathbf{h}$ in length.

- The linear map L in (TD) is indeed uniquely determined: If L_1 and L_2 satisfy (TD), we must have $L_1(\mathbf{h})-L_2(\mathbf{h}) = \mathrm{o}(\mathbf{h})$ for $\mathbf{h}\to\mathbf{0}$. Now let $\mathbf{h}=h\mathbf{v}$ with $\mathbf{v}\in\mathbb{R}^n$ a fixed nonzero vector. Since $h\mathbf{v}\to\mathbf{0}$ for $h\to 0$, the quotient

  $$\frac{|L_1(h\mathbf{v})-L_2(h\mathbf{v})|}{|h\mathbf{v}|} = \frac{\bigl|h\bigl(L_1(\mathbf{v})-L_2(\mathbf{v})\bigr)\bigr|}{|h\mathbf{v}|} = \frac{|L_1(\mathbf{v})-L_2(\mathbf{v})|}{|\mathbf{v}|}$$

  tends to 0 for $h\to 0$, which can't be unless $L_1(\mathbf{v})=L_2(\mathbf{v})$.

  Question: Where have we used that $\mathbf{x}_0$ is an inner point of D? *Answer:* To have $\mathbf{x}_0+h\mathbf{v}\in D$ for small $|h|$.

Notes cont'd

- Linear maps $L\colon \mathbb{R} \to \mathbb{R}$ have the form $L(h) = ah$ for some $a \in \mathbb{R}$. Hence in the case $m = n = 1$ (TD) reduces to the familiar

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

  which just says $f'(x_0) = a$. In other words, differentiable functions $f\colon I \to \mathbb{R}$, $I \subseteq \mathbb{R}$, in the old sense remain differentiable in the new sense and have differential $x_0 \mapsto \mathrm{d}f(x_0)\colon \mathbb{R} \to \mathbb{R}$, $h \mapsto f'(x_0)h$.
  For curves $f\colon I \to \mathbb{R}^n$ the same is true (mutatis mutandis).
- As we have seen, linear maps $L\colon \mathbb{R}^n \to \mathbb{R}^m$ have the form $L(\mathbf{x}) = \mathbf{A}\mathbf{x}$ for some (uniquely determined) matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$. Hence the condition in (TD) can be rephrased as: There exists a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ such that

  $$f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \mathbf{A}\mathbf{h} + \mathrm{o}(\mathbf{h}) \quad \text{for} \quad \mathbf{h} \to \mathbf{0}.$$

  The matrix $\mathbf{A}$, which is uniquely determined by the previous note, is called *Jacobi(an) matrix* of f at $\mathbf{x}_0$ and denoted by $\mathbf{J}_f(\mathbf{x_0})$.

Notes cont'd

- A vector-valued function $f = (f_1, \dots, f_m)$ is differentiable at $\mathbf{x}_0$ iff each coordinate function f_i is differentiable at $\mathbf{x}_0$. This is due to the fact that limits of vector-valued functions can be computed coordinate-wise.
- Finally a note on the various sets $\overline{D}$, ∂D, D', D° defined for any set $D \subseteq \mathbb{R}^n$. In what follows, $\uplus$ denotes the *disjoint union* of sets, i.e., $M = S \uplus T$ means $M = S \cup T$ and $S \cap T = \emptyset$.

  For any D we have $D^\circ \subseteq D \subseteq \overline{D}$, $D^\circ \subseteq D' \subseteq \overline{D}$, $\overline{D} = D \cup D' = D \cup \partial D$, and the decomposition

$$\begin{aligned}
\mathbb{R}^n &= D^\circ \uplus \partial D \uplus (\mathbb{R}^n \setminus D)^\circ \\
&= \overline{D} \uplus (\mathbb{R}^n \setminus D)^\circ && (\overline{D} = D^\circ \uplus \partial D) \\
&= D^\circ \uplus \overline{\mathbb{R}^n \setminus D}. && \text{(by symmetry)}
\end{aligned}$$

  The boundary $\partial D = \partial(\mathbb{R}^n \setminus D)$ consists of those points $\mathbf{x}$ for which every ball around $\mathbf{x}$ contains points of D as well as points of $\mathbb{R}^n \setminus D$. Points in $D \cap \partial D$ must be accumulation points of $\mathbb{R}^n \setminus D$ but need not be accumulation points of D.

The Five Examples reconsidered

1. $f(x) = x^3$ is differentiable in $\mathbb{R}$ with differential
$$\mathrm{d}f(x)\colon \mathbb{R} \to \mathbb{R},\ h \mapsto 3x^2 h.$$
2. $f(x) = x^3 - 3xy^2$ is differentiable in $\mathbb{R}^2$ with differential
$$\mathrm{d}f(x,y)\colon \mathbb{R}^2 \to \mathbb{R},\ (h_1, h_2) \mapsto 3(x^2 - y^2)h_1 - 6xyh_2.$$
3. $V(x,y,z) = xyz$ is differentiable in $\mathbb{R}^3$ with differential
$$\mathrm{d}V(x,y,z)\colon \mathbb{R}^3 \to \mathbb{R},\ (h_1, h_2, h_3) \mapsto yzh_1 + xzh_2 + xyh_3.$$
4. $s(x,y) = (x^2 - y^2, 2xy)$ is differentiable in $\mathbb{R}^2$ with differential
$$\mathrm{d}s(x,y)\colon \mathbb{R}^2 \to \mathbb{R}^2,\ \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \mapsto \underbrace{\begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}}_{\mathbf{J}_s(x,y)} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}.$$
5. $f(x,y) = \mathrm{e}^{xy}$ is differentiable in $\mathbb{R}^2$ with differential
$$df(x,y)\colon \mathbb{R}^2 \to \mathbb{R},\ (h_1, h_2) \mapsto y\mathrm{e}^{xy}h_1 + x\mathrm{e}^{xy}h_2.$$

Further Examples

Example (linear maps)

Consider a linear map $f\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$. Here we have

$$f(\mathbf{x}_0 + \mathbf{h}) = \mathbf{A}(\mathbf{x}_0 + \mathbf{h}) = \mathbf{A}\mathbf{x}_0 + \mathbf{A}\mathbf{h},$$

and there is no remainder term. This implies that f is differentiable at $\mathbf{x}_0$ with differential $\mathrm{d}f(\mathbf{x}_0)\colon \mathbf{h} \mapsto \mathbf{A}\mathbf{h}$.

In other words, a linear map f is differentiable everywhere and the differential $\mathrm{d}f(\mathbf{x})$ coincides with f at any point $\mathbf{x} \in \mathbb{R}^n$.

For affine maps $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ the same is true (except that the differential doesn't coincide with f if $\mathbf{b} \neq \mathbf{0}$), because the constant vector $\mathbf{b}$ does not matter for differentation.

Example (quadratic forms)

A *quadratic form* on $\mathbb{R}^n$ is a map $q\colon \mathbb{R}^n \to \mathbb{R}$ of the form $q(\mathbf{x}) = \sum_{1\leq i\leq j\leq n} q_{ij}x_ix_j$ (i.e., a homogeneous polynomial of degree 2).

Setting $a_{ii} = q_{ii}$ and $a_{ij} = a_{ji} = q_{ij}/2$ for $i < j$ and viewing $\mathbf{x} \in \mathbb{R}^n$ as a column vector, we have

$$q(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \quad \text{with } \mathbf{A} \in \mathbb{R}^{n\times n},\ \mathbf{A} = \mathbf{A}^\mathsf{T}.$$

This representation is best suited for differentiating:

$$\begin{aligned} q(\mathbf{x}+\mathbf{h}) &= (\mathbf{x}+\mathbf{h})^\mathsf{T}\mathbf{A}(\mathbf{x}+\mathbf{h}) \\ &= \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{h}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{h} + \mathbf{h}^\mathsf{T}\mathbf{A}\mathbf{h} \\ &= q(\mathbf{x}) + 2\,\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{h} + \mathbf{h}^\mathsf{T}\mathbf{A}\mathbf{h} \qquad (\text{since } \mathbf{A} = \mathbf{A}^\mathsf{T}) \end{aligned}$$

$\mathbf{h}^\mathsf{T}\mathbf{A}\mathbf{h} = q(\mathbf{h})$ is a sum of terms $q_{ij}h_ih_j$, and we have

Example (cont'd)

$$\frac{|q_{ij}h_ih_j|}{|\mathbf{h}|} \leq |q_{ij}|\,|h_j| \leq |q_{ij}|\,|\mathbf{h}|\,.$$

This shows $\mathbf{h}^{\mathsf{T}}\mathbf{A}\mathbf{h} = \mathrm{o}(\mathbf{h})$ for $\mathbf{h} \to \mathbf{0}$ and hence that q is differentiable everywere with differential

$$\mathrm{d}q(\mathbf{x})\colon \mathbb{R}^n \to \mathbb{R},\ \mathbf{h} \mapsto 2\,\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{h} = 2(\mathbf{A}\mathbf{x})^{\mathsf{T}}\mathbf{h}.$$

In other words, the differential of q at $\mathbf{x} \in \mathbb{R}^n$ is "taking the dot product with the column vector $2(\mathbf{A}\mathbf{x})$", which represents a linear map.

In contrast with the linear case, however, the differential $\mathrm{d}q(\mathbf{x})$ of a quadratic form depends on the particular point $\mathbf{x}$.

As a concrete example, in the two-variable case

$$q(x,y) = ax^2 + 2bxy + cy^2 = \begin{pmatrix} x \\ y \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

we have $\mathrm{d}q(x,y)(h_1,h_2) = 2(ax+by)h_1 + 2(bx+cy)h_2$.

