Chapter 2 Differentiation

Now that we have a good understanding of the Euclidean space $\mathbb{R}^n$, we are ready to discuss the concept of differentiation in multivariable calculus. We are going to deal with functions defined on one Euclidean space with values in another Euclidean space. We shall use the shorthand notation

$$\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$$

to describe such a situation. This means that f can be written in the form

$$f(x) = (f_1(x), f_2(x), \ldots, f_m(x)) \,,$$

where $x = (x_1, x_2, \ldots, x_n)$ and each *coordinate function* f_i is a real-valued function on $\mathbb{R}^n$. In these situations it may be that f is not defined on *all* of $\mathbb{R}^n$, but we'll continue with the above shorthand.

There are two important special cases: $n = 1$ and $m = 1$, respectively. We shall quickly see that the case $n = 1$ is much, much simpler than all other cases. We shall also learn that the case $m = 1$ already contains almost all the interesting mathematics that we investigate — the generalization to $m > 1$ will prove to be very easy indeed.

A. Functions of one real variable (n = 1)

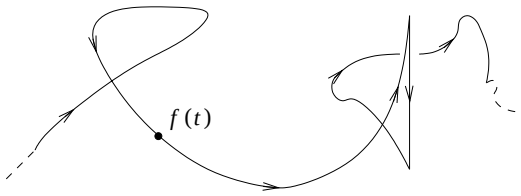
In the situation

$$\mathbb{R} \xrightarrow{f} \mathbb{R}^m$$

we shall typically denote the real numbers in the domain of f by the letter t, and the points in $\mathbb{R}^m$ in the usual manner by $x = (x_1, x_2, \ldots, x_m)$. As we mentioned above, we can represent f in terms of its coordinate functions:

$$f(t) = (f_1(t), f_2(t), \ldots, f_m(t)) \,.$$

This formula displays the vector $f(t)$ in terms of its coordinates, so that the function f can be regarded as comprised of m real-valued functions f_1, $f_2, \ldots, f_m$. We often like to think of real-valued functions in terms of their graphs, but when $m > 1$ this viewpoint seems somewhat cumbersome. A more useful way to think of f in these higher dimensions is to imagine the points $f(t)$ "plotted" in $\mathbb{R}^m$ with regard to the independent variable t. In case $f(t)$ depends continuously on t, the points $f(t)$ then form some sort of continuous curve in $\mathbb{R}^m$:



We have placed arrows on our picture to indicate the direction of increasing t. Thus the points $f(t)$ form a sort of "curve" (whatever that may mean) in $\mathbb{R}^m$.

We need to understand well the definition of *limit* as $t \to t_0$ and/or *continuity* at t_0. As we are somehow interested in the size of $f(t) - f(t_0)$, we can merely use the definition of continuity in the case of real-valued functions, modified so that instead of absolute value we use the norm. Thus we have the

DEFINITION. Let $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$. Then f is *continuous* at t_0 if for each $\epsilon > 0$ there exists $\delta > 0$ such that

$$|t - t_0| < \delta \Rightarrow \|f(t) - f(t_0)\| < \epsilon.$$

PROBLEM 2–1. Let $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$ and let $L \in \mathbb{R}^m$. Write out the correct definition of

$$\lim_{t \to t_0, t \neq t_0} f(t) = L.$$

Then prove that f is continuous at $t_0 \iff$

$$\lim_{t \to t_0, t \neq t_0} f(t) = f(t_0).$$

In preparation for the important characterization of continuity by means of the coordinate functions, work the following

PROBLEM 2–2. Prove that for all $x \in \mathbb{R}^m$

$$\max_{1 \leq i \leq m} |x_i| \leq \|x\| \leq \sqrt{m} \max_{1 \leq i \leq m} |x_i|.$$

THEOREM. *Let $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$. Then f is continuous at $t_0 \iff$ all the coordinate functions f_1, $f_2, \ldots, f_m$ are continuous at t_0.*

PROOF. $\Longrightarrow$: Let $\epsilon > 0$. Then the continuity of f guarantees that there exists $\delta > 0$ such that

$$|t - t_0| < \delta \Longrightarrow \|f(t) - f(t_0)\| < \epsilon.$$

By Problem 2–2, conclude that

$$|t - t_0| < \delta \Longrightarrow |f_i(t) - f_i(t_0)| < \epsilon$$

for each i. Thus each f_i is continuous at t_0.

PROBLEM 2–3. Write out in *careful* detail the proof of the other half ($\Longleftarrow$) of the theorem.

QED

Though this result reduces continuity to that of real-valued functions, we prefer the original definition in terms of the norm of $f(t) - f(t_0)$. For that definition is more "geometric" and does not involve the coordinates of $\mathbb{R}^m$ at all.

We shall always assume that $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$ is at least continuous and defined on an interval in $\mathbb{R}$, which could be all of $\mathbb{R}$ itself.

DEFINITION. A *curve* in $\mathbb{R}^m$ is a continuous function f from an interval $[a, b]$ into $\mathbb{R}^m$. The independent variable $a \leq t \leq b$ is sometimes called a *parameter* for the curve.

Here are some specific examples:

- Straight line: $f(t) = x + tv$, where $x, v \in \mathbb{R}^n$, $v \neq 0$.
- Unit circle in $\mathbb{R}^2$: $f(t) = (\cos t, \sin t)$.
- Unit circle in $\mathbb{R}^2$: $f(t) = (\cos t, -\sin t)$.
- Helix in $\mathbb{R}^3$: $f(t) = (\cos t, \sin t, t)$.
- A curve in $\mathbb{R}^2$: $f(t) = (t^2, t^3)$.

REMARK. Our definition of "curve" is perhaps somewhat unusual. Normally we think of a curve in $\mathbb{R}^m$ as some sort of subset of $\mathbb{R}^m$ which has a continuous one-dimensional shape. This would correspond to the set which is the *image* of f in our actual definition,

$$\{f(t) \mid a \leq t \leq b\}.$$

However, it seems best to keep our definition, which provides the extra information of a parameter t for the curve. It might be better to call the function f a *parametrized curve*, but that just seems too cumbersome.

Now we turn to the basic definition which introduces calculus in this context.

DEFINITION. Let $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$. Then f is *differentiable* at t_0 if the following limit exists:

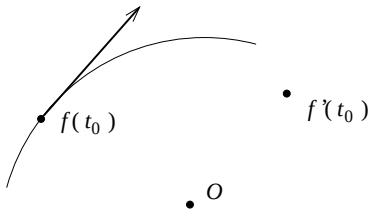
$$f'(t_0) = \frac{df}{dt}(t_0) = \lim_{t \to t_0, t \neq t_0} \frac{f(t) - f(t_0)}{t - t_0}.$$

In terms of coordinates for $\mathbb{R}^m$ the result is like that for continuity, namely, f is differentiable $\iff$ all the coordinate functions are differentiable. Moreover,

$$\frac{d}{dt}(f_1, \ldots, f_m) = \left(\frac{df_1}{dt}, \ldots, \frac{df_m}{dt} \right).$$

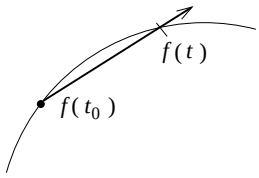
We also say f is *differentiable* if it is differentiable at t_0 for every t_0. This coordinatewise calculation of the derivative $f'(t_0)$ is valid because of the corresponding theorem we proved above.

A helpful way to visualize $f'(t_0)$ is to draw the "arrow" from 0 to $f'(t_0)$. In drawing this picture we like to position the vector $f'(t_0)$ so that its "tail" is at $f(t_0)$:



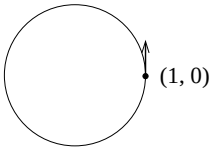
We then employ the phrase *tangent vector* at $t = t_0$. This nice geometrical picture is connected with the "finite" picture of the *secant* vector

$$\frac{f(t) - f(t_0)}{t - t_0} :$$



Thus the tangent vector at $t = t_0$ is the limit of the secant vector as $t \to t_0$.

EXAMPLE. $f(t) = \left(\cos\frac{t}{2}, \sin\frac{t}{2}\right)$. Here $f'(0) = (0, \frac{1}{2})$.



PROBLEM 2–4. Consider the circle in $\mathbb{R}^2$ described as

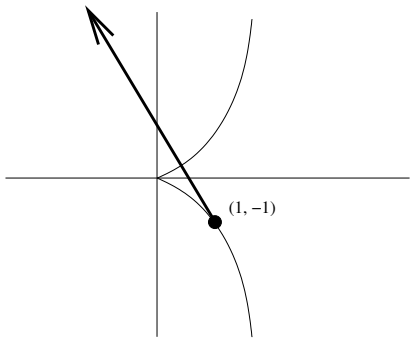
$$f(t) = (a_1 + r\cos\alpha t,\ a_2 + r\sin\alpha t)\,.$$

This is a parametrization of what we have denoted by $S(a, r)$ in Section 1F. Here $\alpha \neq 0$ is a real constant. Prove that

$$f'(t) \bullet (f(t) - a) = 0.$$

What is the geometrical interpretation of this result?

EXAMPLE. $f(t) = (t^2, t^3)$. Here $f'(-1) = (-2, 3)$.



Notice that in this example we have $f'(0) = (0, 0)$. In a sense this explains why the image in $\mathbb{R}^2$ has a nonsmooth appearance at the origin though the curve f is differentiable.

PROBLEM 2–5. Let $f(t) = (t^2, t|t|)$ for $-\infty < t < \infty$. Show that f is differentiable and sketch its image in $\mathbb{R}^2$.

PROBLEM 2–6. Consider the "figure 8 curve" given as the set of points $(x, y) \in \mathbb{R}^2$ which satisfy the equation

$$(x^2 + y^2)^2 = x^2 - y^2.$$

a. Sketch this set reasonably accurately (you might use the corresponding polar coordinate equation $r^2 = \cos 2\theta$).

b. Show that the curve

$$f(t) = \left(\frac{\cos t}{1 + \sin^2 t}, \frac{\sin t \cos t}{1 + \sin^2 t} \right), \quad 0 \leq t < 2\pi,$$

is a parametrization with the feature that $f(s) = f(t) \Rightarrow s = t$ or $s = \frac{\pi}{2}$, $t = \frac{3\pi}{2}$ or $s = \frac{3\pi}{2}$, $t = \frac{\pi}{2}$.

c. From (b) we have $f\left(\frac{\pi}{2}\right) = f\left(\frac{3\pi}{2}\right) = (0, 0)$. Show that

$$\begin{aligned} f'\left(\frac{\pi}{2}\right) &= \left(-\frac{1}{2}, -\frac{1}{2}\right), \\ f'\left(\frac{3\pi}{2}\right) &= \left(\frac{1}{2}, -\frac{1}{2}\right). \end{aligned}$$

d. Conclude that f is differentiable at every t, and that it provides two distinct tangent vectors at the geometric point $(0, 0)$ on the original figure 8 curve.

PROBLEM 2–7. Sketch the set of points in $\mathbb{R}^2$ described by the equation $y^2 = x^2(x + 1)$. Show that the curve

$$f(t) = (t^2 - 1, t^3 - t), \qquad -\infty < t < \infty,$$

gives all points on the given set. Indicate on your sketch the three tangent vectors $f'(0)$, $f'(1)$, and $f'(-1)$.

We close this section with the following useful observation. There is a useful theorem in single-variable calculus which asserts that differentiability $\Longrightarrow$ continuity. The same result is valid in the present context, and the same proof applies:

THEOREM. *If* $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$ *is differentiable at* t_0*, then* f *is continuous at* t_0*.*

PROOF. We use the fact that the limit of a product equals the product of the limits. Thus

$$\begin{aligned}\lim_{t\to t_0}(f(t)-f(t_0)) &= \lim_{t\to t_0}(t-t_0)\frac{f(t)-f(t_0)}{t-t_0}\\ &= \lim_{t\to t_0}(t-t_0)\lim_{t\to t_0}\frac{f(t)-f(t_0)}{t-t_0}\\ &= 0f'(t_0)\\ &= 0.\end{aligned}$$

QED

B. **Lengths of curves**

We continue with our discussion of the special case $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$. We assume that f is differentiable.

KINEMATIC TERMINOLOGY. We often think of the independent variable t as *time*. Then the derivative gives the important quantities

$$\begin{aligned}f'(t) &= \textit{velocity (vector)} \text{ of the curve at time } t,\\ \|f'(t)\| &= \textit{speed} \text{ of the curve at time } t.\end{aligned}$$

In terms of the coordinate functions, the speed is

$$\sqrt{(f_1')^2+(f_2')^2+\cdots+(f_m')^2}.$$

Taking our cue from the basic fact that distance = speed × time, we next define the *length* of the curve $[a,b] \xrightarrow{f} \mathbb{R}^m$ to be the following:

DEFINITION. Assume that the curve $[a,b] \xrightarrow{f} \mathbb{R}^m$ is differentiable. Then its *length* is the definite integral

$$\int_a^b \|f'(t)\|dt,$$

provided that this integral exists.

We are not going to discuss in any detail the issue of the existence of this integral. The quantity $\|f'(t)\|$ might not be a continuous function of t, and in fact it might happen that the

integral assigns the value ∞ to the length. Examples of this behavior are easily found. Here is one:

PROBLEM 2–8. Let f be the curve in $\mathbb{R}^2$ defined by

$$f(t) = \begin{cases} (t^2 \cos t^{-2}, t^2 \sin t^{-2}) & \text{for } 0 < t \leq 1, \\ (0,0) & \text{for } t = 0. \end{cases}$$

a. Prove that f is differentiable (even at $t = 0$).

b. Prove that

$$\|f'(t)\| = 2\sqrt{t^2 + t^{-2}}.$$

c. Prove that

$$\int_0^1 \|f'(t)\| dt = \infty.$$

(HINT: use a lower bound for the integrand.)

(Incidentally, the existence of the integral has nothing to do with whether we can evaluate it in closed form. In fact, lengths of curves are usually very difficult to calculate, because of the square root involved in computing $\|f'(t)\|$.)