Example

The length function $\mathbb{R}^n \to \mathbb{R}$, $\mathbf{x} \mapsto |\mathbf{x}|$ is differentiable at any point $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ but not differentiable at the origin.

We restrict ourselves to the case $n = 2$. (The proof in the general case can be easily inferred from this.)

First we show that $\mathbf{x} \mapsto |\mathbf{x}|$ is not differentiable at $\mathbf{0} = (0,0)$. For the special choice $\mathbf{h} = (h,0) = h\,\mathbf{e}_1$ we have $L(\mathbf{h}) = h\,L(\mathbf{e}_1)$ but

$$|\mathbf{0}+\mathbf{h}| - |\mathbf{0}| = |\mathbf{h}| = |(h,0)| = |h| = \pm h,$$

which cannot be approximated within $\mathrm{o}(h)$ by a single linear map. (We should define $L(\mathbf{e}_1) = 1$ for $h > 0$, but $L(\mathbf{e}_1) = -1$ for $h < 0$.)

Now consider $\mathbf{x} = (x_1, x_2) \neq (0,0)$. Here we must estimate

$$\begin{aligned}
|\mathbf{x}+\mathbf{h}| - |\mathbf{x}| &= \sqrt{(x_1+h_1)^2 + (x_2+h_2)^2} - \sqrt{x_1^2+x_2^2} \\
&= \sqrt{x_1^2+x_2^2+2(x_1h_1+x_2h_2)+h_1^2+h_2^2} - \sqrt{x_1^2+x_2^2} \\
&= \frac{2x_1h_1+2x_2h_2+h_1^2+h_2^2}{\sqrt{x_1^2+x_2^2+2(x_1h_1+x_2h_2)+h_1^2+h_2^2}+\sqrt{x_1^2+x_2^2}}
\end{aligned}$$

Example (cont'd)

This has the form

$$|\mathbf{x}+\mathbf{h}|-|\mathbf{x}|=\frac{2x_1h_1+2x_2h_2+h_1^2+h_2^2}{g(h_1,h_2)},$$

where $g(h_1,h_2)$ is continuous at $(0,0)$ and $g(0,0)=2\sqrt{x_1^2+x_2^2}\neq 0$.

We now show that the linear map $L\colon\mathbb{R}^2\to\mathbb{R}$ defined by

$$\begin{aligned}L(\mathbf{h})=L(h_1,h_2)&=\frac{2x_1h_1+2x_2h_2}{g(0,0)}\\&=\frac{x_1}{\sqrt{x_1^2+x_2^2}}\cdot h_1+\frac{x_2}{\sqrt{x_1^2+x_2^2}}\cdot h_2\end{aligned}$$

has the required approximation property:

$$\begin{aligned}|\mathbf{x}+\mathbf{h}|-|\mathbf{x}|-L(\mathbf{h})&=(2x_1h_1+2x_2h_2)\left(\frac{1}{g(h_1,h_2)}-\frac{1}{g(0,0)}\right)+\frac{h_1^2+h_2^2}{g(h_1,h_2)}\\&=\mathrm{O}(|\mathbf{h}|)\,\mathrm{o}(1)+\mathrm{O}(|\mathbf{h}|^2)=\mathrm{o}(|\mathbf{h}|)=\mathrm{o}(\mathbf{h}),\end{aligned}$$

as claimed.

Partial derivatives

How to get the entries of the Jacobi matrix

In the preceding example we have obtained the Jacobi matrix of the length function $f\colon \mathbb{R}^2 \to \mathbb{R}$, $\mathbf{x} \mapsto \sqrt{x_1^2 + x_2^2}$:

$$\mathbf{J}_f(\mathbf{x}) = \left(\frac{x_1}{\sqrt{x_1^2 + x_2^2}}, \frac{x_2}{\sqrt{x_1^2 + x_2^2}} \right).$$

Question

How to to obtain the entries of $\mathbf{J}_f(\mathbf{x})$, and hence the differential $\mathrm{d}f(\mathbf{x})\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{h} \mapsto \mathbf{J}_f(\mathbf{x})\mathbf{h}$, in general?

The answer uses only Calculus I and can be found by inspecting our earlier examples. It involves the so-called partial derivatives of f.

Example (squaring map continued)

We have seen that $s(x,y) = (x^2 - y^2, 2xy)$ is differentiable in $\mathbb{R}^2$ with differential

$$\mathrm{d}s(x,y)(h_1,h_2) = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix}.$$

The entries of the Jacobi matrix can be obtained without the (rather complicated) expansion step by setting $h_1 = 0$, respectively, $h_2 = 0$ in the approximation formula

$$s(x+h_1, y+h_2) = s(x,y) + \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix} \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} + \mathrm{o}\big((h_1,h_2)\big).$$

For example, setting $h_2 = 0$ gives

$$s(x+h_1, y) = s(x,y) + h_1 \begin{pmatrix} 2x \\ 2y \end{pmatrix} + \mathrm{o}(h_1).$$

$$\implies \begin{pmatrix} 2x \\ 2y \end{pmatrix} = \lim_{h_1 \to 0} \frac{s(x+h_1,y) - s(x,y)}{h_1} = \frac{\mathrm{d}}{\mathrm{d}x}\big(x \mapsto s(x,y)\big).$$

Definition

Let $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$, be a real-valued function. The *partial derivative* of f with respect to the variable x_j, $1 \leq j \leq n$, is the function that assigns to $\mathbf{x} = (x_1, \dots, x_n) \in D$ the derivative of $t \mapsto (x_1, \dots, x_{j-1}, t, x_{j+1}, \dots, x_n)$ at $t = x_j$. The partial derivatives of f are denoted by f_{x_j} or $\frac{\partial f}{\partial x_j}$.

Notes

- According to the definition of derivatives in Calculus I we have

$$f_{x_j}(\mathbf{x}) = \frac{\partial f}{\partial x_j}(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\,\mathbf{e}_j) - f(\mathbf{x})}{h}.$$

- The partial derivatives of f are obtained by viewing f as a function of one variable x_j (keeping all other variables fixed) and applying the usual rules for computing derivatives learned in Calculus I to this function (resp., to its coordinate functions).

Notes cont'd

- The (maximal) domain D_j of f_{x_j} consists of all $\mathbf{x} \in D$ for which the limit $\lim_{h\to 0} \frac{f(\mathbf{x}+h\,\mathbf{e}_j)-f(\mathbf{x})}{h}$ exists.
- If $\mathbf{x} \in D$ is such that all partial derivatives $f_{x_j}(\mathbf{x})$, $1 \leq j \leq n$, exist (i.e., $\mathbf{x} \in \bigcap_{j=1}^n D_j$), we say that f is *partially differentiable* at $\mathbf{x}$ (and partially differentiable per se if this is true for all $\mathbf{x} \in D$).
- We will see in a moment that differentiability implies partial differentiability (at a point $\mathbf{x} \in D$) but not conversely. To make this difference clear, differentiability is also referred to as "*total* differentiability".
- Partial derivatives $\frac{\partial f}{\partial x_j}$ for vectorial functions $f = (f_1, \dots, f_m)$ are defined in the same way. The rules for computing limits of vectorial functions imply that $\frac{\partial f}{\partial x_j} = \left(\frac{\partial f_1}{\partial x_j}, \dots, \frac{\partial f_m}{\partial x_j}\right)$. Thus, anticipating Part (2) of the theorem on the next slide, we can say that the entries of $\mathbf{J}_f(\mathbf{x})$ are the scalar partial derivatives $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$, and the columns of $\mathbf{J}_f(\mathbf{x})$ are the vectorial partial derivatives $\frac{\partial f}{\partial x_j}(\mathbf{x})$. (Recall that we should consider a vectorial function f as a column vector $(f_1, \dots, f_m)^{\mathsf{T}}$.)

Theorem

Let $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, be a function with coordinate functions $f_1, \dots, f_m$ and $\mathbf{x} \in D^\circ$.

1. *If f is differentiable at $\mathbf{x}$ then f is continuous at $\mathbf{x}$.*
2. *If f is differentiable at $\mathbf{x}$ then the partial derivatives $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ exist for $1 \leq i \leq m$, $1 \leq j \leq n$, and*

$$\mathbf{J}_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{pmatrix}.$$

3. *Conversely, if all partial derivatives $\frac{\partial f_i}{\partial x_j}$ exist near $\mathbf{x}$ (i.e., f is partially differentiable in some ball around $\mathbf{x}$) and are continuous at $\mathbf{x}$, then f is differentiable at $\mathbf{x}$.*

Proof.

(1) With $L = \mathrm{d}f(\mathbf{x})$ we have

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + L(\mathbf{h}) + \mathrm{o}(\mathbf{h}) = f(\mathbf{x}) + L(\mathbf{h}) + \mathrm{o}(1)$$

for $\mathbf{h} \to \mathbf{0}$, and it remains to show that

$$\lim_{\mathbf{h}\to\mathbf{0}\in\mathbb{R}^n} L(\mathbf{h}) = \mathbf{0} \in \mathbb{R}^m.$$

This amounts to L being continuous at $\mathbf{0} \in \mathbb{R}^n$ and is easily verified from the matrix representation $L(\mathbf{h}) = \mathbf{A}\mathbf{h}$, $\mathbf{A} = \mathbf{J}_f(\mathbf{x})$.

(2) Specializing the approximation property to $\mathbf{h} = h\,\mathbf{e}_j$, $h \in \mathbb{R}$, gives

$$f(\mathbf{x}+h\,\mathbf{e}_j) = f(\mathbf{x}) + L(h\,\mathbf{e}_j) + \mathrm{o}(h\,\mathbf{e}_j) = f(\mathbf{x}) + h\,L(\mathbf{e}_j) + \mathrm{o}(h).$$

Subtracting $f(\mathbf{x})$ and dividing by h gives further

$$\frac{f(\mathbf{x}+h\,\mathbf{e}_j) - f(\mathbf{x})}{h} = L(\mathbf{e}_j) + \mathrm{o}(1), \quad \text{i.e.} \quad \lim_{h\to 0} \frac{f(\mathbf{x}+h\,\mathbf{e}_j) - f(\mathbf{x})}{h} = L(\mathbf{e}_j).$$

Passing to the coordinate functions f_i then shows that $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ exists for all i,j and forms the (i,j) entry of $\mathbf{J}_f(\mathbf{x})$. □

Proof cont'd.

(3) We assume $n = 2$ and $m = 1$ for simplicity (but the proof in the general case can be easily inferred from this case).

For sufficiently small $\mathbf{h} = (h_1, h_2)$ we have

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) &= f(x_1+h_1, x_2+h_2) - f(x_1, x_2) \\ &= f(x_1+h_1, x_2+h_2) - f(x_1, x_2+h_2) + f(x_1, x_2+h_2) - f(x_1, x_2). \end{aligned}$$

Applying the Mean Value Theorem from Calculus I to the functions $g_1(s) = f(s, x_2+h_2)$ and $g_2(t) = f(x_1, t)$ shows the existence of $\xi_1 \in (x_1, x_1+h_1)$, $\xi_2 \in (x_2, x_2+h_2)$ such that

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) &= g_1'(\xi_1)h_1 + g_2'(\xi_2)h_2 \\ &= \frac{\partial f}{\partial x_1}(\xi_1, x_2+h_2)h_1 + \frac{\partial f}{\partial x_2}(x_1, \xi_2)h_2. \end{aligned}$$

Finally, the continuity of $\frac{\partial f}{\partial x_1}$, $\frac{\partial f}{\partial x_2}$ at $\mathbf{x}$ gives for $\mathbf{h} \to \mathbf{0}$ the estimate

$$\begin{aligned} f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) &= \left(\frac{\partial f}{\partial x_1}(x_1, x_2) + \mathrm{o}(1)\right) h_1 + \left(\frac{\partial f}{\partial x_2}(x_1, x_2) + \mathrm{o}(1)\right) h_2 \\ &= \frac{\partial f}{\partial x_1}(x_1, x_2)h_1 + \frac{\partial f}{\partial x_2}(x_1, x_2)h_2 + \underbrace{\mathrm{o}(1)h_1 + \mathrm{o}(1)h_2}_{=\mathrm{o}(\mathbf{h})}. \end{aligned}$$

Afternote

After class I realized that students have some problems with the use of Big-0/little-o notation in the wider setting of vectorial and multivariable functions, including mixed-dimension cases. Here is the definition in more detail:

Suppose $D \subseteq \mathbb{R}^n$, $f\colon D \to \mathbb{R}^{m_1}$, $g\colon D \to \mathbb{R}^{m_2}$, and $\mathbf{x}_0 \in D^\circ$.

1. We say $f(\mathbf{x}) = \mathrm{O}\bigl(g(\mathbf{x})\bigr)$ for $\mathbf{x} \to \mathbf{x}_0$ if there exist constants $C, \delta > 0$ such that $|f(\mathbf{x})| \leq C\,|g(\mathbf{x})|$ for all $\mathbf{x} \in \mathrm{B}_\delta(\mathbf{x}_0) \setminus \{\mathbf{x}_0\}$
2. We say $f(\mathbf{x}) = \mathrm{o}\bigl(g(\mathbf{x})\bigr)$ for $\mathbf{x} \to \mathbf{x}_0$ if for every $\epsilon > 0$ there exists $\delta = \delta(\epsilon) > 0$ such that $|f(\mathbf{x})| \leq \epsilon\,|g(\mathbf{x})|$ for all $\mathbf{x} \in \mathrm{B}_\delta(\mathbf{x}_0) \setminus \{\mathbf{x}_0\}$.

Here $|f(\mathbf{x})|$, $|g(\mathbf{x})|$ denote the Euclidean lengths of $f(\mathbf{x})$, $g(\mathbf{x})$, and it is tacitly assumed that δ is chosen in such a way that $\mathrm{B}_\delta(\mathbf{x}_0) \subseteq D$.

"$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + L(\mathbf{h}) + \mathrm{o}(\mathbf{h})$ for $\mathbf{h} \to \mathbf{0}$" is a special case of (2): Here $\mathbf{x}_0 = \mathbf{0}$, the vectorial variable is $\mathbf{h}$ instead of $\mathbf{x}$, the function $\mathbf{h} \mapsto f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x}) - L(\mathbf{h})$ plays the role of f, and $g(\mathbf{h}) = \mathbf{h}$.

"$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + L(\mathbf{h}) + \mathrm{o}(1)$" is also a special case of (2): Use $g(\mathbf{h}) = 1$.

"$\mathrm{o}(\mathbf{h}) = \mathrm{o}(1)$" means "if $f(\mathbf{h}) = \mathrm{o}(\mathbf{h})$ then $f(\mathbf{h}) = \mathrm{o}(1)$". Note that …

Afternote cont'd

... the equality sign doesn't obey the usual rules here but is used informally just like "is" is often in common English: Any function that is $\mathrm{o}(\mathbf{h})$ for $\mathbf{h} \to \mathbf{0}$ *is* also $\mathrm{o}(1)$, but not conversely.

The main purpose of using Big-O/little-o notation in the lecture is to make limit calculations more concise. Compare the final part of the proof of Part (3) of the theorem to the following:

$$\begin{aligned}
f(\mathbf{x}+\mathbf{h})-f(\mathbf{x}) &= \frac{\partial f}{\partial x_1}(\xi_1, x_2+h_2)h_1 + \frac{\partial f}{\partial x_2}(x_1,\xi_2)h_2 \\
&= \frac{\partial f}{\partial x_1}(x_1,x_2)h_1 + \frac{\partial f}{\partial x_2}(x_1,x_2)h_2 + \underbrace{\left(\frac{\partial f}{\partial x_1}(\xi_1,x_2+h_2) - \frac{\partial f}{\partial x_1}(x_1,x_2)\right)}_{\to 0} h_1 \\
&\quad + \underbrace{\left(\frac{\partial f}{\partial x_2}(x_1,\xi_2) - \frac{\partial f}{\partial x_2}(x_1,x_2)\right)}_{\to 0} h_2 \\
&= \frac{\partial f}{\partial x_1}(x_1,x_2)h_1 + \frac{\partial f}{\partial x_2}(x_1,x_2)h_2 + \mathrm{o}(\mathbf{h}) \quad \text{for } \mathbf{h}=(h_1,h_2)\to(0,0).
\end{aligned}$$

The two $\mathrm{o}(1)$'s have been replaced. Now imagine we want to get rid of the $\mathrm{o}(\mathbf{h})$ as well!

Example

We show that the length function $f\colon \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x} \mapsto |\mathbf{x}|$ is differentiable at every point $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ with

$$\mathrm{d}f(\mathbf{x})(\mathbf{h}) = \frac{\mathbf{x} \cdot \mathbf{h}}{|\mathbf{x}|}.$$

For the proof we use Part 3 of the theorem.

$$\begin{aligned} f_{x_j}(x_1, \dots, x_n) &= \frac{\partial}{\partial x_j} \sqrt{x_1^2 + \cdots + x_j^2 + \cdots + x_n^2} \\ &= \frac{2x_j}{2\sqrt{x_1^2 + \cdots + x_n^2}} = \frac{x_j}{|\mathbf{x}|}. \end{aligned}$$

$\Longrightarrow$ The partial derivatives of f exist and are continuous on $\mathbb{R}^n \setminus \{\mathbf{0}\}$.
$\Longrightarrow$ f is differentiable on $\mathbb{R}^n \setminus \{\mathbf{0}\}$ with differential

$$\mathrm{d}f(\mathbf{x})(\mathbf{h}) = \left(\frac{x_1}{|\mathbf{x}|}, \dots, \frac{x_n}{|\mathbf{x}|} \right) \mathbf{h} = \frac{\mathbf{x} \cdot \mathbf{h}}{|\mathbf{x}|}, \quad \mathbf{h} \in \mathbb{R}^n.$$

Example

We compute the differential of the polar coordinate map

$$f(r,\phi)=\begin{pmatrix} r\cos\phi \\ r\sin\phi \end{pmatrix},\quad (r,\phi)\in\mathbb{R}^+\times\mathbb{R}.$$

$$\mathbf{J}_f(r,\phi)=\begin{pmatrix} \frac{\partial(r\cos\phi)}{\partial r} & \frac{\partial(r\cos\phi)}{\partial\phi} \\ \frac{\partial(r\sin\phi)}{\partial r} & \frac{\partial(r\sin\phi)}{\partial\phi} \end{pmatrix}=\begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix}$$

Since the entries of $\mathbf{J}_f(r,\phi)$ are continuous functions of (r,ϕ), the polar coordinate map f is differentiable in $\mathbb{R}^+\times\mathbb{R}$ with differential

$$\mathrm{d}f(r,\phi)(\mathbf{h})=\begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & r\cos\phi \end{pmatrix}\begin{pmatrix} h_1 \\ h_2 \end{pmatrix}=\begin{pmatrix} h_1\cos\phi-h_2r\sin\phi \\ h_1\sin\phi+h_2r\cos\phi \end{pmatrix}$$

for $\mathbf{h}=(h_1,h_2)^\mathsf{T}\in\mathbb{R}^2$.