REMARKS. Most curves that actually arise in calculus are *piecewise continuously differentiable*. This means that there is a partition of the parameter interval $a \leq t_0 < t_1 < \cdots < t_k = b$ such that f is differentiable on each closed interval $[t_{i-1}, t_i]$ and f' is a continuous function there. Strictly speaking, f itself is not necessarily differentiable at the points t_i, but this causes no problem with computing the above integral. Thus we think of curves which may have "corners," but for which the tangent vectors have limits as we approach the corners (but the limits may be different from one another as $t \to t_i$ from $t < t_i$ or from $t > t_i$). The formula for length in this case is then given as the sum of the lengths of the pieces,

$$\sum_{i=1}^{k} \int_{t_{i-1}}^{t_i} \|f'(t)\| dt.$$

This is really the same as the original definition in view of the fact that the integral from a to b is independent of the values the integrand takes (or does not have) at the finitely many points $t_1, \ldots, t_{k-1}$.

EXAMPLE. A circle of radius R is described parametrically by

$$f(t) = (R\cos t, R\sin t), \qquad 0 \le t \le 2\pi.$$

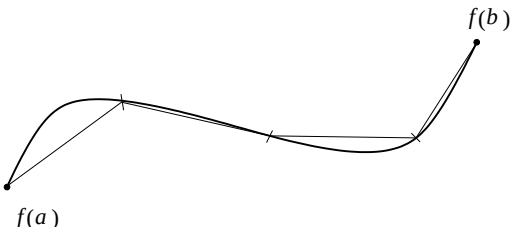
The velocity is $f'(t) = (-R\sin t, R\cos t)$ and the speed is therefore $\|f'(t)\| = R$. Thus the length of this circle is

$$\int_0^{2\pi} R dt = 2\pi R.$$

There is a related way to view the definition of length of a curve. Namely, think of a "polygon" "inscribed" in the given curve. Such a polygon may be defined by choosing a partition $a = t_0 < t_1 < \cdots < t_k = b$ of the parameter interval and using the line segments $[f(t_{i-1}), f(t_i)]$, $1 \le i \le k$, to approximate the arc. Then the length of this polygon is

$$\sum_{i=1}^{k} \|f(t_i) - f(t_{i-1})\|.$$

By the way, the polygon is an example of the piecewise continuously differentiable curves we discussed above, and Problem 2–10 below shows that the sum given here is indeed its length.



Since f is differentiable, we know that the norm

$$\frac{\|f(t_i) - f(t_{i-1})\|}{t_i - t_{i-1}}$$

is as close as we please to $\|f'(t_{i-1})\|$, provided $t_i - t_{i-1}$ is sufficiently small. Thus the length $\|f(t_i) - f(t_{i-1})\|$ is very well approximated by $\|f'(t_{i-1})\|(t_i - t_{i-1})$. Thus we expect the *Riemann sum*

$$\sum_{i=1}^{k} \|f'(t_{i-1})\|(t_i - t_{i-1})$$

to be a good approximation to the length of the polygon. On the other hand, the length of the polygon should be a good approximation of the length of the curve, so that the Riemann sum should be a good approximation of the length.

All of the above can be made rigorous, if needed, with some moderate hypothesis on f. However, as we are giving a *definition* of length, we choose not to pause to give the proof.

In fact, the idea of looking at inscribed polygons can be made into a definition of length which doesn't even mention the derivative at all. Suppose $f : [a, b] \to \mathbb{R}^m$ is any curve (still required to be continuous). Then we can *define*

$$\text{length of } \ f \ = \ \sup\{L\},$$

where $\{L\}$ stands for the set of numbers formed by all possible lengths L of polygons inscribed in f. It could happen that the length of f is ∞; in case it is finite we say that f is *rectifiable*. It is then a theorem that if f is piecewise continuously differentiable, then f is rectifiable and the two definitions of length produce the same number.

PROBLEM 2–9. Find the length of the curve $f(t) = (t^2, t^3)$ for $-1 \le t \le 0$. [Answer: $\frac{13^{3/2}-8}{27}$]

PROBLEM 2–10. The curve $f(t) = x + t(y - x)$, $0 \le t \le 1$, represents the line segment from x to y. Check that its length is $\|y - x\|$.

PROBLEM 2–11. Find the length of the *helix* in $\mathbb{R}^3$ given by

$$f(t) = (R\cos t, R\sin t, at), \qquad 0 \le t \le 2\pi.$$

PROBLEM 2–12. Find the length of the parabolic arch in $\mathbb{R}^2$ described by $f(t) = (t, a^2 - t^2)$, $-a \le t \le a$.

PROBLEM 2–13. Find the length of the exponential curve in $\mathbb{R}^2$ given as $f(t) = (t, e^t)$, $0 \le t \le 1$.

PROBLEM 2–14. Find the length of the logarithmic curve in $\mathbb{R}^2$ given as $f(t) = (\log t, t)$, $1 \leq t \leq e$.

PROBLEM 2–15. A curve in $\mathbb{R}^2$ called a *cycloid* is described by $f(t) = (t - \sin t, 1 - \cos t)$, $0 \leq t \leq 2\pi$. Find its length.

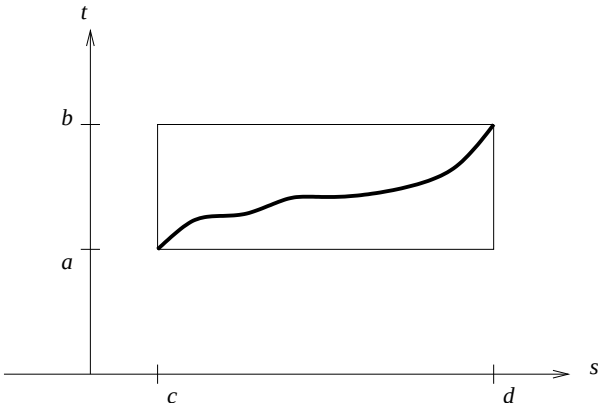
PROBLEM 2–16. A *hypocycloid* in $\mathbb{R}^2$ is described as the set of all points (x, y) satisfying $x^{2/3} + y^{2/3} = 1$. Draw a sketch of this set. Define the associated curve by $f(t) = (\cos^3 t, \sin^3 t)$ for $0 \leq t \leq 2\pi$, and compute its length.

PROBLEM 2–17. A curve in $\mathbb{R}^3$ is described by $f(t) = (t - \sin t, 1 - \cos t, 4 \sin \frac{1}{2}t)$. Show that its speed is 2.

PROBLEM 2–18. The preceding problem is an example of a curve invented just so its length can be calculated easily. Choose the constant a just right to render the length of the curve $f(t) = (t, t^2, at^3)$ easily computable.

PROBLEM 2–19. Find the length of the *catenary* described as $f(t) = (t, \cosh t)$, $-a \leq t \leq a$.

It is quite important to realize and exploit the fact that the length of a curve is unchanged if a reasonable change of the independent variable is made. We now explain and prove this feature. We suppose that $[a, b] \xrightarrow{f} \mathbb{R}^m$ is the curve, so $a \leq t \leq b$. We also suppose that another "parameter" s is to be used, $c \leq s \leq d$. And we suppose these are related by a differentiable function φ, so that $t = \varphi(s)$. We further suppose that φ is **increasing**:



Then we have, strictly speaking, a different curve g given by the composition

$$g = f \circ \varphi; \quad \text{that is }, \ g(s) = f(\varphi(s)).$$

Now we begin to compute the length of g. We first notice the consequence of the *chain rule*,

$$\begin{array}{ccccc} g'(s) & = & f'(\varphi(s)) & & \varphi'(s). \\ \uparrow & & \uparrow & & \uparrow \\ \text{vector} & & \text{vector} & & \text{scalar} \end{array}$$

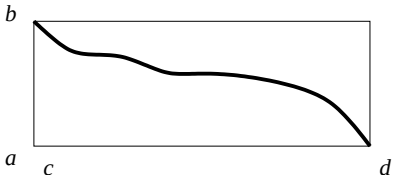
This is a consequence of the chain rule of single-variable calculus, and is proved by simply writing down the corresponding equation for each coordinate function of g. (Notice that we have written vector times scalar on the right side — it's the same product as the usual scalar times vector.) Since $\varphi' \geq 0$, the norms are related by

$$\|g'(s)\| = \|f'(\varphi(s))\|\varphi'(s).$$

Thus we have

$$\begin{aligned} \text{length of } g &= \int_c^d \|g'(s)\| ds \\ &= \int_c^d \|f'(\varphi(s))\|\varphi'(s) ds \\ &\overset{t=\varphi(s)}{=} \int_a^b \|f'(t)\| dt \\ &= \text{length of } f. \end{aligned}$$

In case the orientation is reversed, so that $\varphi' \leq 0$, we get the same result:

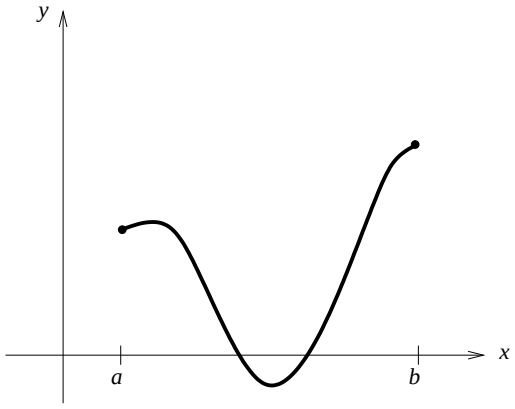


$$\begin{aligned}\int_c^d \|g'(s)\|ds &= \int_c^d \|f'(\varphi(s))\|(-\varphi'(s))ds \\ &= \int_b^a \|f'(t)\|(-dt) = \int_a^b \|f'(t)\|dt.\end{aligned}$$

Because of this invariance, if we deal with a set of points in $\mathbb{R}^m$ that is clearly equal to the image of some curve in a one-to-one fashion, we say that its length is the length of the corresponding curve. Thus, we have no qualms about saying the length of the circle $S(a,r) \subset \mathbb{R}^2$ is $2\pi r$, even though we haven't displayed a parametrization of $S(a,r)$.

For instance, the length of a **graph** $y = F(x)$, $a \leq x \leq b$, in $\mathbb{R}^2$ is given by the usual formula

$$\int_a^b \sqrt{1 + \left(\frac{dF}{dx}\right)^2}\, dx.$$



This is seen by using the parametrization

$$x \longrightarrow (x, F(x))\,.$$

There are various calculations we need to perform with this derivative of functions of one real variable. The chain rule, which we have already seen, is quite important. Other important

ones are versions of the *product rule*:

$$\begin{aligned} \text{Scalar Times Vector}: &\quad (hf)' = hf' + h'f, \\ \text{Vector Dot Vector}: &\quad (f \bullet g)' = f \bullet g' + f' \bullet g. \end{aligned}$$

PROBLEM 2–20. Prove the two versions of the product rule.

PROBLEM 2–21. Suppose a curve in $\mathbb{R}^m$ lies on a sphere. That is, $f(t) \in S(a,r)$ for all t. Prove that the velocity $f'(t)$ is tangent to the sphere and the acceleration vector satisfies $(f(t)-a) \bullet f''(t) \leq 0$. What is the kinematic interpretation of that inequality?

A special case of the latter version of the product rule is frequently of great use: for a curve in $\mathbb{R}^m$

$$\frac{d}{dt}\|f\|^2 = 2f \bullet f'.$$

A nice kinematic fact follows from this. If a curve $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$ has an *acceleration* f'' which exists, and if it has *constant speed*, then its acceleration is orthogonal to the curve. The proof is easy: we apply the above formula to f' rather than f and use the fact that $\|f'\|^2$ is constant. Thus

$$\begin{aligned} 0 &= \frac{d}{dt}\|f'\|^2 \\ &= 2f' \bullet f''. \end{aligned}$$

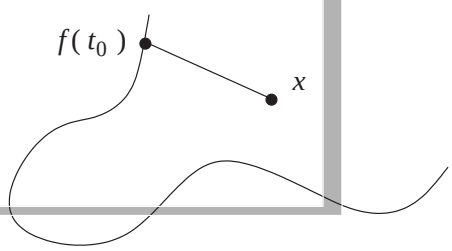
That is, f'' is orthogonal to the tangent vector f'.

PROBLEM 2–22. Let $\mathbb{R} \xrightarrow{f} \mathbb{R}^m$ be a differentiable curve and assume $x \in \mathbb{R}^m$ is a point which is not "on" the curve. Suppose $f(t_0)$ is a point on the curve which is closest to x: that is,

$$\|f(t_0) - x\| \leq \|f(t) - x\| \qquad \text{for all } \ t.$$

Prove that

$$f(t_0) - x \quad \text{is orthogonal to} \quad f'(t_0).$$



PROBLEM 2–23. Consider the parabola $y = x^2$ in $\mathbb{R}^2$. Let $-\infty < a < \infty$ and find the point(s) on the parabola which is (are) closest to $(0, a)$.
[Careful: you should discover two cases.]

PROBLEM 2–24. For any number $0 \leq a < 2\pi$ define the curve f_a in $\mathbb{R}^4$ by

$$f_a(t) = (\cos t, \sin t, \cos(t + a), \sin(t + a)), \quad 0 \leq t \leq 2\pi.$$

a. Show that for each fixed a the image $\{f_a(t) | 0 \leq t < 2\pi\}$ is a circle C_a which lies in a certain two-dimensional plane in $\mathbb{R}^4$.

b. Show that each C_a has center 0 and radius $\sqrt{2}$.

c. Show that if $a \neq b$, then $C_a \cap C_b = \emptyset$.

C. **Directional derivatives**

As we have just observed, it is very easy to develop calculus for vector-valued functions of one real variable. We now turn to the much more intriguing situation of functions of several variables, $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ with $n > 1$. For our first look at this situation we shall set things up to use the $n = 1$ case in a significant way.