Example (squaring map in $\mathbb{C}$)

Consider again the squaring map $f\colon \mathbb{C} \to \mathbb{C}$, $z \mapsto z^2$, i.e.,

$$f(z) = f(x+y\mathrm{i}) = (x+y\mathrm{i})^2 = x^2 - y^2 + 2xy\,\mathrm{i} = \begin{pmatrix} x^2 - y^2 \\ 2xy \end{pmatrix}.$$

We have

$$\mathbf{J}_f(x,y) = \begin{pmatrix} \frac{\partial(x^2-y^2)}{\partial x} & \frac{\partial(x^2-y^2)}{\partial y} \\ \frac{\partial(2xy)}{\partial x} & \frac{\partial(2xy)}{\partial y} \end{pmatrix} = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix},$$

$$\mathrm{d}f(x,y)(\mathbf{h}) = \begin{pmatrix} 2x & -2y \\ 2y & 2x \end{pmatrix}\begin{pmatrix} h_1 \\ h_2 \end{pmatrix} = \begin{pmatrix} 2xh_1 - 2yh_2 \\ 2yh_1 + 2xh_2 \end{pmatrix}.$$

Switching back to $\mathbb{C}$,

$$\mathrm{d}f(z)(h) = \mathrm{d}f(x+y\mathrm{i})(h_1+h_2\mathrm{i}) = 2(xh_1-yh_2)+2(yh_1+xh_2)\mathrm{i} = 2zh,$$

so that—as we have mentioned before—the (real) differential $\mathrm{d}f(z)$ is multiplication by the complex derivative $f'(z) = 2z$.

This holds more generally for complex functions whose complex derivative exists.

Example

Consider again the functions $f, g\colon \mathbb{R}^2 \setminus \{(0,0)\} \to \mathbb{R}$ defined by

$$f(x,y) = \frac{xy}{x^2+y^2}, \qquad g(x,y) = \frac{xy^2}{x^2+y^2}.$$

We extend f, g to $\mathbb{R}^2$ by defining $f(0,0) = g(0,0) = 0$. Then, as we have seen earlier, g is continuous in $(0,0)$ but f is not.

Reasoning as in the previous examples, one can easily show that f and g are differentiable in $\mathbb{R}^2 \setminus \{(0,0)\}$.
It turns out, however, that f and g are only partially but not totally differentiable at $(0,0)$. We show this for f and leave the case of g as a worksheet exercise.

By Part 1 of the theorem, since f is not continuous at $(0,0)$, it cannot be differentiable at $(0,0)$.

Since $f(x,0) = f(0,y) = f(0,0) = 0$ for all $x, y \in \mathbb{R}$, we have

$$f_x(0,0) = \lim_{h\to 0} \frac{f(h,0)-f(0,0)}{h} = 0, \quad f_y(0,0) = \lim_{h\to 0} \frac{f(0,h)-f(0,0)}{h} = 0,$$

i.e., the partial derivatives of f exist also at $(0,0)$.

Example (cont'd)

Why can Part 3 of the theorem not be applied here?

$$f_x(x,y)=\frac{y(x^2+y^2)-xy(2x)}{(x^2+y^2)^2}=\frac{y^3-x^2y}{(x^2+y^2)^2}$$

Substituting $y=mx$, $m\in\mathbb{R}$ fixed, gives

$$f_x(x,mx)=\frac{(mx)^3-mx^3}{(x^2+m^2x^2)^2}=\frac{m^3-m}{(1+m^2)^2x}\quad\text{for } x\in\mathbb{R}\setminus\{0\}.$$

$$\implies \lim_{x\to 0\pm} f_x(x,mx)=\pm\infty \text{ or } \mp\infty \quad \text{if } m\neq 0,\pm 1$$

and $\lim_{(x,y)\to(0,0)} f_x(x,y)$ does not exist (not even in the improper sense).

$\implies f_x$ is discontinuous at $(0,0)$.

Similarly, f_y is discontinuous at $(0,0)$.

Directional Derivatives

Definition

Suppose $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, is a function, $\mathbf{x} \in D^\circ$ and $\mathbf{u} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$. The *derivative* of f at $\mathbf{x}$ in the direction $\mathbf{u}$ is defined as

$$f_{\mathbf{u}}(\mathbf{x}) = \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{u}) - f(\mathbf{x})}{t}.$$

Other notations in use for $f_{\mathbf{u}}(\mathbf{x})$ are $\mathrm{D}_{\mathbf{u}} f(\mathbf{x})$ and $\frac{\partial f}{\partial \mathbf{u}}(\mathbf{x})$.

Notes

- Partial derivatives form a special case of directional derivatives: $\frac{\partial f}{\partial x_j}(\mathbf{x}) = f_{\mathbf{e}_j}(\mathbf{x})$.
- In general, $f_{\mathbf{u}}(\mathbf{x})$ is equal to the derivative at $t = 0$ of the function $t \mapsto f(\mathbf{x}+t\mathbf{u})$, which describes the behaviour of f on the line $\mathbf{x} + \mathbb{R}\mathbf{u}$. Note, however, that different choices of the direction vector for this line result in different values of the directional derivative (except in the case $f_{\mathbf{u}}(\mathbf{x}) = 0$).
- For functions $f\colon D \to \mathbb{R}$ (i.e., $m = 1$), the quantity $f_{\mathbf{u}}(\mathbf{x})$ measures the slope of G_f at $\mathbf{x}$ in the direction $\mathbf{u}$; equality holds in the case $|\mathbf{u}| = 1$.

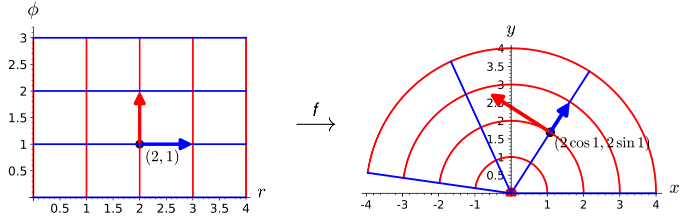
Notes cont'd

- If f is differentiable at $\mathbf{x}$, then all directional derivatives $f_{\mathbf{u}}(\mathbf{x})$, $\mathbf{u} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, exist and are obtained as $f_{\mathbf{u}}(\mathbf{x}) = \mathrm{d}f(\mathbf{x})(\mathbf{u})$.
  This follows from
  $$f(\mathbf{x} + t\mathbf{u}) - f(\mathbf{x}) = L(t\mathbf{u}) + \mathrm{o}(t\mathbf{u}) = t\,L(\mathbf{u}) + \mathrm{o}(t)$$
  for $t \to 0$, where $L = \mathrm{d}f(\mathbf{x})$.
- Returning to the case $m = 1$, the slope of G_f at $\mathbf{x}$ is maximized if the direction vector $\mathbf{u}$ (assumed to have unit length) is taken as a positive multiple of $\mathbf{J}_f(\mathbf{x})$ (and minimized for negative multiples). This follows from
  $f_{\mathbf{u}}(\mathbf{x}) = \mathbf{J}_f(\mathbf{x})\mathbf{u} = \mathbf{J}_f(\mathbf{x})^{\mathsf{T}} \cdot \mathbf{u} = \left|\mathbf{J}_f(\mathbf{x})^{\mathsf{T}}\right| |\mathbf{u}| \cos\theta = \left|\mathbf{J}_f(\mathbf{x})^{\mathsf{T}}\right| \cos\theta$.
- The preceding theorem has a coordinate-independent generalization, which uses directional derivatives.
  For example, Part 3 generalizes to:
  Suppose there exists a basis $\mathbf{u}_1, \dots, \mathbf{u}_n$ of $\mathbb{R}^n$ such that the directional derivatives $f_{\mathbf{u}_j}(\mathbf{x})$, $1 \leq j \leq n$, exist and are continuous at $\mathbf{x}$. Then f is differentiable at $\mathbf{x}$, and
  $\mathrm{d}f(\mathbf{x})\left(\sum_{j=1}^n h_j \mathbf{u}_j\right) = \sum_{j=1}^n f_{\mathbf{u}_j}(\mathbf{x}) h_j \quad$ *for $(h_1, \dots, h_n) \in \mathbb{R}^n$.*

Remark (columns of the Jacobi matrix)

For a differentiable map $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, the vectorial partial derivatives $\frac{\partial f}{\partial x_j}(\mathbf{x}) = \lim_{t\to 0} \frac{f(\mathbf{x}+t\mathbf{e}_j)-f(\mathbf{x})}{t}$ provide tangent vectors to the curves $t \mapsto f(\mathbf{x}+t\mathbf{e}_j)$ (images of the coordinate lines under f) in $\mathbf{x}$.

For example, the polar coordinate map $f(r,\phi) = \begin{pmatrix} r\cos\phi \\ r\sin\phi \end{pmatrix}$, which has $\mathbf{J}_f(r,\phi) = \begin{pmatrix} \cos\phi & -r\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}$, maps the two coordinate lines through $(2,1)$ to curves through $f(2,1) = (2\cos 1, 2\sin 1)$ with tangent vectors $\begin{pmatrix} \cos 1 \\ \sin 1 \end{pmatrix}$, respectively, $\begin{pmatrix} -2\sin 1 \\ 2\cos 1 \end{pmatrix}$.



In general the tangent vector of an image curve is obtained by applying the differential at the corresponding point to the tangent vector of the original curve (in our case $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ resp. $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$); cf. next lecture.

The Gradient

We consider a real-valued function $f\colon \mathbb{R}^n \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$. Suppose f is differentiable at $\mathbf{x}$.

Observation

For $\mathbf{h} \in \mathbb{R}^n$ (represented as a column vector) we have

$$\begin{aligned} \mathrm{d}f(\mathbf{x})(\mathbf{h}) = \mathbf{J}_f(\mathbf{x})\mathbf{h} &= \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}) h_j \\ &= \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{pmatrix} \cdot \begin{pmatrix} h_1 \\ \vdots \\ h_n \end{pmatrix}, \end{aligned}$$

i.e., the differential $\mathrm{d}f(\mathbf{x})$ is "taking the dot product with the column vector $\mathbf{J}_f(\mathbf{x})^\mathsf{T}$ of partial derivatives".

Definition

The column vector $\mathbf{J}_f(\mathbf{x})^\mathsf{T} = \left(\frac{\partial f}{\partial x_1}(\mathbf{x}), \dots, \frac{\partial f}{\partial x_n}(\mathbf{x})\right)^\mathsf{T} \in \mathbb{R}^n$ is called *gradient* of f at $\mathbf{x}$ and denoted by $\nabla f(\mathbf{x})$ or $\operatorname{grad} f(\mathbf{x})$.

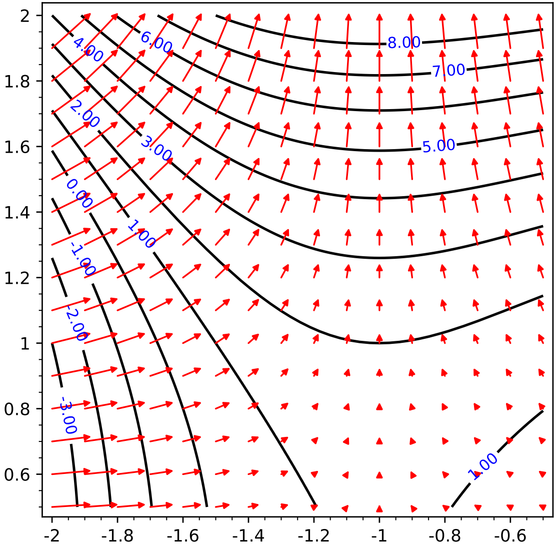


Figure: Contours of $f(x,y) = 2x^3 + 3x^2 + y^3$ and gradients $\nabla f(x,y) = (6x^2 + 6x, 3y^2)^{\mathsf{T}}$ scaled by 0.01

Notes

- In terms of the gradient, the approximation property of the differential takes the form $f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot \mathbf{h} + \mathrm{o}(\mathbf{h})$ for $\mathbf{h} \to \mathbf{0}$. Here $\mathbf{x}$ and $\mathbf{h}$ are viewed as column vectors.
- $\nabla f(\mathbf{x})$ contains of course the same information as $\mathbf{J}_f(\mathbf{x})$, but it "lives" in the ambient space of D and interacts with the points in D through vector arithmetic; cf. the picture.
- The gradient $\nabla f(\mathbf{x})$ points into the direction of the steepest ascent of the graph G_f at $\mathbf{x}$.
  More precisely, the slope m of the one-variable function $t \mapsto f(\mathbf{x}+t\mathbf{u})$, $|\mathbf{u}| = 1$, at $t = 0$ is maximized for $\mathbf{u} = \frac{\nabla f(\mathbf{x})}{|\nabla f(\mathbf{x})|}$, as follows from $m = f_{\mathbf{u}}(\mathbf{x}) = \nabla f(\mathbf{x}) \cdot \mathbf{u} = |\nabla f(\mathbf{x})|\,|\mathbf{u}| \cos\theta$.
  The maximal slope is $|\nabla f(\mathbf{x})|$.
- The gradient $\nabla f(\mathbf{x})$ is perpendicular to the contour of f through $\mathbf{x}$ (provided that the contour admits a parametrization that is smooth at $\mathbf{x}$); see the corollary to the Chain Rule.

Tangent Spaces

Suppose $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, is differentiable at $\mathbf{x}_0$. Setting $\mathbf{x} = \mathbf{x}_0 + \mathbf{h}$, the approximation property can be written as

$$f(\mathbf{x}) = f(\mathbf{x}_0) + L(\mathbf{x} - \mathbf{x}_0) + \mathrm{o}(\mathbf{x} - \mathbf{x}_0) \quad \text{for } \mathbf{x} \to \mathbf{x}_0.$$

This says that the affine map $A\colon \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{x} \mapsto f(\mathbf{x}_0) + L(\mathbf{x} - \mathbf{x}_0)$, $L = \mathrm{d}f(\mathbf{x}_0)$, approximates f very well near $\mathbf{x}_0$. The graph G_A appears to touch G_f in $\big(\mathbf{x}_0, f(\mathbf{x}_0)\big)$.

Definition

The graph G_A is called *(affine) tangent space* of G_f at $\mathbf{x}_0$.

Notes

- In parametric form the tangent space is given as

$$\begin{aligned}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} &= \begin{pmatrix}\mathbf{x}\\ \mathbf{y}_0 + \mathbf{A}(\mathbf{x} - \mathbf{x}_0)\end{pmatrix} = \begin{pmatrix}\mathbf{0}\\ \mathbf{y}_0 - \mathbf{A}\mathbf{x}_0\end{pmatrix} + \begin{pmatrix}\mathbf{I}_n\\ \mathbf{A}\end{pmatrix}\mathbf{x}\\ &= \begin{pmatrix}\mathbf{x}_0\\ \mathbf{y}_0\end{pmatrix} + \begin{pmatrix}\mathbf{I}_n\\ \mathbf{A}\end{pmatrix}\mathbf{h}, \qquad \text{(Subst. } \mathbf{x} = \mathbf{x}_0 + \mathbf{h}\text{)}\end{aligned}$$

  where $\mathbf{y}_0 = f(\mathbf{x}_0)$ and $\mathbf{A} = \mathbf{J}_f(\mathbf{x}_0)$. It follows that $\dim(\mathrm{G}_A) = n$.

Notes cont'd

- An equational form for the tangent space is $\mathbf{y} - \mathbf{y}_0 = \mathbf{A}(\mathbf{x} - \mathbf{x}_0)$ or, as a proper linear system of equations,

$$\begin{pmatrix} -\mathbf{A} & \mathbf{I}_n \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \mathbf{y}_0 - \mathbf{A}\mathbf{x}_0.$$

  In the case $m = 1$ the tangent space is a hyperplane of $\mathbb{R}^{n+1}$ (*tangent hyperplane*). It is then defined by the single equation $y - y_0 = \mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$, or

$$x_{n+1} = f(\mathbf{x}^{(0)}) + \sum_{j=1}^{n} \frac{\partial f}{\partial x_j}(\mathbf{x}^{(0)})(x_j - x_j^{(0)}),$$

  where we have written y as x_{n+1} and, in order to avoid double subscripts, $\mathbf{x}_0$ as $\mathbf{x}^{(0)} = (x_1^{(0)}, \dots, x_n^{(0)})$.
  If f is a function of two variables, we can use x, y, z-notation instead. The *tangent plane* to the graph of f in (x_0, y_0, z_0), $z_0 = f(x_0, y_0)$ is then given by

$$z = z_0 + f_x(x_0, y_0)(x - x_0) + f_y(x_0, y_0)(y - y_0).$$

Example

We determine the tangent plane of the parabolic cylinder P in $\mathbb{R}^3$ with equation $z = x^2 + y^2$ in the point $(1,1,2) \in P$.

P is the graph of $f\colon \mathbb{R}^2 \to \mathbb{R}$, $(x,y) \mapsto x^2 + y^2$.
We compute $f_x(x,y) = 2x$, $f_y(x,y) = 2y$ and hence
$\mathbf{J}_f(x,y) = (2x, 2y)$.
$\Longrightarrow$ An equation for the tangent plane of P in $(1,1,2)$ is

$$\begin{aligned} z &= f(1,1) + f_x(1,1)(x-1) + f_y(1,1)(y-1) \\ &= 2 + 2(x-1) + 2(y-1) = 2x + 2y - 2. \end{aligned}$$

The corresponding parametric form is

$$\begin{aligned} \begin{pmatrix} x \\ y \\ z \end{pmatrix} &= \begin{pmatrix} x \\ y \\ 2x+2y-2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ -2 \end{pmatrix} + x \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + y \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix} \\ &= \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix} + h_1 \begin{pmatrix} 1 \\ 0 \\ 2 \end{pmatrix} + h_2 \begin{pmatrix} 0 \\ 1 \\ 2 \end{pmatrix}. \end{aligned}$$

Generalization

Our earlier definition of the tangent line to a (smooth) parametric curve doesn't fit the present definition of "tangent space", since curves usually are not specified as graphs of maps $D \to \mathbb{R}^{n-1}$, $D \subseteq \mathbb{R}$.

Definition (Parametric surface)

A map $g\colon D \to \mathbb{R}^n$, $D \subseteq \mathbb{R}^k$, is called a parametric surface in $\mathbb{R}^n$. The parametric surface is said to be *smooth* and of *dimension* k if g is differentiable and $\mathbf{J}_g(\mathbf{x})$ has full column rank k for all $\mathbf{x} \in D$.