Though we are facing a situation here that we may never have seen, something significant comes to mind. Namely, we could view the function values $f(x_1, x_2, \ldots, x_n)$ as depending on the single real variable x_i if we just regard all the other independent variables $x_1, \ldots, x_{i-1}$, $x_{i+1}, \ldots, x_n$ as fixed. Then we can perform "ordinary" differentiation with respect to x_i. The result of this differentiation could be denoted in the usual way as

$$\frac{df}{dx_i},$$

but the universally accepted and time-honored notation is instead

$$\frac{\partial f}{\partial x_i}.$$

Here then is the actual

DEFINITION. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$, and let $1 \leq i \leq n$. Then the *partial derivative of f with*

respect to x_i is

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

(provided the limit exists).

Notice that if $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$, then $\partial f / \partial x_i$ is a vector in $\mathbb{R}^m$.

EXAMPLES.

$$\begin{aligned}
\frac{\partial}{\partial x_1}(e^{x_1 x_2}) &= x_2 e^{x_1 x_2}; \\
\frac{\partial}{\partial x}(y \sin x) &= y \cos x; \\
\frac{\partial}{\partial y}(y \sin x) &= \sin x; \\
\frac{\partial}{\partial x}(x^y) &= y x^{y-1}; \\
\frac{\partial}{\partial y}(x^y) &= x^y \log x.
\end{aligned}$$

So we now have the concept of "partial" differentiation as being "ordinary" differentiation in coordinate directions. So far, so good, but we can do much better. After all, why be restricted to coordinate directions only? Why not investigate all directions in $\mathbb{R}^n$? We now explore this vast generalization, which will indeed free us from a coordinate system entirely.

Assume $x \in \mathbb{R}^n$ is fixed, and assume f is defined at least in a neighborhood of x, say a small closed ball $B(x, r)$.

We then consider a vector $h \in \mathbb{R}^n$ which will serve as a "direction." This means we look at the line through x in that direction, parametrized as the set of points $x + th$, $-\infty < t < \infty$. We then restrict attention to the behavior of f on this line. That is, we consider the function of t given as $f(x + th)$. This function is defined at least for all sufficiently small $|t|$. We then compute the t-derivative of this function at $t = 0$, if it exists.

DEFINITION. The *directional derivative* of f at x in the *direction* h is

$$\frac{d}{dt} f(x + th) \Big|_{t=0} = \lim_{t \to 0} \frac{f(x + th) - f(x)}{t}.$$

We shall use the notation

$$Df(x; h)$$

for this limit. Notice that if $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$, then the directional derivative $Df(x;h) \in \mathbb{R}^m$.

We stress that we have restricted our attention to an arbitrary straight line through x and have thus been enabled to use differentiation for a function of a single real variable. Notice that if $h = 0$ then we are not dealing with a straight line at all, but instead $f(x + t0) = f(x)$ is constant with respect to t and our definition yields

$$Df(x;0) = 0.$$

PROBLEM 2–25. Prove that for any scalar a,

$$Df(x;ah) = aDf(x;h).$$

(Notice that this result implies $Df(x;0) = 0$.)

Directional derivatives are not very interesting for functions of a single real variable, as all the "directions" just lie on $\mathbb{R}$. The directional derivative is just a scalar multiple of the ordinary derivative f':

PROBLEM 2–26. In the special case of $\mathbb{R} \xrightarrow{f} \mathbb{R}^n$, show that for any $h \in \mathbb{R}$

$$Df(x;h) = hf'(x).$$

PROBLEM 2–27. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ be given by

$$f(x_1, x_2) = (x_1 + x_2)e^{x_1 - x_2}.$$

Calculate the directional derivatives

$$Df\left((1,1);h\right) = 3h_1 - h_2,$$
$$Df\left((1,0);h\right) = 2eh_1.$$

PROBLEM 2–28. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^2$ be given by

$$f(r, \theta) = (r \cos \theta, r \sin \theta).$$

Calculate

$$Df\left((1, \theta); h\right) = (h_1 \cos \theta - h_2 \sin \theta, h_1 \sin \theta + h_2 \cos \theta).$$

PROBLEM 2–29. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^2$ be given by

$$f(x, y) = \left(\sqrt{x^2 + y^2}, \arctan \frac{y}{x}\right).$$

Calculate

$$Df\left((1, 0); h\right) = h.$$

PROBLEM 2–30. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be given by

$$f(x) = u \bullet x + a,$$

where $u \in \mathbb{R}^n$ and $a \in \mathbb{R}$ are constants. Calculate

$$Df(x; h) = u \bullet h.$$

The next two problems give directional derivative versions of the product rule.

PROBLEM 2–31. Suppose $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ and $\mathbb{R}^n \xrightarrow{g} \mathbb{R}$. Prove that

$$D(gf)(x; h) = g(x)Df(x; h) + Dg(x; h)f(x).$$

PROBLEM 2–32. Suppose $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ and $\mathbb{R}^n \xrightarrow{g} \mathbb{R}^m$. Prove that

$$D(g \bullet f)(x;h) = g(x) \bullet Df(x;h) + Dg(x;h) \bullet f(x).$$

There is also a version of the *chain rule*:

PROBLEM 2–33. Suppose that $\mathbb{R}^n \xrightarrow{f} \mathbb{R} \xrightarrow{g} \mathbb{R}$ and that g is a differentiable function. The composite function $g \circ f$ is defined by the equation $g \circ f(x) = g(f(x))$. Prove that

$$D(g \circ f)(x;h) = g'(f(x))Df(x;h).$$

PROBLEM 2–34. Let $f(x) = \|x\|^2$. Calculate

$$Df(x;h) = 2x \bullet h.$$

PROBLEM 2–35. Combine the two preceding problems to show that for any real number α and any $x \in \mathbb{R}^n$ with $x \neq 0$,

$$D(\|x\|^\alpha)(x;h) = \alpha\|x\|^{\alpha-2}x \bullet h.$$

(In particular,

$$D(\|x\|)(x;h) = \frac{x \bullet h}{\|x\|}.)$$

PROBLEM 2–36. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^n$ be given by

$$f(x) = \frac{x}{\|x\|^2}.$$

Calculate

$$Df(x;h) = \frac{h}{\|x\|^2} - \frac{2x \bullet h}{\|x\|^4}x.$$

PARTIAL DERIVATIVES. Still working in the context of directional derivatives, we frequently pay special attention to the *unit coordinate* directions e_1, $e_2, \ldots, e_n$. Here each vector $e_i \in \mathbb{R}^n$ is given by

$$e_i = (0, \ldots, 0, 1, 0, \ldots, 0),$$

where the single "1" appears in the i^{th} position.

PROBLEM 2–37. Show that

$$Df(x;e_i) = \frac{\partial f}{\partial x_i}.$$

We also frequently use a special notation:

$$D_i f(x) = Df(x;e_i) = \frac{\partial f}{\partial x_i}.$$

The notation $D_i f(x)$ has the advantage over $\partial f/\partial x_i$ of not having to name the coordinates a special way. We just have to keep track of the order in which they are written. For instance, if $f(m,p,a) = am^2p^3$, then $D_2 f = 3am^2p^2$.

Still another useful special notation represents partial derivatives with *subscripts*, so that

$$f_{x_i} = \frac{\partial f}{\partial x_i}.$$

PROBLEM 2–38. For the function of Problem 2–29 show that

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

Incidentally, in physics and engineering a special notation for the unit coordinate vectors in $\mathbb{R}^3$ is in vogue:

$$\hat{i} = (1, 0, 0),$$
$$\hat{j} = (0, 1, 0),$$
$$\hat{k} = (0, 0, 1).$$

In fact, in physics special attention is paid to *unit* vectors (vectors with norm 1), in that each such vector is dubbed with a circumflex ($\hat{}$). Thus it is correct in $\mathbb{R}^4$ to write $h = (\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2})$ as $\hat{h}$, whereas it is incorrect in $\mathbb{R}^2$ to write $h = (1, 1)$ as $\hat{h}$. We shall often employ this notation. Thus if you see a symbol $Df(x; \hat{h})$, you are assured that the norm $\|\hat{h}\| = 1$.

D. **Pathology**

"It is good for me that I have been afflicted, that I might learn thy statutes"

Psalm 119[71]

The purpose of this entire section is to present some examples of functions which have directional derivatives with certain strange properties. (Such examples are often called "counterexamples.") The reason for doing this is not my own love for the perverse, but rather to make sure we fully appreciate the tremendous usefulness of the concept of *differentiability* which will be discussed in the next section. These examples also serve to illustrate the inadequacy of the concept of directional differentiation, however appealing and useful it is.

All our examples are going to require that $n > 1$, and it so happens that $n = 2$ gives us enough room for the strange behavior we want to illustrate. Therefore, in this entire section we deal with

$$\mathbb{R}^2 \xrightarrow{f} \mathbb{R},$$

and we denote points in $\mathbb{R}^2$ with the notation (x, y). We still denote the directions as $h = (h_1, h_2)$. We shall also arrange things so that the pathology occurs at the origin in all cases.

Before going on, notice that all the examples and problems in Section C had the feature that the directional derivatives were *linear* functions of h. We'll have much more to say about this in the next section, but for now we just note that a linear function of h is a function of the form $c_1h_1 + c_2h_2$, where c_1 and c_2 are constants.

QUESTION 1. **Is it possible that a continuous function have a directional derivative which is a *nonlinear* function of the direction**?

ANSWER. **Yes**!

EXAMPLE.

$$f(x,y) = \left(x^{1/3} + y^{1/3}\right)^3.$$

Continuity is clear. And it is easily seen that

$$Df(0;h) = \left(h_1^{1/3} + h_2^{1/3}\right)^3 \ldots \text{ a nonlinear function of } \ h.$$

Incidentally, notice for example that

$$Df\left((1,8);h\right) = 9h_1 + \frac{9}{4}h_2 \text{ is a linear function of } \ h.$$

Here's another:

$$f(x,y) = \begin{cases} \frac{x^2y}{x^2+y^2} & \text{if } \ (x,y) \neq (0,0), \\ 0 & \text{if } \ (x,y) = (0,0). \end{cases}$$

Continuity is clear except at the origin. But notice that

$$|f(x,y)| = \frac{x^2|y|}{x^2+y^2} \leq |y|,$$

so f is continuous at 0. (Also, $|f(x,y)| \leq \frac{1}{2}|x|$.)

Now we compute

$$\begin{aligned} \frac{f(th) - f(0)}{t} &= \frac{\frac{t^3h_1^2h_2}{t^2h_1^2+t^2h_2^2} - 0}{t} \\ &= \frac{h_1^2h_2}{h_1^2+h_2^2}. \end{aligned}$$

This doesn't even depend on t, so certainly for $h \neq 0$

$$Df(0;h) = \frac{h_1^2h_2}{h_1^2+h_2^2} \ldots \text{ a nonlinear function of } \ h.$$

QUESTION 2. **Is it possible that a function have directional derivatives in every direction and *not* be continous**?

ANSWER. **Yes**!

EXAMPLE.

$$f(x,y) = \begin{cases} \frac{x^2y}{x^4+y^2} & \text{for } \ (x,y) \neq (0,0), \\ 0 & \text{for } \ (x,y) = (0,0). \end{cases}$$

PROBLEM 2–39. Show that f is discontinuous at the origin. Show that $Df(0;h)$ exists for all h and is given by

$$Df(0;h) = \begin{cases} h_1^2/h_2 & \text{if } \ h_2 \neq 0, \\ 0 & \text{if } \ h_2 = 0. \end{cases}$$

One might think that perhaps the trouble with the preceding example is that the directional derivative is nonlinear. Here's a somewhat more sophisticated counterexample.

QUESTION 3. **Is it possible that a function have directional derivative equal to zero in every direction and *not* be continous**?

ANSWER. **Yes**!

EXAMPLE.

Let

$$f(x,y) = \begin{cases} \frac{x^5 y}{x^8+y^4} & \text{for } \ (x,y) \neq 0, \\ 0 & \text{for } \ (x,y) = 0. \end{cases}$$

PROBLEM 2–40. Verify that f satisfies the conditions we have asserted for it.

MORAL. Unlike single-variable calculus, existence of directional derivatives does not imply continuity of the function. More subtly, the directional derivative is not necessarily a linear function of the direction. We shall soon discover how wonderful it is for the directional derivative to depend linearly on the direction, so we shall incorporate this property into the definition in the following section.

E. **Differentiability of real-valued functions**

At last we turn to the actual definition we shall employ. First, we need the important definition of linearity. We shall discuss this thoroughly in Section I, but for now we give a

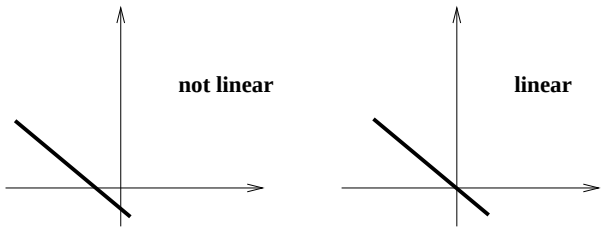
PROVISIONAL DEFINITION. A *linear* function from $\mathbb{R}^n$ to $\mathbb{R}$ is a function L of the form

$$L(h) = c_1 h_1 + \cdots + c_n h_n,$$

where the numbers $c_1, \ldots, c_n$ are constants. Assembling the coefficients $c_1, \ldots, c_n$ as the coordinates of a vector $c \in \mathbb{R}^n$, we can use our scalar product notation to write L in the form

$$L(h) = c \bullet h.$$

This probably doesn't quite agree with your usual terminology for linear functions. Here are two graphs of functions from $\mathbb{R}$ to $\mathbb{R}$:



A function from $\mathbb{R}$ to $\mathbb{R}$ of the form $f(x) = ax + b$ is said to be **affine**, and is therefore linear $\iff b = 0$. More generally, a function from $\mathbb{R}^n$ to $\mathbb{R}$ of the form $f(h) = c \bullet h + d$ is said to be **affine**. Thus every linear function is affine, and an affine function f is linear $\iff f(0) = 0$.

The next thing is to remind ourselves of the definition of differentiability of a function from $\mathbb{R}$ to $\mathbb{R}$. It is that the limit

$$c = \lim_{y \to 0} \frac{f(x+y) - f(x)}{y}$$

exists. Rewrite: there exists a number c such that

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

Here's a slight modification. Replace the denominator by $|y|$ (since the limit is 0, this doesn't change the definition):

linear function of y

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - cy}{|y|} = 0.$$

(To repeat: replacing the denominator y by its absolute value does not change the value 0 of the limit.)

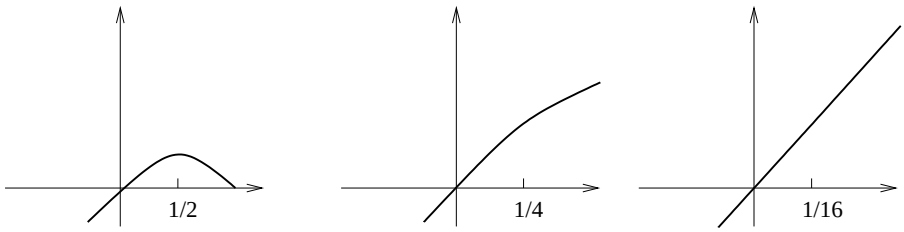
This is perfect! Notice the linear function in the numerator! Now that we have transformed the definition cleverly, we can immediately generalize to functions on $\mathbb{R}^n$, as follows.

★ **DEFINITION.** ★ **Let $x \in \mathbb{R}^n$ be a fixed point. Assume $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ is defined at least in a neighborhood of x (a ball $B(x,r)$). Then f is *differentiable* at x if there exists a vector $c \in \mathbb{R}^n$ such that**

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - c \bullet y}{\|y\|} = 0.$$

REMARK. This definition is absolutely crucial. Notice the vast difference between it and the concept of "directional" derivative. There is nothing "directional" in this definition — the variable point y tends to 0 in norm (which is true $\Longleftrightarrow$ all coordinates of y tend to 0) *with no restriction on its direction*.

REMARK. There is a nice geometric interpretation of this definition. In the case of a function from $\mathbb{R}$ to $\mathbb{R}$, the existence of $f'(x)$ has the familiar "tangent line" interpretation, namely that the graph of f near the point x looks affine on the microscopic scale. That is, $f(x) + f'(x)y$ is a very good approximation to $f(x+y)$ for small y. For instance, here are three sketches of the graph of $x - x^2$ near $x = 0$:



The same sort of thing is true in our case: the affine function of y given by $f(x) + c \bullet y$ is a very good approximation to the function $f(x+y)$ for small $\|y\|$. The difference between the given function and the affine function tends to zero as $\|y\| \to 0$, and it does so at a faster rate than $\|y\|$ itself: the *quotient* of the two even tends to zero.

Here's an easy fact:

if f is differentiable at x, then the directional derivatives exist at x.

Let y be restricted to have the form th in order to prove this (assuming ($h \neq 0$). Then the differentiability of f at x implies that

$$0 = \lim_{t \to 0} \frac{f(x+th) - f(x) - c \bullet th}{|t| \ \|h\|}.$$

Multiply by the number $\|h\|$ and multiply by $|t|/t$:

$$\begin{aligned} 0 &= \lim_{t\to 0} \frac{f(x+th) - f(x) - tc \bullet h}{t} \\ &= \lim_{t\to 0} \frac{f(x+th) - f(x)}{t} - c \bullet h. \end{aligned}$$

Thus, $Df(x;h) = c \bullet h$. In particular, if $h = \hat{e}_i$, then the i^{th} component of c is

$$c_i = \frac{\partial f}{\partial x_i}(x).$$

(Thus c is uniquely determined by f.)

DEFINITION. The vector c is called the **gradient** of f at x and is written with two different notations:

$$(\text{grad} f)(x) = (\nabla f)(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right).$$

Thus we have shown that if f is differentiable at x, then

$$\boxed{Df(x;h) = \nabla f(x) \bullet h.}$$

In particular, notice the very pleasant situation that $Df(x;h)$ **is a linear function of** h.

The notations we have chosen for the gradient are quite standard. The symbol ∇, an upside down delta, is called *del*. It also has the rather obsolete name *nabla*, and its properties are still sometimes called *the nabla calculus*.

We can now easily state some expected calculus formulas for the gradient. We assume that f and g are real-valued functions defined on subsets of $\mathbb{R}^n$. Then we have

$$\begin{aligned} \nabla(f+g) &= \nabla f + \nabla g; \\ \nabla(af) &= a\nabla f \quad \text{if } a \in \mathbb{R} \text{ is constant;} \\ \nabla f &= 0 \quad \text{if } f \text{ is constant;} \\ \nabla(fg) &= f\nabla g + g\nabla f. \end{aligned}$$

We do not treat the chain rule at the present time, as we shall discuss it thoroughly in Section K.

PROBLEM 2–41. Prove the above formulas. In addition, state and prove the corresponding quotient rule.

PROBLEM 2–42.

a. Let f be an **affine** function: $f(x) = c \bullet x + d$. Show that f is differentiable at any x, and $(\nabla f)(x) = c$.

b. Let f be the **quadratic** function: $f(x) = \|x\|^2$. Show that f is differentiable at any x, and $(\nabla f)(x) = 2x$. Thus

$$\nabla(c \bullet x + d) = c,$$
$$\nabla(\|x\|^2) = 2x.$$

(Solution: (a) $f(x+y) - f(x) - c \bullet y = c \bullet (x+y) + d - c \bullet x - d - c \bullet y = 0$. Thus

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - c \bullet y}{\|y\|} = 0.$$

(b) $f(x+y) = \|x+y\|^2 = \|x\|^2 + 2x \bullet y + \|y\|^2 = f(x) + 2x \bullet y + \|y\|^2$. Thus

$$\frac{f(x+y) - f(x) - 2x \bullet y}{\|y\|} = \|y\| \longrightarrow 0 \quad \text{as} \quad y \to 0.$$

Thus $(\nabla f)(x) = 2x$.)

We now quickly show that just as in the special case of Section B ($\mathbb{R} \xrightarrow{f} \mathbb{R}^m$), differentiability $\Longrightarrow$ continuity:

THEOREM. *If f is differentiable at x, then f is continuous at x.*

PROOF. This is *tres* simple: consider the two limits,

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - c \bullet y}{\|y\|} = 0,$$
$$\lim_{y \to 0} \|y\| = 0.$$

Multiply them, using the fact that the limit of a product is the product of the limits:

$$\lim_{y \to 0} \Big(f(x+y) - f(x) - c \bullet y \Big) = 0.$$

But of course $c \bullet y$ has limit zero, so we conclude

$$\lim_{y \to 0} \big(f(x+y) - f(x) \big) = 0;$$

i.e.,

$$\lim_{y\to 0} f(x+y) = f(x).$$

QED

We are surely elated that in the case of differentiability, $Df(x;h)$ is indeed a linear function of h. However, the converse is not valid, as the counterexample for Question 3 had $Df(0;h) = 0$ for all h (thus linear) and yet f was not even continuous at 0; and we now know that f could certainly therefore not be differentiable at 0.

You may say, "Aha! Suppose we assume that $Df(x;h)$ exists and is a linear function of h *and* f is continuous at x. Then perhaps f is differentiable at x":

QUESTION 4. **Is it possible that a *continuous* function with $Df(0;h) = 0$ for all h not be differentiable at** 0?

ANSWER. Yes!

Here's an example:

PROBLEM 2–43. Let

$$f(x,y) = \begin{cases} \frac{x^3 y}{x^4+y^2} & \text{for } (x,y) \neq (0,0), \\ 0 & \text{for } (x,y) = (0,0). \end{cases}$$

a. Show that f is continuous at $(0,0)$.

b. Show that $Df(0;h) = 0$ for all h.

c. Show that f is not differentiable at 0.

(The hard part is c. Here's a proof by contradiction. If f were differentiable at 0, show that it would follow that $\nabla f(0) = 0$. Conclude that

$$\lim_{(x,y)\to 0} \frac{f(x,y)}{\sqrt{x^2+y^2}} = 0.$$

Now show that this is not true.)

Therefore our *necessary* conditions that f be differentiable at x: (1) f is continuous at x and (2) $Df(x;h)$ is a linear function of h, turn out not to be *sufficient* to ensure the differentiability.

PROBLEM 2–44. As might be expected, the fourth calculus formula given above, the product rule, is more complicated to prove than the other easy calculus rules. Here is a lemma which will be useful in giving a proof: show that if f is differentiable at x, then there exists a constant C such that

$$|f(x+y) - f(x)| \ \leq \ C\|y\|$$

for all sufficiently small $y \in \mathbb{R}^n$.

The above inequality is called a *Lipschitz condition* for f at the point x. It is named for the German mathematician Rudolf Otto Sigismund Lipschitz. Notice that the inequality gives another proof of the continuity of a differentiable function.

Now it is easy to prove the product rule:

PROBLEM 2–45. Prove that if f and g are differentiable at x, then the product fg is differentiable at x, and

$$\nabla(fg)(x) = f(x)(\nabla g)(x) + g(x)(\nabla f)(x).$$

(HINT: write $f = f_1 + f(x)$ and $g = g_1 + g(x)$ (x is a fixed point). Express fg as a sum of four products and use the preceding problem to show that $\nabla(f_1 g_1)(x) = 0$.)

PROBLEM 2–46. What is wrong with this "proof" for the preceding problem?
a. We know

$$\nabla f(x) = (\partial f/\partial x_1, \ldots, \partial f/\partial x_n),$$
$$\nabla g(x) = (\partial g/\partial x_1, \ldots, \partial g/\partial x_n).$$

b. We also know that

$$\frac{\partial}{\partial x_i}(fg) = f\frac{\partial g}{\partial x_i} + g\frac{\partial f}{\partial x_i}.$$

c. Therefore,

$$\begin{aligned}\nabla(fg) &= (\partial(fg)/\partial x_1, \ldots, \partial(fg)/\partial x_n)\\ &= (f\partial g/\partial x_1 + g\partial f/\partial x_1, \ldots)\\ &= f(\partial g/\partial x_1, \ldots) + g(\partial f/\partial x_1, \ldots)\\ &= f\nabla g + g\nabla f.\end{aligned}$$

QED?

We have now made significant progress in our understanding of multivariable calculus, thanks to the all-important definition of differentiability. In the next section we shall learn how it can be quite useful.

F. **Sufficient condition for differentiability**

The property of differentiability is absolutely crucial in multivariable calculus. The definition is so technically involved that it is hard to verify directly. Happily, there is a *sufficient* condition that is extremely useful, and handles almost all cases that we ever need. We now state and prove it.

★**THEOREM.**★ **Assume $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ is defined in a neighborhood of x, and assume the partial derivatives $\partial f/\partial x_1, \ldots, \partial f/\partial x_n$ all exist in this neighborhood. Furthermore, assume these partial derivatives are all continuous at x. Then f is differentiable at x.**

PROOF. The main tool we shall employ in this proof is the famous *mean value theorem* of single-variable calculus. This theorem asserts (under the proper hypothesis) that for $-\infty < a < b < \infty$

$$g(b) - g(a) = g'(c)(b - a),$$

where c is some point in the interval $a < c < b$.

To verify the differentiability of f at x, we need to compare $f(x+y)$ and $f(x)$. We do this by moving from x to $x+y$ along n "steps" taken one coordinate at a time (a "taxicab" trip). Since there is nothing essentially different between the cases $\mathbb{R}^n$ and $\mathbb{R}^2$, I am content for the sake of brevity to perform the proof for the case $n = 2$ only.

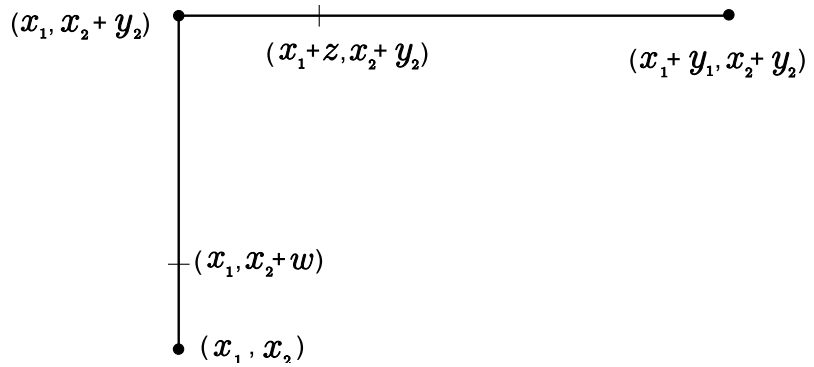
We use the notation $c = \nabla f(x) = (\partial f/\partial x_1, \partial f/\partial x_2)$. (Of course, the hypothesis guarantees the existence of these partial derivatives.) Then for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$\|y\| < \delta \Longrightarrow \left| \frac{\partial f}{\partial x_i}(x+y) - c_i \right| < \epsilon/2;$$

this is where we use the hypothesis of continuity of $\partial f/\partial x_i$ at x.

Now we write in two "steps"

$$f(x+y) - f(x) = [f(x_1+y_1, x_2+y_2) - f(x_1, x_2+y_2)] + [f(x_1, x_2+y_2) - f(x_1, x_2)]\,.$$



This is perfectly arranged to use the mean value theorem on each term to produce for any $\|y\| < \delta$,

$$f(x+y) - f(x) = \frac{\partial f}{\partial x_1}(x_1+z, x_2+y_2)\, y_1 + \frac{\partial f}{\partial x_2}(x_1, x_2+w)\, y_2,$$

where z is between 0 and y_1 and w is between 0 and y_2. Therefore,

$$\begin{aligned}|f(x+y)-f(x)-c\bullet y| &= \left|\left[\frac{\partial f}{\partial x_1}(x_1+z,x_2+y_2)-c_1\right]y_1+\left[\frac{\partial f}{\partial x_2}(x_1,x_2+w)-c_2\right]y_2\right| \\ &\le \frac{1}{2}\epsilon|y_1|+\frac{1}{2}\epsilon|y_2| \\ &\le \frac{1}{2}\epsilon\|y\|+\frac{1}{2}\epsilon\|y\| \\ &= \epsilon\|y\|.\end{aligned}$$

Therefore,

$$0<\|y\|<\delta \Longrightarrow \frac{|f(x+y)-f(x)-c\bullet y|}{\|y\|} \;\le\; \epsilon.$$

Thus we have proved that

$$\lim_{y\to 0}\frac{f(x+y)-f(x)-c\bullet y}{\|y\|}=0.$$

Therefore, f is differentiable at x (and $c=\nabla f(x)$).