Just like a parametric curve, a parametric surface has a geometric object associated with it, viz. the range $g(D)$. This is called a non-parametric surface.

Parametric surfaces will be discussed later in more detail, when we do surface integration.

For a parametric surface we can define the tangent space in a way similar to the definition of the tangent line to a parametric curve. If g is differentiable in $\mathbf{x}_0$, the distance between $g(\mathbf{x})$ and the linear approximation $\mathbf{x} \mapsto g(\mathbf{x}_0) + \mathbf{J}_g(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ near $\mathbf{x}_0$ is much smaller than $|\mathbf{x} - \mathbf{x}_0|$, so that the range of the linear approximation appears to touch the surface in $g(\mathbf{x}_0)$.

Definition

Suppose $g\colon D \to \mathbb{R}^n$, $D \subseteq \mathbb{R}^k$, is differentiable in $\mathbf{x}_0 \in D$ and $\operatorname{rk} \mathbf{J}_g(\mathbf{x}_0) = k$. Then the range

$$T = \left\{g(\mathbf{x}_0) + \mathbf{J}_g(\mathbf{x}_0)\mathbf{h}; \mathbf{h} \in \mathbb{R}^k\right\}$$

of the linear approximation of g in $\mathbf{x}_0$ is called *(affine) tangent space* of g in $\mathbf{x}_0$, or of the non-parametric surface $g(D)$ associated with g.

By assumption, T is a k-dimensional affine subspace of $\mathbb{R}^n$, whose associated linear subspace ("direction space") is the column space of $\mathbf{J}_g(\mathbf{x}_0)$.

The last part of the definition actually requires justification (which is omitted): Show that if a non-parametric surface S is represented as $S = g_1(D_1) = g_2(D_2)$ with parametrizations g_1, g_2 then the tangent spaces T_1, T_2 of g_1, g_2 at $\mathbf{y} = g_1(\mathbf{x}_1) = g_2(\mathbf{x}_2)$ according to the previous definition are equal.

Example

Consider the (non-parametric) surface

$$S = \{(u+v, u^2+v^2, u^3+v^3); u, v, \in \mathbb{R}\}.$$

Here we have

$$g(u,v) = \begin{pmatrix} u+v \\ u^2+v^2 \\ u^3+v^3 \end{pmatrix}, \quad \mathbf{J}_g(u,v) = \begin{pmatrix} 1 & 1 \\ 2u & 2v \\ 3u^2 & 3v^2 \end{pmatrix}.$$

Transforming $\mathbf{J}_g(u,v)$ into column-echelon form, viz.
$\begin{pmatrix} 1 & 0 \\ 2u & 2(u-v) \\ 3u^2 & 3(u-v)(u+v) \end{pmatrix}$, shows that $\mathbf{J}_g(u,v)$ has rank 2 if $u \neq v$.
The points on S with $u = v$, i.e., $g(u,u) = (2u, 2u^2, 2u^3)$ form a twisted cubic (enlarged by the factor 2), and S (which one may call "twisted cubic surface") appears to have this curve C as a 1-dimensional boundary; cf. subsequent picture. Removing this curve from S results in a smooth 2-dimensional parametric surface, which can be parametrized bijectively by restricting the domain of g to $\{(u,v) \in \mathbb{R}^2; u > v\}$, say.

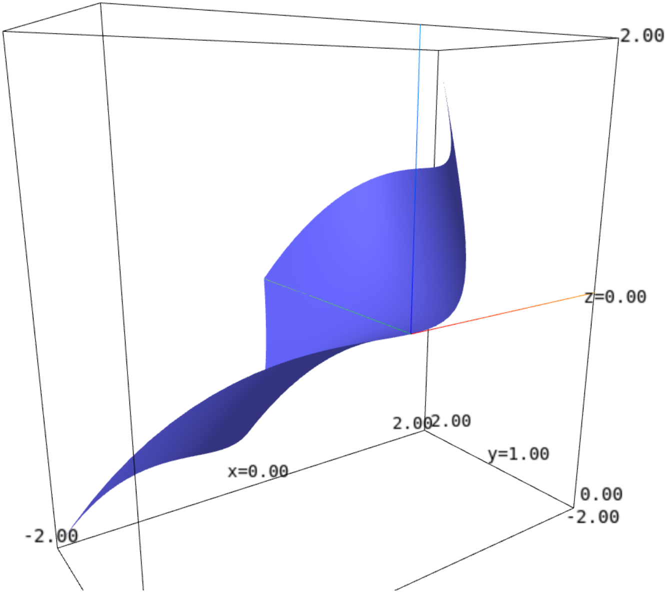


Figure: The twisted-cubic surface with parameter domain restricted to $[-1, 1]^2$

Example (cont'd)

As an example for computing tangent planes we consider the point $g(1,0)=(1,1,1)$. Since $\mathbf{J}_g(1,0)=\left(\begin{smallmatrix}1&1\\2&0\\3&0\end{smallmatrix}\right)$, the tangent plane to S in $(1,1,1)$ has parametric form

$$\begin{pmatrix}1\\1\\1\end{pmatrix}+\mathbb{R}\begin{pmatrix}1\\2\\3\end{pmatrix}+\mathbb{R}\begin{pmatrix}1\\0\\0\end{pmatrix}.$$

and equation $3y-2z=1$.

Exercise

1. Show that the restriction of $g(u,v)=(u+v,u^2+v^2,u^3+v^3)$ maps $\{\{(u,v)\in\mathbb{R}^2; u>v\}$ bijectively onto $S\setminus C$.
2. Show that S can be represented as graph of a function $z=h(x,y)$, compute the tangent plane to G_h in $(x,y,z)\in S$ according to our earlier definition, and verify that both definitions yield the same tangent planes.

   Hint: Eliminate u,v from $x=u+v$, $y=u^2+v^2$, $z=u^3+v^3$.

What are $\mathrm{d}x$, $\mathrm{d}x_j$, $\mathrm{d}\mathbf{x}$, $\mathrm{d}z$?

Have you ever wondered what "$\mathrm{d}x$" in the notation for integrals, e.g., in $\int_0^1 x^2\,\mathrm{d}x$ means?

Answer: $\mathrm{d}x$ and its friends are differentials.

- $\mathrm{d}x$ denotes the differential of $\mathbb{R}\to\mathbb{R}$, $x\mapsto x$ (the identity map $\mathrm{id}_\mathbb{R}$ of $\mathbb{R}$), which is $\mathbb{R}\to\mathbb{R}$, $x\mapsto \mathrm{id}_\mathbb{R}$. Thus $\mathrm{d}x(h)=\mathrm{d}x(x_0)(h)=h$ for all $x_0\in\mathbb{R}$ and $h\in\mathbb{R}$.
  Moreover, $\mathrm{d}f(x)=f'(x)\mathrm{d}x$ in the sense that $\mathrm{d}f(x)(h)=f'(x)h=f'(x)\mathrm{d}x(h)$ for all x at which f is differentiable and all $h\in\mathbb{R}$.
- $\mathrm{d}x_j$ denotes the differential of the coordinate projection $\mathbb{R}^n\to\mathbb{R}^n$, $\mathbf{x}\mapsto x_j$, which is given by $\mathrm{d}x_j(\mathbf{h})=\mathrm{d}x_j(\mathbf{x}_0)(\mathbf{h})=h_j$ for all $\mathbf{x}_0\in\mathbb{R}^n$ and $\mathbf{h}\in\mathbb{R}^n$.
  Moreover, $\mathrm{d}f=\sum_{j=1}^n\frac{\partial f}{\partial x_j}\,\mathrm{d}x_j$ in the sense that if f is differentiable at $\mathbf{x}$ then

$$\mathrm{d}f(\mathbf{x})(\mathbf{h})=\sum_{j=1}^n\frac{\partial f}{\partial x_j}(\mathbf{x})\,\mathrm{d}x_j(\mathbf{h})\quad\text{for }\mathbf{h}\in\mathbb{R}^n.$$

Answer (cont'd)

- $\mathrm{d}\mathbf{x}$ is the differential of $\mathbb{R}^n \to \mathbb{R}^n$, $\mathbf{x} \mapsto \mathbf{x}$ (the identity on $\mathbb{R}^n$), which is given by $\mathrm{d}\mathbf{x}(\mathbf{h}) = \mathrm{d}\mathbf{x}(\mathbf{x}_0)(\mathbf{h}) = \mathbf{h}$ for all $\mathbf{x}_0 \in \mathbb{R}^n$, $\mathbf{h} \in \mathbb{R}^n$.
- $\mathrm{d}z$ is the differential of $\mathbb{C} \to \mathbb{C}$, $z \mapsto z$ (the identity on $\mathbb{C}$) and thus equal to $\mathrm{d}\mathbf{x}$ for $n = 2$, provided that $\mathbb{C}$ is identified with $\mathbb{R}^2$. We have $\mathrm{d}f(z) = f'(z)\mathrm{d}z$ for those z in the domain of f for which the complex derivative $f'(z) = \lim_{h\to 0} \frac{f(z+h)-f(z)}{h}$ exists.

Caution

In [Ste16] the differential $\mathrm{d}f$ of a two-variable scalar function $z = f(x, y)$ is sometimes denoted by $\mathrm{d}z$ as well. In the lecture we won't use $\mathrm{d}z$ in this sense. Instead, when writing $x = x_1$, $y = y_2$, $z = x_3$, we denote the corresponding differentials by $\mathrm{d}x$, $\mathrm{d}y$, and $\mathrm{d}z$. So in the lecture, $\mathrm{d}z$ has two different meanings: (i) $\mathrm{d}z = \mathrm{d}x_3$; (ii) the complex differential $\mathrm{d}z = \mathrm{d}x + \mathrm{i}\,\mathrm{d}y$.