QED

The theorem we have just proved is quite wonderful, as is its proof. This is just the sort of mathematical proof that essentially works itself. Once we decided to use the *partial derivatives*, then the diagram in the body of the proof suggests itself, as does the use of the mean value theorem.

Whereas the next problem holds no interest for calculus that I am aware of, working through it may enhance your understanding of the above important proof. It is an "improvement" of the theorem to be sure, as it provides the same conclusion with weaker hypotheses.

PROBLEM 2–47. Assume $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ is defined in a neighborhood of x, and assume the partial derivatives $\partial f/\partial x_1,\ldots,\partial f/\partial x_{n-1}$ all exist in this neighborhood and are continuous at x. Assume also that $\partial f/\partial x_n$ exists *at* x. Prove that f is differentiable at x.
(HINT: limit attention to $n=2$ and apply the mean value theorem only to the term $f(x_1+y_1,x_2+y_2)-f(x_1,x_2+y_2)$ in the above proof.)

It so happens that in practice an even *weaker* theorem than the one we have proved is definitive in an amazing variety of situations:

COROLLARY. **Assume $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ is defined in an open ball $B(x,r)$, and that the partial derivatives $\partial f/\partial x_1,\ldots,\partial f/\partial x_n$ all exist and are continuous functions in $B(x,r)$. Then f is differentiable at every point of $B(x,r)$.**

This result is an immediate corollary of the theorem. The reason it is so useful is that we can very often tell at a glance that a function has these continuous partial derivatives, and thus it must be differentiable. For instance,

$$f(x_1, x_2, x_3, x_4) = \log\left(x_1^2 + x_2^4 + x_3^6 + e^{x_1 - x_4}\right) + \sin x_3 \cos(\sin x_1)$$

gives a function f that clearly satisfies the conditions on all of $\mathbb{R}^4$; for all we need do is the mental exercise of thinking about how to compute the four partial derivatives $\partial f/\partial x_i$. For that we use the full power of one-variable calculus — the chain rule, the product rule etc. — and realize that the formulas we thereby obtain give continuous functions on $\mathbb{R}^4$. The only possible "trouble" that could arise in this particular example would occur because the argument of log might be 0. As this argument is clearly always positive, that cannot happen. We conclude that each $\partial f/\partial x_i$ is continuous and thus that f is differentiable at every point of $\mathbb{R}^4$.

EXAMPLE. Here's a reworking of Problem 2–27:

$$f(x_1, x_2) = (x_1 + x_2)e^{x_1 - x_2}$$

We see at a glance that f satisfies the conditions of the corollary. If we want the directional derivative at any $x \in \mathbb{R}^2$, we can use the formula on p. 2–26:

$$\frac{\partial f}{\partial x_1} = (1 + x_1 + x_2)e^{x_1 - x_2},$$
$$\frac{\partial f}{\partial x_2} = (1 - x_1 - x_2)e^{x_1 - x_2}.$$

Thus

$$Df(x; h) = \left[(1 + x_1 + x_2)\, h_1 + (1 - x_1 - x_2)\, h_2\right] e^{x_1 - x_2}.$$

No difference quotients needed! No need to substitute $x + th$ and compute the t-derivative and substitute $t = 0$. It's just algebra!

DEFINITION. A function which satisfies the hypothesis of the above corollary is said to be *continuously differentiable*. It is also said to be of *class* C^1. (A continuous function is of class C^0.)

Here we give an example which illustrates the power of this corollary. Consider the function given as the norm, $f(x) = \|x\|$ for $x \in \mathbb{R}^n$, $x \neq 0$. Since $f(x)$ is the square root of the positive quantity $x_1^2 + \cdots + x_n^2$, the chain rule of single-variable calculus makes it clear that all the

partial derivatives $\partial f/\partial x_i$ are continuous for $x \neq 0$; thus f is of class C^1 for $x \neq 0$. We already have from Problem 2–35 the directional derivative

$$Df(x;h) = \frac{x \bullet h}{\|x\|}.$$

Since f is differentiable (by the corollary),

$$\nabla f(x) \bullet h = Df(x;h).$$

Therefore,

$$\nabla f(x) \bullet h = \frac{x \bullet h}{\|x\|} \quad \text{for all} \quad h \in \mathbb{R}^n.$$

We can now conclude that $\nabla f(x) = x/\|x\|$. (See Problem 1-11.) We record this result in the form

$$\boxed{\nabla \|x\| = \frac{x}{\|x\|}.}$$

PROBLEM 2–48. Show that for any real number α

$$\nabla \|x\|^\alpha = \alpha \|x\|^{\alpha-2} x \qquad \text{for} \quad x \neq 0.$$

For which α is this equation also valid for $x = 0$?

There are two common misconceptions concerning differentiability. One is the idea that our sufficient condition is also necessary. The following example shows this not to be the case even for single-variable calculus:

PROBLEM 2–49. Define $\mathbb{R} \xrightarrow{f} \mathbb{R}$ by

$$f(x) = \begin{cases} x^2 & \text{if } x \text{ is irrational,} \\ 0 & \text{if } x \text{ is rational.} \end{cases}$$

Prove the following:

a. f is continuous *only* at 0.

b. f is differentiable at 0, and $f'(0) = 0$.

c. f is differentiable only at 0.

The second misconception is that the mere existence of the partial derivatives in a ball automatically implies their continuity. This again is not true even for single-variable calculus. Here is an example found in almost all calculus texts:

PROBLEM 2–50. Define $\mathbb{R} \xrightarrow{f} \mathbb{R}$ by

$$f(x) = \begin{cases} x^2 \sin \frac{1}{x} & \text{for } \ x \neq 0, \\ 0 & \text{for } \ x = 0. \end{cases}$$

Prove that

a. f is differentiable on all of $\mathbb{R}$;

b. $f'(0) = 0$;

c. $f'(x)$ is not a continuous function of x at $x = 0$.

G. **A first look at critical points**

We'll eventually present critical points in great depth, and in fact won't finish the discussion until Chapter 4, but already we are able to discuss the concept to some extent.

DEFINITION. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be differentiable at x. Then x is called a *critical point* of f if $\nabla f(x) = 0$.

This terminology should already be familiar to you from single-variable calculus. Notice that the condition for x to be a critical point can be expressed in terms of partial derivatives:

$$\frac{\partial f}{\partial x_i}(x) = 0 \quad \text{for } \ 1 \leq i \leq n.$$

Another important concept:

DEFINITION. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be defined on a set $A \subset \mathbb{R}^n$ and let $x_0 \in A$. We say that f has a *global maximum* at x_0 if

$$f(x) \leq f(x_0) \quad \text{for all } \ x \in A.$$

We say that f has a *local maximum* at x_0 if there exists $r > 0$ such that

$$f(x) \leq f(x_0) \quad \text{for all } \ x \in A \cap B(x_0, r).$$

(Notice that a global maximum is certainly a local maximum, but not conversely.) We define similarly global *minimum* and local *minimum*. We say that f has a local *extreme value* at x_0 if it has a local maximum or a local minimum at x_0.

As in single-variable calculus, we have the easy

THEOREM. *Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be differentiable at x_0 and also have a local extreme value at x_0. Then x_0 is a critical point of f. The converse statement does not necessarily hold, even if $n = 1$.*

PROOF. We are content to handle the case of a local minimum. Then for any $h \in \mathbb{R}^n$,

$$f(x_0 + th) \geq f(x_0) \quad \text{for all sufficiently small } |t|.$$

Therefore for some $\epsilon > 0$

$$\frac{f(x_0 + th) - f(x_0)}{t} \text{ is } \begin{cases} \geq 0 & \text{for } 0 < t < \epsilon, \\ \leq 0 & \text{for } -\epsilon < t < 0. \end{cases}$$

Now let $t \to 0$ to achieve both the inequalities

$$Df(x_0; h) \geq 0 \quad \text{and} \quad \leq 0.$$

Therefore, $Df(x_0; h) = 0$ for all directions $h \in \mathbb{R}^n$. In particular, $\partial f / \partial x_i = 0$ at x_0 for $1 \leq i \leq n$. Thus x_0 is a critical point. For the statement about a converse, most people's favorite example is $f(x) = x^3$ for $x \in \mathbb{R}$.

QED

EXAMPLE. If $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ is given by $f(x, y) = xy^2$, then $\nabla f = (y^2, 2xy)$. Thus *every* point of the form $(x, 0)$ is a critical point. Notice that critical points do not have to be isolated.

EXAMPLE. Suppose we want to find all the critical points of the function given as $f(x, y) = 5x^2y + xy^2 + 15xy$. Then we compute the two partial derivatives to get the two scalar equations

$$\begin{cases} 10xy + y^2 + 15y &= 0, \\ 5x^2 + 2xy + 15x &= 0. \end{cases}$$

There are two things to notice before proceeding. Namely, we have two equations for two unknowns (x and y) and they are *nonlinear*. This is the usual situation, and usually it will be

difficult or impossible to solve the equations explicitly. However, in this particular example it's pretty simple. The equations can be rewritten

$$\begin{cases} y(10x + y + 15) &= 0, \\ x(5x + 2y + 15) &= 0. \end{cases}$$

The first equation asserts that

$$y = 0 \quad \text{or} \quad 10x + y = -15,$$

and the second that

$$x = 0 \quad \text{or} \quad 5x + 2y = -15.$$

There are four possibilities. The least obvious one is the one in which the two affine equations are satisfied, and elimination gives the solution $x = -1$, $y = -5$. Thus the four critical points of f are

$$(0, 0),\ (0, -15),\ (-3, 0),\ (-1, -5).$$

The next eleven problems give functions defined on (a subset of) $\mathbb{R}^n$. All of them are of class C^1, and thus are differentiable. You are to find all the critical points of each.

PROBLEM 2–51.

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

PROBLEM 2–52.

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

Here f is defined only for $x \neq 0$, $y \neq 0$.

PROBLEM 2–53.

$$f(x, y) = x^2y^3(2x + 3y - 6).$$

PROBLEM 2–54.

$$f(x, y) = \frac{1}{xy} - \frac{a}{x^2y} - \frac{b}{xy^2} + 17.$$

Here a, b are nonzero constants.

PROBLEM 2–55.

$$f(x,y,z) = \frac{x^3+y^3+z^3}{xyz}.$$

PROBLEM 2–56.

$$f(x,y) = xy + x^{-1} + y^{-1}.$$

PROBLEM 2–57.

$$f(x,y,z) = xyz + ax^{-1} + by^{-1} + cz^{-1}.$$

Here a, b, c are nonzero constants. There are two qualitatively different cases, depending on the sign of abc.

PROBLEM 2–58.

$$f(x,y,z,w) = xyzw + ax^{-1} + by^{-1} + cz^{-1} + dw^{-1}.$$

Here a, b, c, d are nonzero constants.

PROBLEM 2–59.

$$f(x,y) = (y-x^2)(y-2x^2).$$

PROBLEM 2–60.

$$f(x) = (a_1x_1^2 + \cdots + a_nx_n^2)e^{-\|x\|^2}.$$

Here $\|x\|$ is the norm of x, and the constants satisfy $a_1 > a_2 > \cdots > a_n > 0$.
(There are 2n+1 critical points.)

PROBLEM 2–61.

$$f(x,y) = (x+y)e^{-\sqrt{x^2+y^2}}.$$

(There are two critical points.)

PROBLEM 2–62. The preceding problem brings up an issue.

a. Prove that the function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ given as the norm, $f(x) = \|x\|$, is differentiable at every $x \neq 0$, and is not differentiable at $x = 0$.

b. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be defined as

$$f(x) = x_1 e^{-\|x\|}.$$

Prove that f is differentiable even at $x = 0$, and calculate $(\nabla f)(0)$.

PROBLEM 2–63. Let a be a fixed real number and find the critical points of the function on $\mathbb{R}^n$ defined by $f(x) = \|x\|^2 + x_1 + a\|x\|$. (There will be three cases depending on what a is.)

EXAMPLE. The critical point structure of this function will prove to be very enlightening. Define $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ by

$$f(x, y) = (x^2 - 1)^2 + (x^2 - e^y)^2.$$

As f is given as a sum of two squares, it is clear that $f(x, y) \geq 0$. Moreover, $f(x, y) = 0 \iff x^2 = 1$ and $x^2 = e^y$. Thus f attains its global minimum value precisely at the two points

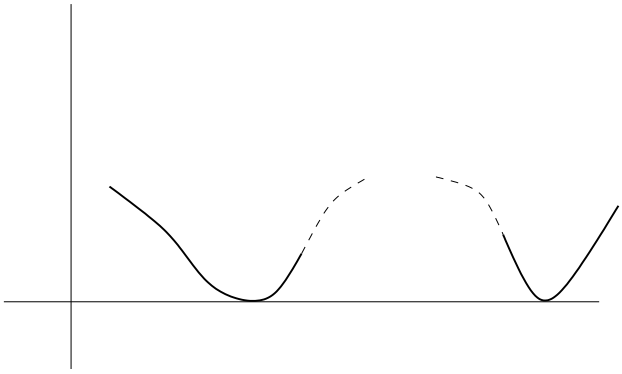
$$(1, 0) \quad \text{and} \quad (-1, 0).$$

We know these are critical points. Let us see if there are others:

$$\partial f/\partial x = 4x(x^2 - 1) + 4x(x^2 - e^y) = 0,$$
$$\partial f/\partial y = -2e^y(x^2 - e^y) = 0.$$

The second of these equations of course requires $x^2 = e^y$, and then the first requires $4x(x^2-1) = 0$. Thus $x = 0$, 1, or -1. The value $x = 0$ is excluded by $x^2 = e^y$. We then have $1 = e^y$, so $y = 0$. Thus the two points we found by inspection are the only critical points.

MORAL. Situations can be much more complicated in two variables than in one. A differentiable function defined on an interval in $\mathbb{R}$ cannot have two global minima unless it has at least one more critical point (a local maximum), as indicated in the following sketch. But with two independent variables there is more room to "maneuver." We might say that there doesn't need to be a mountain peak between two lakes.



PROBLEM 2–64. Define $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ by

$$f(x, y) = 3xe^y - x^3 - e^{3y}.$$

a. Show that $(1, 0)$ is the only critical point.

b. Show that $(1, 0)$ is a local maximum. (You *may* want to show first that $f(x, y) \leq 2x^{3/2} - x^3$ for all $x > 0$, $-\infty < y < \infty$.)

c. Show that $(1, 0)$ is not a global maximum.

d. Give the *moral* of this example.

H. **Geometric significance of the gradient**

In this section we return to the general situation

$$\mathbb{R}^n \xrightarrow{f} \mathbb{R}$$

in which f is differentiable at the point x. We have defined the gradient $\nabla f(x)$ and we know "algebraically" how to compute it in terms of partial derivatives,

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right),$$

but we now want to explore the *geometry* which is contained in this vector.

First, it could be that x is a critical point of f. By definition this means that $\nabla f(x) = 0$ and the geometrical situation is quite clear.

Thus we assume from now on that x is not a critical point; that is,

$$\nabla f(x) \neq 0.$$

We want to examine thoroughly the equation (p. 2–26) which relates the directional derivative and the gradient,

$$Df(x;h) = \nabla f(x) \bullet h.$$

To recapitulate, the definition of $Df(x;h)$ is very geometrical in nature, and so is that of $\nabla f(x)$. We also have a way of computing ∇f almost algebraically, using the coordinates of $\mathbb{R}^n$ and the partial derivatives of f at x. And then computing the directional derivative really does become a matter of linear algebra:

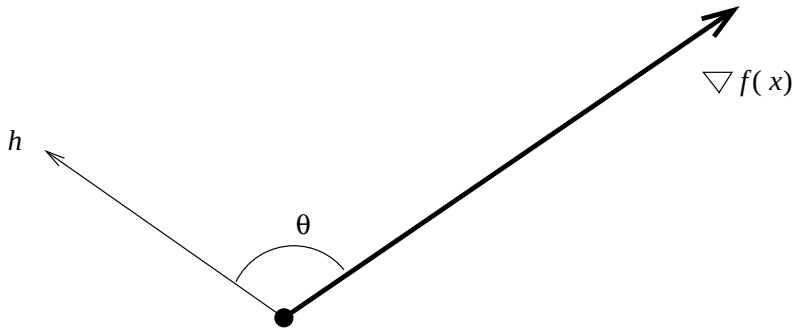
$$Df(x;h) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_j}(x) h_i.$$

What we are going to do now is add to this situation our geometrical understanding of the dot product so that we emerge with a geometrical understanding of the gradient.

Therefore we now consider the behavior of $\nabla f(x) \bullet h$ as the *direction* of h varies. In order to make this significant, we shall now work exclusively with *unit* vectors $\hat{h}$ (see p. 2–21 for this notation).

Geometrically, this means that we are actually looking at "directions" $\hat{h}$ in $\mathbb{R}^n$, emanating from x, and the directional derivative $Df(x;\hat{h})$ is a measure of the "rate of increase of f at x in the direction $\hat{h}$."

Let θ denote the angle between the gradient $\nabla f(x)$ and the direction $\hat{h}$. We then recall from p. 1–17 that



$Df(x;h) = \|\nabla f(x)\| \cos\theta.$

(Remember: $\|\hat{h}\| = 1$.) This equation is very revealing. It shows that the maximum value of $Df(x;\hat{h})$ is the norm of the gradient $\|\nabla f(x)\|$, and this maximum rate of increase is realized $\iff \hat{h}$ is the unit vector in the same direction as $\nabla f(x)$.

(Also, the minimum of $Df(x;\hat{h})$ is $-\|\nabla f(x)\|$, and this occurs $\iff \hat{h}$ is the unit vector in the same direction as $-\nabla f(x)$.)

Not only is this useful geometric information about the directional derivative, but also it gives us a way to describe $\nabla f(x)$ in purely geometric terms, with no reference whatsoever to a coordinate system! Namely, still assuming x is not a critical point,

> $\nabla f(x)$ is the unique vector at x in $\mathbb{R}^n$ determined as follows:
>
> - its *direction* is the direction of maximal increase of f at x,
> - its *magnitude* or norm is the rate of this maximal increase.

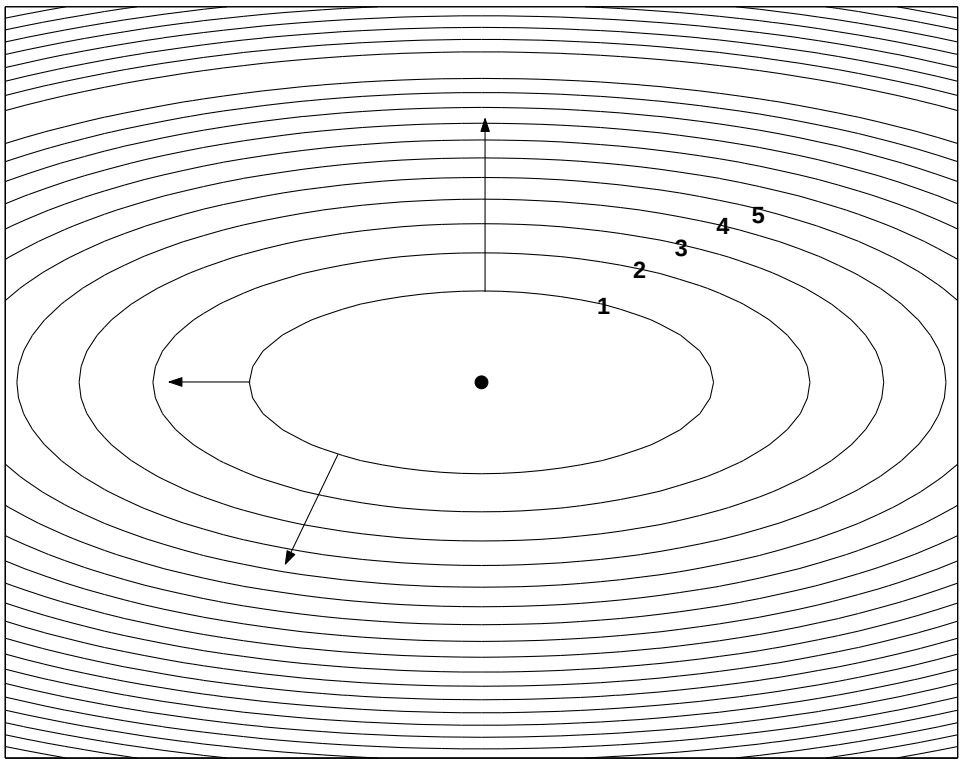
This is indeed a wonderful situation, one that frequently happens in mathematics. We have an important quantity (in this case, the gradient of a function) which, on the one hand, has a completely geometric description, and which, on the other hand, can be computed in a coordinate system in a very routine and useful manner (in this case, as $(\partial f/\partial x_1, \ldots, \partial f/\partial x_n)$). Truly a double-edged sword!

Level sets. As you may realize, the graph of a function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ seems to be less visually helpful for $n > 1$ than for $n = 1$. However, the idea of *level sets* of f seems quite useful. By definition, these are sets of the form

$$\{x \in \mathbb{R}^n \mid f(x) = a\},$$

where a is any fixed real number. These sets are obviously disjoint (as a varies) and fill up all of the domain of f. Sketching them is clearly a different matter from sketching the graph of f (a subset of $\mathbb{R}^{n+1}$). They seem especially convenient when $n = 2$.

EXAMPLE. Let $f(x, y) = \frac{x^2}{4} + y^2$. The level sets of f are of course ellipses with center $(0, 0)$ and with the same shape. (Unless $a < 0$, in which case the level set is empty; or $a = 0$, in which case the level set is just $(0, 0)$.) Here are rough sketches of a few level sets, where the numbers refer to the value f attains along that particular level set.

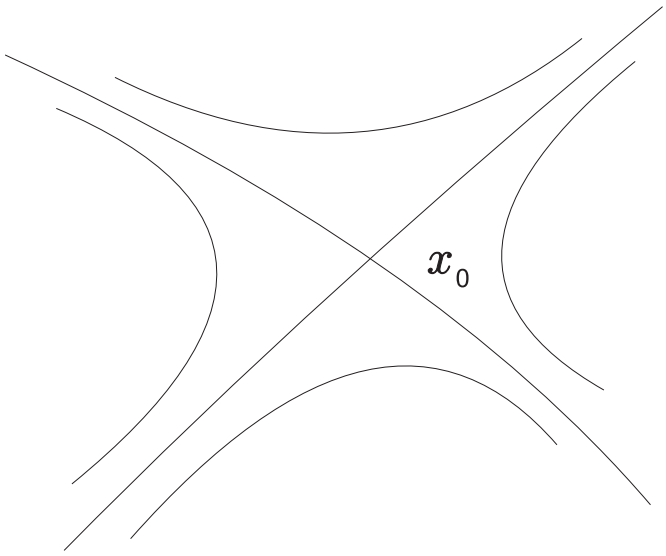


I have sketched $\nabla f(x, y)$ at three points of the level set $f(x, y) = 1$.

Notice that ∇f is always *orthogonal* to the level sets in the above sketch. This feature is always true, but we need to wait for the discussion of the *chain rule* in Section K to see this. Also notice that the more tightly spaced the level sets are, the larger the gradient must be. This is of course because the norm of the gradient measures the rate of increase of f in that

direction, and when the level sets are close together, f must be changing rapidly. When you hike in Colorado with a survey map that shows curves of constant altitude, it's when they are close together that your hike is strenuous.

PROBLEM 2–65. Here is a rough sketch of some level sets of a certain function $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$:



Suppose that f is differentiable at the indicated point x_0. Prove that x_0 is a critical point for f.

PROBLEM 2–66. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ be differentiable at x_0. Suppose that

$$Df\left(x_0; \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)\right) = 1$$
$$Df\left(x_0; \left(\frac{1}{2}, \frac{-\sqrt{3}}{2}\right)\right) = 1.$$

Calculate $(\nabla f)(x_0)$.

PROBLEM 2–67. Let $\mathbb{R}^4 \xrightarrow{f} \mathbb{R}$ be differentiable at x_0. Suppose that

$$Df\left(x_0; \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}\right)\right) = 0,$$
$$Df\left(x_0; \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}\right)\right) = 1,$$
$$Df\left(x_0; \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right)\right) = 2,$$
$$Df\left(x_0; \left(\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}\right)\right) = 3.$$

Calculate $(\nabla f)(x_0)$.

PROBLEM 2–68. Sketch a few level sets of the function defined on $\mathbb{R}^2$ by $f(x, y) = x^2 - y^2 + 3$. Also give accurate sketches of ∇f at several points in $\mathbb{R}^2$.

PROBLEM 2–69. Repeat Problem 2–68 but with $f(x, y) = y^2 - x^3 - x^2$.

I. **A little matrix algebra**

In the next section we are going to discuss the concept of differentiation for a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. The key concept we shall have to understand is the idea of a *linear* function from $\mathbb{R}^n$ to $\mathbb{R}^m$. The reason is that the derivative in this general case is intimately connected to a type of affine approximation to the given function. This is yet another instance of the intimate connection that calculus provides between algebraic and geometric concepts. In the present section we want to provide the necessary elementary algebraic structure. We first need to modify the provisional definition we gave in Section E:

DEFINITION. A *linear* function from $\mathbb{R}^n$ to $\mathbb{R}^m$ is a function $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$ which is compatible with vector addition and scalar multiplication, in the sense that

$$F(x + y) = F(x) + F(y),$$
$$F(ax) = aF(x),$$

for all x, $y \in \mathbb{R}^n$ and all $a \in \mathbb{R}$.

REMARK. Often linear functions are also called linear *transformations* or linear *operators*.

PROBLEM 2–70. Here are some immediate consequences of the definition, showing further how a linear function respects the algebraic properties of $\mathbb{R}^n$ and $\mathbb{R}^m$: prove that if F is linear, then

$$\begin{aligned} F(0) &= 0, \\ F(-x) &= -F(x), \\ F\left(a_1x^{(1)} + a_2x^{(2)} + \cdots + a_kx^{(k)}\right) &= a_1F\left(x^{(1)}\right) + a_2F\left(x^{(2)}\right) + \cdots + a_kF\left(x^{(k)}\right). \end{aligned}$$

PROBLEM 2–71. Prove that we could have defined linear function by requiring F to "preserve linear combinations," in the sense that

$$F(ax + by) = aF(x) + bF(y)$$

for all x, $y \in \mathbb{R}^n$ and all a, $b \in \mathbb{R}$.

PROBLEM 2–72. Show that our provisional definition in the case $\mathbb{R}^n \xrightarrow{F} \mathbb{R}$, namely that $F(x) = c \bullet x$, gives linear functions. Moreover, prove that if $\mathbb{R}^n \xrightarrow{F} \mathbb{R}$ is linear, then there exists a unique $c \in \mathbb{R}^n$ such that $F(x) = c \bullet x$.

PROBLEM 2–73. Just as on p. 2–20, we need to understand the difference between *affine* and *linear* functions. We define $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$ to be *affine* if

$$F\left(ax + (1-a)y\right) = aF(x) + (1-a)F(y)$$

for all vectors x and y and all scalars a. Prove that F is affine $\iff$ there exists a linear function F_0 and a fixed vector $w \in \mathbb{R}^m$ such that

$$F(x) = w + F_0(x) \quad \text{for all} \quad x.$$

Prove that F_0 and w are uniquely determined by F.