The differentials $\mathrm{d}\mathbf{x}$, $\mathrm{d}z$, which belong to vector-valued functions, will be rarely used in the sequel. The coordinate differentials $\mathrm{d}x_j$, or $\mathrm{d}x$, $\mathrm{d}y$, dz, however, are so convenient to use (and will be used frequently).

Example

Consider $f(x,y) = x^3 - 3xy^2$ defined on $D = \mathbb{R}^2$.

$$f_x(x,y) = 3x^2 - 3y^2, \quad f_y(x,y) = -6xy,$$

$$\mathbf{J}_f(x,y) = \begin{pmatrix} 3x^2 - 3y^2 & -6xy \end{pmatrix}, \quad \nabla f(x,y) = \begin{pmatrix} 3x^2 - 3y^2 \\ -6xy \end{pmatrix}$$

$$\mathrm{d}f = f_x \,\mathrm{d}x + f_y \,\mathrm{d}y = (3x^2 - 3y^2)\,\mathrm{d}x - 6xy\,\mathrm{d}y$$

The purpose of the differential is to approximate

$$f(x+h_1, y+h_2) - f(x,y) \approx \mathrm{d}f(x,y)(h_1,h_2) = (3x^2 - 3y^2)h_1 - 6xyh_2$$

for small h_1, h_2.

Mnemonic: In order to apply $\mathrm{d}f(x,y)$ to $\mathbf{h} = (h_1, h_2)$, substitute the coordinates h_1, h_2 for $\mathrm{d}x$, $\mathrm{d}y$ in the expression $\mathrm{d}f = (3x^2 - 3y^2)\,\mathrm{d}x - 6xy\,\mathrm{d}y$.

The Multivariable Chain Rule

generalizing $(g \circ f)'(x) = g'(y)f'(x)$

Theorem

Suppose $f\colon D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^n$, is differentiable in $\mathbf{x}_0 \in D$, $g\colon E \to \mathbb{R}^p$, $E \subseteq \mathbb{R}^m$, is differentiable in $\mathbf{y}_0 \in E$, and f satisfies $f(D) \subseteq E$, $f(\mathbf{x}_0) = \mathbf{y}_0$. Then $g \circ f\colon D \to \mathbb{R}^p$ is differentiable in $\mathbf{x}_0$, and its differential satisfies

$$\mathrm{d}(g \circ f)(\mathbf{x}_0) = \mathrm{d}g(\mathbf{y}_0) \circ \mathrm{d}f(\mathbf{x}_0).$$

In other words, the differential of a composition is the composition of the differentials (evaluated at the respective points).

Proof.

Writing $L = \mathrm{d}f(\mathbf{x}_0)$, $M = \mathrm{d}g(\mathbf{y}_0)$, we have $L\colon \mathbb{R}^n \to \mathbb{R}^m$, $M\colon \mathbb{R}^m \to \mathbb{R}^p$, $M \circ L\colon \mathbb{R}^n \to \mathbb{R}^p$ and must show that $\mathrm{d}(g \circ f)(\mathbf{x}_0) = M \circ L$ (hence at least the dimensions match).

The differentiability conditions say that for $\mathbf{h} \to \mathbf{0}$, $\mathbf{k} \to \mathbf{0}$,

$$\begin{aligned} f(\mathbf{x}_0 + \mathbf{h}) &= f(\mathbf{x}_0) + L(\mathbf{h}) + \phi(\mathbf{h}), \quad \phi(\mathbf{h}) = \mathrm{o}(\mathbf{h}), \\ g(\mathbf{y}_0 + \mathbf{k}) &= g(\mathbf{y}_0) + M(\mathbf{k}) + \psi(\mathbf{k}), \quad \psi(\mathbf{k}) = \mathrm{o}(\mathbf{k}). \end{aligned}$$

Proof cont'd.

Hence

$$\begin{aligned} g\big(f(\mathbf{x}_0+\mathbf{h})\big) &= g\big(f(\mathbf{x}_0)+L(\mathbf{h})+\phi(\mathbf{h})\big) \\ &= g\big(f(\mathbf{x}_0)\big)+M\big(L(\mathbf{h})+\phi(\mathbf{h})\big)+\psi\big(L(\mathbf{h})+\phi(\mathbf{h})\big) \\ &= g\big(f(\mathbf{x}_0)\big)+M\big(L(\mathbf{h})\big)+M\big(\phi(\mathbf{h})\big)+\psi\big(L(\mathbf{h})+\phi(\mathbf{h})\big) \end{aligned}$$

for $\mathbf{h}\to\mathbf{0}$, and it remains to show that (i) $M\big(\phi(\mathbf{h})\big)=\mathrm{o}(\mathbf{h})$ for $\mathbf{h}\to\mathbf{0}$ and (ii) $\psi\big(L(\mathbf{h})+\phi(\mathbf{h})\big)=\mathrm{o}(\mathbf{h})$ for $\mathbf{h}\to\mathbf{0}$.

(i) This part is easy and can be done as follows:

$$\frac{M\big(\phi(\mathbf{h})\big)}{|\mathbf{h}|}=M\left(\frac{\phi(\mathbf{h})}{|\mathbf{h}|}\right)\to M(\mathbf{0})=\mathbf{0}\quad\text{for } \mathbf{h}\to\mathbf{0},$$

using the linearity of M, $\lim_{\mathbf{h}\to\mathbf{0}}\phi(\mathbf{h})/|\mathbf{h}|=\mathbf{0}$, and the continuity of M in $\mathbf{0}$.

Proof cont'd.

(ii) This part is more difficult. We have

$$\frac{\psi\big(L(\mathbf{h})+\phi(\mathbf{h})\big)}{|\mathbf{h}|}=\frac{\psi\big(L(\mathbf{h})+\phi(\mathbf{h})\big)}{|L(\mathbf{h})+\phi(\mathbf{h})|}\cdot\frac{|L(\mathbf{h})+\phi(\mathbf{h})|}{|\mathbf{h}|}.$$

The 1st factor tends to $\mathbf{0}\in\mathbb{R}^p$ for $\mathbf{h}\to\mathbf{0}$, since $L(\mathbf{h})+\phi(\mathbf{h})\to\mathbf{0}$ and $\lim_{\mathbf{k}\to\mathbf{0}}\psi(\mathbf{k})/|\mathbf{k}|=\mathbf{0}$.

The 2nd factor remains bounded for $\mathbf{h}\to\mathbf{0}$, since $|\phi(\mathbf{h})|/|\mathbf{h}|\to 0$ and

$$\begin{aligned}\frac{|L(\mathbf{h})|}{|\mathbf{h}|}&=\frac{|L(h_1\mathbf{e}_1+\cdots+h_n\mathbf{e}_n)|}{|\mathbf{h}|}=\frac{|h_1L(\mathbf{e}_1)+\cdots+h_nL(\mathbf{e}_n)|}{|\mathbf{h}|}\\&\leq|L(\mathbf{e}_1)|+\cdots+|L(\mathbf{e}_n)|.\end{aligned}$$

This shows $\psi\big(L(\mathbf{h})+\phi(\mathbf{h})\big)=\mathrm{o}(\mathbf{h})$ for $\mathbf{h}\to\mathbf{0}$ and completes the proof of the chain rule. $\square$

Corollary (chain rule in matrix form)

Under the assumptions of the theorem we have

$$\mathbf{J}_{g\circ f}(\mathbf{x}_0) = \mathbf{J}_g(\mathbf{y}_0)\mathbf{J}_f(\mathbf{x}_0).$$

Proof.

Use $L(\mathbf{h}) = \mathbf{J}_f(\mathbf{x}_0)\mathbf{h}$, $M(\mathbf{k}) = \mathbf{J}_g(\mathbf{y}_0)\mathbf{k}$, and the fact that the composition of two linear maps is represented by the product of the corresponding matrices. $\square$

Remark

If g is scalar-valued ($p = 1$), we have, suppressing arguments, $\mathbf{J}_g = \left(\frac{\partial g}{\partial y_1}, \dots, \frac{\partial g}{\partial y_m}\right)$ and similarly, writing $h = g\circ f$, $\mathbf{J}_h = \left(\frac{\partial h}{\partial x_1}, \dots, \frac{\partial h}{\partial x_n}\right)$. Since $\mathbf{J}_f = \left(\frac{\partial f_i}{\partial x_j}\right)$, the corollary gives

$$\frac{\partial h}{\partial x_j} = (\mathbf{J}_g\mathbf{J}_f)_j = \sum_{i=1}^{m} \frac{\partial g}{\partial y_i}\frac{\partial f_i}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial u}{\partial y_i}\frac{\partial y_i}{\partial x_j}.$$

In terms of the variables $y_i = f_i(x_1, \dots, x_n)$, $u = g(y_1, \dots, y_m) = = h(x_1, \dots, x_n)$ this can be written as $\frac{\partial u}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial u}{\partial y_i}\frac{\partial y_i}{\partial x_j}$, recovering the “general version” of the chain rule in [Ste16], Ch. 14.5, p. 940.