PROBLEM 2–74. Prove that if F is affine, then

$$F\left(a_1x^{(1)} + a_2x^{(2)} + \cdots + a_kx^{(k)}\right) = a_1F\left(x^{(1)}\right) + a_2F\left(x^{(2)}\right) + \cdots + a_kF\left(x^{(k)}\right)$$

whenever $a_1 + a_2 + \cdots + a_k = 1$.

It is virtually impossible to provide meaningful pictorial descriptions of linear functions. However, there is a classic bookkeeping device for handling them, namely, the algebra of *matrices*.

DEFINITION. An $m \times n$ *matrix* of real numbers is a rectangular array A having m *rows* and n *columns*. A standard notation for A is

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}.$$

We say that the real number a_{ij} is the *entry* of A in row i and column j. We also say that

the i^{th} *row* of A is $(a_{i1}\ a_{i2} \dots a_{in})$ (a $1 \times n$ matrix)

and the j^{th} *column* of A is $\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$ (an $m \times 1$ matrix).

We say that the matrix A has *shape* $m \times n$. If we are in a context in which the shape of A is known, we may abbreviate

$$A = (a_{ij}).$$

Here are some basic algebra definitions:

- two matrices A and B are equal $\Longleftrightarrow$ they have the same entries at each position (A and B must therefore have the same shape).
- $A + B$ is defined entry-wise (A and B must therefore have the same shape), so that

  $$(a_{ij}) + (b_{ij}) = (a_{ij} + b_{ij});$$

  notice that $A + B = B + A$.

- 0 is the $m \times n$ matrix whose entries are all 0; thus

$$A + 0 = 0 + A = A.$$

- If $c \in \mathbb{R}$, cA is defined entry-wise by

$$cA = (ca_{ij}).$$

- Notice that the matrix $-A$ defined as $(-1)A$ is the unique matrix which when added to A gives the matrix 0.

It is nice to observe that the set of all $m \times n$ matrices is now additively just like $\mathbb{R}^{mn}$. Namely, each $m \times n$ matrix is specified by its mn real entries, arranged in a certain pattern. The same is true of each vector in $\mathbb{R}^{mn}$. Not only that, but also addition of matrices and scalar multiplication are just like they are for $\mathbb{R}^{mn}$.

However, matrices enjoy another algebraic property that far outweighs the above in importance. That is, we can *multiply* them in certain situations. The precise definition is this:

- if $A = (a_{ij})$ is an $m \times p$ matrix and $B = (b_{ij})$ is a $p \times n$ matrix, then the *product* AB is the $m \times n$ matrix whose ij entry is

$$\sum_{k=1}^{p} a_{ik} b_{kj}.$$

  In other words, the entry of AB in row i and column j is obtained from the i^{th} row of A and the j^{th} column of B by a kind of *dot* product of vectors in $\mathbb{R}^p$:

$$(a_{i1}, a_{i2}, \ldots, a_{ip}) \bullet (b_{1j}, b_{2j}, \ldots, b_{pj}).$$

  (Notice the commas!).

Here are a few numerical examples:

$$\begin{pmatrix} 1 & 2 & -3 \end{pmatrix} \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix} = \begin{pmatrix} -6 \end{pmatrix};$$

$$\begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix} \begin{pmatrix} 1 & 2 & -3 \end{pmatrix} = \begin{pmatrix} -1 & -2 & 3 \\ 2 & 4 & -6 \\ 3 & 6 & -9 \end{pmatrix};$$

$$\begin{pmatrix} 1 & 0 & 2 \\ 0 & -1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 \\ 2 & 0 & 0 \\ 1 & 0 & -1 \end{pmatrix} = \begin{pmatrix} 3 & 1 & -1 \\ 1 & 0 & -3 \end{pmatrix};$$

$$\begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 5 & -9 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

To repeat, we can multiply matrices *only* in this case:

$$\underset{m\times p}{A} \text{ times } \underset{p\times n}{B} = \underset{m\times n}{AB}.$$

There is an *identity* for this matrix multiplication. We'll say that I is the $m \times m$ (notice its shape is *square*) matrix

$$I = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & & \\ 0 & 0 & \dots & 1 \end{pmatrix}.$$

(The context will always determine the size of the identity matrix I.) Then

$$\underset{m\times n}{A} \text{ times } \underset{n\times n}{I} = \underset{m\times n}{A},$$

$$\underset{m\times m}{I} \text{ times } \underset{m\times n}{A} = \underset{m\times n}{A}.$$

It is sometimes useful to employ the *Kronecker delta* function to work with I: by definition

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases}$$

Thus,

$$I = (\delta_{ij}).$$

Here are some properties of multiplication:

- it is *associative*:
$$(AB)C = A(BC).$$
- it is *distributive*:
$$(A + B)C = AC + BC,$$
$$A(B + C) = AB + AC.$$
- $0A = A0 = 0$ (the three "zeros" may all be different!).
- it is *not commutative*:
$$AB \neq BA \text{ in general.}$$

The last situation is quite interesting. In the first place, AB is defined only if # of columns of A = # of rows of B; and BA is defined only if # of columns of B = # of rows of A. Thus AB and BA are both defined only if A is $m \times n$ and B is $n \times m$; then AB is $m \times m$ and BA is $n \times n$. Thus $AB = BA$ is possible only if $m = n$. Thus the only possible situation for $AB = BA$ is that A and B are both square matrices of the same size. But even then they might not "commute":

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix},$$
$$\begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

The latter situation also shows that we cannot in general "cancel" matrices from an equation:

$$AB = 0 \not\Rightarrow A = 0 \text{ or } B = 0.$$

PROBLEM 2–75. Show that matrix multiplication does not in general have multiplicative inverses, by showing that there is no 2×2 matrix A such that

$$A \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},$$

even though $\begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$ is not zero.

PROBLEM 2–76. Suppose A is an $n \times n$ matrix which commutes with every $n \times n$ matrix. That is, $AB = BA$ for every $n \times n$ matrix B. Prove that A is a scalar multiple of I.

Column vectors. In presenting the general definition of derivative it is helpful to introduce a convention for writing linear functions from $\mathbb{R}^n$ to $\mathbb{R}^m$. The convention is described as follows: in these situations we modify the notational scheme for points in our basic vector space $\mathbb{R}^n$ by expressing them as *columns* rather than *rows*, so that we write

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \text{an } n \times 1 \text{ matrix.}$$

The reason for doing this is simple. We want to have a compact expression for a linear function from $\mathbb{R}^n$ to $\mathbb{R}^m$. If A is an $m \times n$ matrix, then the matrix product Ay is a well-defined $m \times 1$ matrix. Thus matrix multiplication produces the desired linear function $y \mapsto Ay$.

(If we used the row vector notation we have been following, then the corresponding linear function would have to send $1 \times n$ matrices to $1 \times m$ matrices, so it would need to be written in the form $y \mapsto yA$, where A is an $n \times m$ matrix. As we usually prefer thinking of functions as operating on the left, this would go against our custom. Hence the sudden change to column vectors.)

Thus if y is an $n \times 1$ column vector, and A is an $m \times n$ matrix, then Ay is the column vector in $\mathbb{R}^m$ whose i^{th} entry is

$$\sum_{k=1}^{n} a_{ik} y_k.$$

As a special case, note the interesting result involving $\hat{e}_j$, the j^{th} coordinate vector for $\mathbb{R}^n$:

$$A\hat{e}_j = \text{ the } j^{\text{th}} \text{ column of } A.$$

Here of course we are writing $\hat{e}_j$ as a column vector:

$$\hat{e}_j = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \quad (n \times 1 \text{ matrix, 1 in position } j).$$

PROBLEM 2–77. Notice that in the above situation, Ay is certainly a *linear* function of y, in the sense of the definition at the beginning of this section. This exercise establishes the converse. Namely, show that if $\mathbb{R}^n \xrightarrow{F} \mathbb{R}^m$ is linear, then there is a unique $m \times n$ matrix A such that

$$F(y) = Ay \quad \text{for all} \quad y \in \mathbb{R}^n.$$

I want to stress that the matrix A is not the linear function here — it's not a function at all. It is just that when A multiplies vectors as above, the resulting function F is linear.

PROBLEM 2–78. Consider this situation:

$$\mathbb{R}^n \xrightarrow{G} \mathbb{R}^p \xrightarrow{F} \mathbb{R}^m,$$

where F is linear and G is linear. Let the corresponding matrices be A and B, so that

$$F(x) = Ax \quad \text{for all} \quad x \in \mathbb{R}^p,$$
$$G(y) = By \quad \text{for all} \quad y \in \mathbb{R}^n.$$

Prove that the composite function $F \circ G$ is linear, and that its corresponding matrix is the product AB.

REMARK. Problem 2–78 gives the actual reason for the strange-looking definition of matrix multiplication.

J. **Derivatives for functions** $\mathbb{R}^n \to \mathbb{R}^m$

In this section we are going to attain our goal of defining derivatives in general. Recall that in Section A we very easily took care of the case $n = 1$, which essentially was like one-variable calculus; the generalization to values of functions in $\mathbb{R}^m$ rather than $\mathbb{R}$ was quite simple.

However, the case $m = 1$ and general n was quite another matter. We devoted Sections C, D, and E just to getting the definition of differentiability correct.

Based on the above two paragraphs, you might expect that our current generalization from $m = 1$ to general m will be completely straightforward. This is indeed the case. We can even take our cue from the great definition on p. 2–24, as we recognize that the crucial form of the numerator,

$$f(x + y) - f(x) - c \bullet y,$$

involves the *linear* function of y represented there by the scalar product. Literally, all we need

to do now is insert the correct form of linear function of y, namely one that maps $\mathbb{R}^n$ into $\mathbb{R}^m$. We know from Section I that such linear functions can be realized as products of the form Ay, where A is an $m \times n$ matrix and $y \in \mathbb{R}^n$ is written as an $n \times 1$ matrix (a *column* vector). Here then is what we require:

★ **DEFINITION.** ★ **Let $x \in \mathbb{R}^n$ be a fixed point. Assume $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ is defined at least in a neighborhood of x (a ball $B(x,r)$). Then f is *differentiable* at x if there exists an $m \times n$ matrix A such that**

$$\lim_{y \to 0} \frac{f(x+y) - f(x) - Ay}{\|y\|} = 0.$$

DISCUSSION

1. Notice that the numerator in this definition makes good sense, as $f(x+y)$, $f(x)$, and Ay all belong to $\mathbb{R}^m$.

2. This isn't quite like p. 2–24 in case $m = 1$. This is only because in the former definition we wrote our linear function in the form of the scalar product $c \bullet y$. The corresponding form we are now using in this case

$$Ay = \underset{1\times n \text{ matrix}}{(c_1\ c_2 \dots c_n)} \quad \underset{n\times 1 \text{ matrix}}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}.$$

   This is of course equal to the number

$$c_1 y_1 + c_2 y_2 + \cdots + c_n y_n,$$

   and in our old notation this is indeed the scalar product

$$(c_1, c_2, \dots, c_n) \bullet (y_1, y_2, \dots, y_n).$$

3. The remark on p. 2–25 still holds in this more general case, and asserts that the affine function of y given by $f(x) + Ay$ is a very good approximation to the function $f(x+y)$ for small $\|y\|$.

4. The proof of p. 2–27 still applies to show that *if f is differentiable at x, then f is continuous at x.*

5. Directional derivatives work the same way as before. If we restrict y to have the form th for a fixed $h \in \mathbb{R}^n$ and let $t \to 0$, we obtain the formula

$$Df(x;h) = Ah, \quad \text{all} \quad h \in \mathbb{R}^n.$$

6. **TERMINOLOGY**. The matrix A is called the *Jacobian matrix* of f at x. This is in honor of the great mathematician,

Carl Gustav Jakob Jacobi, 1804–1851.

We shall use either notation for this matrix,

$$(Df)(x) \quad \text{or} \quad f'(x).$$

7. In particular, let $h = \hat{e}_j =$ the j^{th} coordinate vector. Then we have

$$\frac{\partial f}{\partial x_j}(x) = Df(x)\hat{e}_j = j^{\text{th}} \text{ column of } Df(x)$$

(see p. 2–51). That is, the n columns of $Df(x)$ are given respectively by the n partial derivatives $\partial f/\partial x_1, \ldots, \partial f/\partial x_n$. We can embellish this observation as follows. Write the function f in terms of its coordinate functions (of course, arranged in a column!):

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{pmatrix}.$$

Then the j^{th} column of $Df(x)$ is

$$\begin{pmatrix} \partial f_1/\partial x_j \\ \partial f_2/\partial x_j \\ \vdots \\ \partial f_m/\partial x_j \end{pmatrix}.$$

This gives the all-important formula for the Jacobian matrix of f in terms of partial derivatives,

$$(Df)(x) = \begin{pmatrix} \partial f_1/\partial x_1 & \ldots & \partial f_1/\partial x_n \\ \partial f_2/\partial x_1 & \ldots & \partial f_2/\partial x_n \\ \vdots & & \vdots \\ \partial f_m/\partial x_1 & \ldots & \partial f_m/\partial x_n \end{pmatrix},$$

where of course all the partial derivatives are evaluated at x.

8. The wonderfully useful sufficient condition for differentiability as given on p. 2–30 is still in force. This means that if $f_1, \ldots, f_m$ all have continuous partial derivatives with respect to $x_1, \ldots, x_n$, then we have differentiability of f and we can use the formula for the Jacobian matrix as given above.

9. In particular, if all the coordinate functions f_i are of class C^1 in $B(x, r)$, then f is differentiable at every point of $B(x, r)$.