Example ([Ste16], Ch. 14.5, p. 940 bottom)

Here $w = g(x, y, z, t)$ is composed with $f(u, v) = \begin{pmatrix} x(u,v) \\ y(u,v) \\ z(u,v) \\ t(u,v) \end{pmatrix}$, and the chain rule

$$\begin{aligned} \mathbf{J}_{g\circ f}(u, v) &= \mathbf{J}_g\big(f(u, v)\big)\,\mathbf{J}_f(u, v) \\ &= \mathbf{J}_g\big(x, y, z, t\big)\,\mathbf{J}_f(u, v) \end{aligned}$$

takes the form

$$\begin{pmatrix} \dfrac{\partial w}{\partial u} & \dfrac{\partial w}{\partial v} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial w}{\partial x} & \dfrac{\partial w}{\partial y} & \dfrac{\partial w}{\partial z} & \dfrac{\partial w}{\partial t} \end{pmatrix} \begin{pmatrix} \dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v} \\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v} \\ \dfrac{\partial z}{\partial u} & \dfrac{\partial z}{\partial v} \\ \dfrac{\partial t}{\partial u} & \dfrac{\partial t}{\partial v} \end{pmatrix}.$$

The full story is not visible in this form, since the partial derivatives $\frac{\partial w}{\partial x}$, $\frac{\partial w}{\partial y}$, $\frac{\partial w}{\partial z}$, $\frac{\partial w}{\partial t}$ need to be composed with $f(u, v)$.

Corollary

Suppose $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^2$, is differentiable at $\mathbf{x} = (x_1, x_2)$ and the contour of f through $\mathbf{x}$, viz. $\mathrm{N}_f(k)$ with $k = f(\mathbf{x})$, admits a smooth parametrization g near $\mathbf{x}$. Then $\nabla f(\mathbf{x})$ is perpendicular to the tangent line of $\mathrm{N}_f(k)$ at $\mathbf{x}$.

Proof.

By assumption, there exists $g\colon (a,b) \to D$ and $t_0 \in (a,b)$ such that $g(t_0) = \mathbf{x}$, $g'(t_0) \neq \mathbf{0} \in \mathbb{R}^2$, and $f\bigl(g(t)\bigr) = k$ for $t \in (a,b)$. The Chain Rule gives

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} f\bigl(g(t)\bigr) = \mathrm{d}(f \circ g)(t)(1) = \nabla f\bigl(g(t)\bigr) \cdot g'(t).$$

Plugging in $g(t_0) = \mathbf{x}$ and recalling that the tangent line to the curve at $g(t_0)$ is $g(t_0) + \mathbb{R}g'(t_0)$ completes the proof. $\square$

The proof shows that orthogonality holds at every point $g(t)$, $t \in (a,b)$, but this statement is of course equivalent to the corollary.

Notes on the Corollary

- The Implicit Function Theorem (a pretty advanced result, which is beyond the scope of this course) gives that for a C^1-function f (i.e., the partial derivatives of f exist and are continuous on D) the condition $\nabla f(\mathbf{x}_0) \neq (0,0)$ is sufficient for $\mathrm{N}_f(k)$ admitting a smooth parametrization near $\mathbf{x}_0$.
- In the n-variable case $f\colon D \to \mathbb{R}$, $D \subseteq \mathbb{R}^n$, level sets $\mathrm{N}_f(k)$ locally admit parametrizations by functions g of $n-1$ variables (provided f is C^1 and $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$) and form $(n-1)$-dimensional smooth parametric surfaces in $\mathbb{R}^n$. Reasoning as in the proof of the corollary then shows: The gradient $\nabla f(\mathbf{x}_0)$ is orthogonal to the columns of $\mathbf{J}_g(\omega_0)$, where $\mathbf{x}_0 = g(\omega_0)$. But the columns of $\mathbf{J}_g(\omega_0)$ generate the $n-1$-dimensional direction space of the tangent hyperplane T of g at ω_0 or, equivalently, of $\mathrm{N}_f(k)$ in $\mathbf{x}_0 = g(\omega_0)$. Thus $\nabla f(\mathbf{x}_0)$ serves as normal vector for T, which therefore has equation $\nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) = 0$; cp. [Ste16], Ch. 14.6, Eq. (19).
  The chain rule argument also shows that $\nabla f(\mathbf{x}_0)$ is orthogonal to every curve through $\mathbf{x}_0$ (i.e., orthogonal to its tangent L at $\mathbf{x}_0$) that is entirely contained in the corresponding level surface $\mathrm{N}_f(k)$, $k = f(\mathbf{x}_0)$. Under the asumption $\operatorname{rk} \mathbf{J}_g(\omega_0) = n-1$ this implies that L is contained in the tangent hyperplane of $\mathrm{N}_f(k)$.

Example

The sphere $x^2 + y^2 + z^2 = 9$ contains the point $(1, 2, 2)$. We compute the tangent plane in $(1, 2, 2)$ in two different ways:

1. The sphere is the 9-level surface of $f(x, y, z) = x^2 + y^2 + z^2$. Since $\nabla f(x, y, z) = (2x, 2y, 2z) = 2(x, y, z)$, we can take the point (x, y, z) itself as normal vector and obtain that the tangent plane has equation $(x - 1) + 2(y - 2) + 2(z - 2) = 0$.
2. The upper half sphere $x^2 + y^2 + z^2 = 9 \wedge z > 0$, which contains $(1, 2, 2)$, is parametrized by

$$g(x, y) = \left(x, y, \sqrt{9 - x^2 - y^2}\right), \quad x^2 + y^2 < 9.$$

$$\Longrightarrow \mathbf{J}_g(x, y) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -\frac{x}{\sqrt{9-x^2-y^2}} & -\frac{y}{\sqrt{9-x^2-y^2}} \end{pmatrix}, \ \mathbf{J}_g(1, 2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -\frac{1}{2} & -1 \end{pmatrix}$$

$\Longrightarrow$ A parametric form for the tangent plane is

$$\begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} + \mathbb{R} \begin{pmatrix} 1 \\ 0 \\ -\frac{1}{2} \end{pmatrix} + \mathbb{R} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}.$$

Suppose $f\colon D \to \mathbb{R}^n$, $D \subseteq \mathbb{R}^n$, is a C^1-function and $\mathbf{x}_0 \in D$ satisfies $\mathrm{rk}\bigl(\mathrm{d}f(\mathbf{x}_0)\bigr) = n$ (i.e., the linear map $\mathrm{d}f(\mathbf{x}_0)$ or, equivalently, the Jacobi matrix $\mathbf{J}_f(\mathbf{x}_0)$ has an inverse).
In this case the so-called *Inverse Function Theorem* shows that f itself has a C^1-inverse on a suitably restricted domain $D' \subseteq D$, i.e., there exists an open set $D' \subseteq D$ with $\mathbf{x}_0 \in D'$ and a C^1-function $g\colon E' \to D'$, $E' = f(D')$, such that $f \circ g = \mathrm{id}_{E'}$, $g \circ f = \mathrm{id}_{D'}$.

Corollary

Under the assumptions made above we have

$$\mathrm{d}g(\mathbf{y}) = \mathrm{d}f\bigl(g(\mathbf{y})\bigr)^{-1} \quad \textit{for} \quad \mathbf{y} \in E'$$

or, in terms of Jacobi matrices, $\mathbf{J}_g(\mathbf{y}) = \mathbf{J}_f\bigl(g(\mathbf{y})\bigr)^{-1}$ *for* $\mathbf{y} \in E'$*.*

Proof.

By assumption $g\bigl(f(\mathbf{x})\bigr) = \mathbf{x}$ for all $\mathbf{x} \in D'$. Applying the chain rule gives

$$\mathrm{d}g\bigl(f(\mathbf{x})\bigr) \circ \mathrm{d}f(\mathbf{x}) = \mathrm{d}\mathbf{x} = \mathrm{id}_{\mathbb{R}^n} \quad \text{for} \quad \mathbf{x} \in D'.$$

Similarly, using $f\bigl(g(\mathbf{y})\bigr) = \mathbf{y}$ one shows $\mathrm{d}f\bigl(g(\mathbf{y})\bigr) \circ \mathrm{d}g(\mathbf{y}) = \mathrm{id}_{\mathbb{R}^n}$ for all $\mathbf{y} \in E'$. This provides ample proof of the desired equality $\mathrm{d}g(\mathbf{y}) = \mathrm{d}f(\mathbf{x})^{-1}$, $\mathbf{y} = f(\mathbf{x})$. $\square$

The Chain Rule—A Concrete Example

Example (Differential of the length function re-examined)

The length function

$$h(\mathbf{x}) = |\mathbf{x}| = \sqrt{x_1^2 + \cdots + x_n^2} = \sqrt{\mathbf{x} \cdot \mathbf{x}}$$

is the composition of $f(\mathbf{x}) = \mathbf{x} \cdot \mathbf{x}$ and $g(y) = \sqrt{y}$. Since

$$\mathrm{d}f(\mathbf{x})(\mathbf{h}) = 2\,\mathbf{x}^{\mathsf{T}}\mathbf{h}, \quad g'(y) = \frac{1}{2\sqrt{y}},$$

the chain rule gives

$$\mathbf{J}_h(\mathbf{x}) = g'\big(f(\mathbf{x})\big)\mathbf{J}_f(\mathbf{x}) = \frac{2\,\mathbf{x}^{\mathsf{T}}}{2\sqrt{\mathbf{x} \cdot \mathbf{x}}} = \frac{\mathbf{x}^{\mathsf{T}}}{|\mathbf{x}|}, \quad \textit{or} \quad \nabla h(\mathbf{x}) = \frac{\mathbf{x}}{|\mathbf{x}|}.$$

In other words, the gradient field of h is radial and consists of unit vectors.

$\Longrightarrow$ The graph of h has radial slope 1 (except at the origin), i.e., it is (the surface of) a cone.