10. In the special case of a real-valued function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$, we now have two distinct notions for the derivative. One is the gradient ∇f and the other is the Jacobian matrix Df. It is perhaps unfortunate that these are different, but they are very similar. The Jacobian matrix is by defintion the $1 \times n$ matrix

$$Df(x) = (f_{x_1}\ f_{x_2} \ldots f_{x_n}).$$

(No commas!) On the other hand, since we are thinking of vectors in $\mathbb{R}^n$ as column vectors, we should probably now write

$$\nabla f(x) = \begin{pmatrix} f_{x_1} \\ f_{x_2} \\ \vdots \\ f_{x_n} \end{pmatrix}.$$

There is very little danger of confusion, for the distinction in these two concepts is that between algebra and geometry. The directional derivative is given either way as

$$Df(x; h) = Df(x)h \qquad \text{(matrix product)}$$

and

$$Df(x; h) = \nabla f(x) \bullet h \qquad \text{(dot product)}.$$

EXAMPLE. Here we work out the Jacobian matrices associated with *polar coordinates* for $\mathbb{R}^2$. Denoting $\mathbb{R}^2$ as the $x - y$ plane, the formulas are

$$\begin{cases} x = & r\cos\theta, \\ y = & r\sin\theta. \end{cases}$$

We insist that $0 < r < \infty$, so we are leaving out the origin in $\mathbb{R}^2$. Of course, $-\infty < \theta < \infty$, though a point $(x, y) \in \mathbb{R}^2$ determines θ only up to the addition of integer multiples of 2π.

In order to conform with our column vector notation, we need to write our formulas in column form in terms of a function $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^2$ as

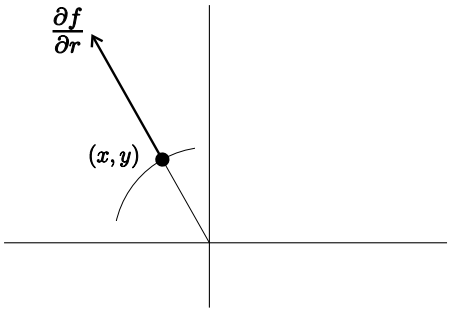
$$f(r, \theta) = \begin{pmatrix} r\cos\theta \\ r\sin\theta \end{pmatrix}.$$

Then

$$Df(r, \theta) = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}.$$

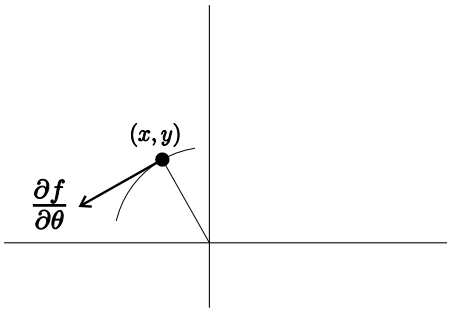
(A small but important point: we have arbitrarily chosen to write the independent variables r, θ in the order displayed. If we had written them in the order θ, r, then the columns of our matrix would be interchanged.)

The two columns of this Jacobian matrix have tremendous geometric significance. Notice first that they are orthogonal. We can think of f as depicting points in the $x - y$ plane as functions of r, θ. Thus $\partial f/\partial r$ describes how the point $f(r, \theta)$ varies with increasing r for fixed θ:



(Notice that $\left\| \frac{\partial f}{\partial r} \right\| = 1$.)

On the other hand, $\partial f/\partial\theta$ describes the motion of $f(r, \theta)$ with increasing θ for fixed r:



(Notice that $\left\| \frac{\partial f}{\partial \theta} \right\| = r$.)

The function f assigns Cartesian coordinates if the polar coordinates are given. Conversely, we now consider the function g which assigns polar coordinates to a point given in Cartesian coordinates:

$$g(x, y) = \begin{pmatrix} r \\ \theta \end{pmatrix},$$

where the formulas $x = r\cos\theta$, $y = r\sin\theta$ need to be solved for r, θ. Of course,

$$r = \sqrt{x^2 + y^2};$$

as for θ, it is only determined "modulo 2π." If we use a formula like $\theta = \arctan(y/x)$, we can readily compute the partial derivatives as

$$\frac{\partial \theta}{\partial x} = \frac{-y}{x^2 + y^2}, \quad \frac{\partial \theta}{\partial y} = \frac{x}{x^2 + y^2}.$$

(There's another way to find these partial derivatives. Namely, start with $x = r\cos\theta$, $y = r\sin\theta$, and compute directly. For instance, keeping y fixed and using the subscript notation for partial derivatives,

$$\begin{cases} 1 &= r_x \cos\theta - r\sin\theta\theta_x, \\ 0 &= r_x \sin\theta + r\cos\theta\theta_x \end{cases}.$$

Now eliminate the "unknown" r_x from this pair of equations:

$$(-\sin\theta)1 + (\cos\theta)0 = r\sin^2\theta\theta_x + r\cos^2\theta\theta_x.$$

Thus

$$-\sin\theta = r\theta_x,$$

so that

$$\theta_x \;=\; \frac{-\sin\theta}{r} \;=\; \frac{-r\sin\theta}{r^2} \;=\; \frac{-y}{x^2+y^2}.$$

Likewise for θ_y. This method has the advantage of avoiding any problems with arctangent when $x = 0$.)

Thus we find the Jacobian matrix

$$Dg(x,y) = \begin{pmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ \frac{-y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{pmatrix}.$$

In terms of the polar coordinates themselves this may be written

$$Dg(x,y) = \begin{pmatrix} \cos\theta & \sin\theta \\ \frac{-\sin\theta}{r} & \frac{\cos\theta}{r} \end{pmatrix}.$$

Notice that $Df(r,\theta)$ and $Dg(x,y)$ are *inverse* matrices:

$$\begin{aligned} Df(r,\theta)Dg(x,y) &= \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ \frac{-\sin\theta}{r} & \frac{\cos\theta}{r} \end{pmatrix} \\ &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \\ &= I. \end{aligned}$$

(We shall see in Section K that this is an illustration of the *chain rule*.)

PROBLEM 2–79. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^n$ be the function given in Problem 2–36:

$$f(x) = \frac{x}{\|x\|^2}.$$

Calculate the Jacobian matrix $Df(\hat{e}_1)$.

PROBLEM 2–80. For the function of the preceding problem show that

$$Df(x) = \|x\|^{-2}I - 2\|x\|^{-4}(x_i x_j).$$

PROBLEM 2–81. For the function of the preceding problem show that for any u, $v \in \mathbb{R}^n$

$$Df(x)u \bullet Df(x)v = \|x\|^{-4} u \bullet v.$$

(Because of this result, the function f is said to be *conformal*.)

PROBLEM 2–82. Suppose $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ is itself a linear function: $f(x) = Ax$. Show that for any $x \in \mathbb{R}^n$

$$Df(x) = A.$$

(In particular, if $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^n$ is the identity function, $f(x) = x$ for all x, then $Df(x) = I$.)

PROBLEM 2–83. For any $m \times n$ matrix $A = (a_{ij})$, let A^t be its *transpose*, namely the $n \times m$ matrix

$$A^t = (a_{ji}).$$

Show that for all x, y in the appropriate spaces

$$Ax \bullet y = x \bullet A^t y$$

(what *are* the appropriate spaces?).

PROBLEM 2–84. Let A be an $n \times n$ matrix and let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be the corresponding *quadratic form* on $\mathbb{R}^n$,

$$f(x) = Ax \bullet x.$$

Show that

$$\nabla f(x) = (A + A^t)x.$$

(In particular, if A is *symmetric*, meaning $A^t = A$, then

$$\nabla f(x) = 2Ax.)$$

PROBLEM 2–85. Suppose $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ and $\mathbb{R}^n \xrightarrow{g} \mathbb{R}^m$ are both differentiable at x. Show that $f + g$ is also differentiable at x, and

$$D(f + g)(x) = Df(x) + Dg(x).$$

PROBLEM 2–86. Here's another product rule. Suppose $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ and $\mathbb{R}^n \xrightarrow{g} \mathbb{R}^m$ are both differentiable at x. Show that fg is also differentiable at x, and

$$D(fg)(x) = f(x)Dg(x) + g(x)Df(x).$$

PROBLEM 2–87. Another product rule: if $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ and $\mathbb{R}^n \xrightarrow{g} \mathbb{R}^m$, then show that

$$D(f \bullet g) = f^t Dg + g^t Df.$$

(Here f^t and g^t refer to the transposed values, so that $f^t(x)$ and $g^t(x)$ are $1 \times m$ matrices (row vectors)).

K. **The chain rule**

The material in this section is very, very important in calculus and its applications, as the results are used constantly in both theory and practice. It all has to do with the basic concept of *composition* of two functions. In general, whenever two functions f and g are given and it makes sense that we are able to define $g(f(x))$, we write the resulting function $g \circ f$:

$$(g \circ f)(x) = g(f(x)).$$

You are surely used to thinking this way, though perhaps not with this notation, from single-variable calculus.

Our abstract framework will involve this situation:

$$\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m \xrightarrow{g} \mathbb{R}^\ell,$$

where as usual we do not require f and g to be defined everywhere. If $n = m = \ell = 1$, then we are in the familiar single-variable calculus situation and we certainly recognize the chain

rule in the form

$$\frac{d}{dx}\left(g(f(x))\right) = g'(f(x))f'(x).$$

Another notation you are probably familiar with is something like

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

This result emphasizes the expected result that the derivative of the composition is the product of the derivatives of the two involved functions.

There is a more geometric way of thinking of this in terms of affine approximations. In the more general situation we are investigating, we think of the function $y \to f(x) + Df(x)y$ as the "best" affine approximation of $f(x+y)$ near $y = 0$. Let us temporarily express this approximation in the following notation:

$$f(x+y) \doteq f(x) + Ay,$$

where $A = Df(x)$. Likewise, if we denote $B = Dg(f(x))$, then near $z = 0$ we have the affine approximation

$$g(f(x)+z) \doteq g(f(x)) + Bz.$$

Therefore we definitely anticipate that we have the approximation near $y = 0$,

$$\begin{aligned}(g \circ f)(x+y) &= g(f(x+y)) \\ &\doteq g(f(x) + Ay) \\ &\doteq g(f(x)) + BAy.\end{aligned}$$

The last expression is an affine function of y and indicates that

$$\begin{aligned}D(g \circ f)(x) &= BA \\ &= Dg(f(x))Df(x).\end{aligned}$$

This is indeed what we shall prove. There's a wonderful moral here: while the composition $g \circ f$ may be very difficult to compute, involving all sorts of complicated operations, the affine approximation of $g \circ f$ is very easy to compute, involving only the very basic algebra of multiplying matrices!

We now state and prove what is also sometimes more accurately termed "the composite function theorem." Our proof does not rely on the single-variable calculus result, but actually handles that as a special case. Though it's a special case, all the essential ingredients are present even there.

THE CHAIN RULE. *Assume*

$$\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m \xrightarrow{g} \mathbb{R}^\ell.$$

Let x be a fixed point in $\mathbb{R}^n$. Assume

$$\begin{aligned} &f \quad \textit{is differentiable at } x; \\ &g \quad \textit{is differentiable at } f(x). \end{aligned}$$

Then

$$g \circ f \quad \textit{is differentiable at } x$$

and

$$\boxed{D(g \circ f)(x) = Dg(f(x))Df(x)}.$$

This formula relating the derivatives is really beautiful, containing not only the three Jacobian matrices, but also expressing the result in terms of the nice definition of matrix multiplication:

$$\underset{\ell\times n}{D(g \circ f)} \;=\; \underset{\ell\times m}{Dg} \quad \underset{m\times n}{Df}.$$

The proof is not hard at all, essentially merely using the definition of differentiability.

PROOF. As above, we denote $A = Df(x)$ ($m\times n$ matrix) and $B = Dg(f(x))$ ($\ell\times m$ matrix). The first step in the proof is going to express g as the sum of two functions, one of which is linear and the other of which varies so little near $f(x)$ that it contributes zero to the derivative at $f(x)$. Namely, we write $g = g_1 + g_2$, where

$$\begin{aligned} g_1(w) &= Bw \quad (\text{a } \textit{linear} \text{ function}), \\ g_2(w) &= g(w) - Bw. \end{aligned}$$

Notice that

$$\begin{aligned} Dg_1(f(x)) &= B \qquad (\text{Problem 2–82}), \\ Dg_2(f(x)) &= Dg(f(x)) - B \qquad (\text{Problem 2–85}) \\ &= B - B \\ &= 0. \end{aligned}$$

Now $g \circ f = g_1 \circ f + g_2 \circ f$, so again we can use Problem 2–85 once we prove that $g_1 \circ f$ and $g_2 \circ f$ are differentiable at x. First we have a straightforward algebra calculation

$$\begin{aligned}
&\lim_{y\to 0} \frac{(g_1 \circ f)(x+y) - (g_1 \circ f)(x) - BAy}{\|y\|} \\
&\quad = \lim_{y\to 0} \frac{Bf(x+y) - Bf(x) - BAy}{\|y\|} \\
&\quad = \lim_{y\to 0} \frac{B\left(f(x+y) - f(x) - Ay\right)}{\|y\|} \\
&\quad = B \lim_{y\to 0} \frac{f(x+y) - f(x) - Ay}{\|y\|} \\
&\quad = B0 \\
&\quad = 0.
\end{aligned}$$

This proves that $g_1 \circ f$ is differentiable at x and

$$D(g_1 \circ f)(x) = BA.$$

Now we are going to show that $g_2 \circ f$ is differentiable at x and

$$D(g_2 \circ f)(x) = 0.$$

(This will finish the proof, thanks to Problem 2–85.) That is, we are going to prove that

$$\lim_{y\to 0} \frac{g_2\left(f(x+y)\right) - g_2\left(f(x)\right)}{\|y\|} \;=\; 0.$$

This is completely expected, thanks to the fact that $Dg_2(f(x)) = 0$. We just need to check that the presence of f doesn't cause any unwelcome surprises.

We first claim that f satisfies a *Lipschitz* condition at the point x. Namely (see Problem 2–44), there exists a constant C and a positive number δ such that

$$\|f(x+y) - f(x)\| \le C\|y\|$$

for all $\|y\| < \delta$. (We'll prove this at the end.)

Now for $\|y\| < \delta$ we consider the quotient

$$\frac{\|g_2\left(f(x+y)\right) - g_2\left(f(x)\right)\|}{\|y\|} \qquad (*)$$

If it happens to equal zero for a particular y, then we are of course quite pleased, as we want to show $(*)$ has limit zero. So we need only be concerned with those y for which $\|y\| < \delta$ and $(*) \neq 0$. In particular, $f(x+y) - f(x) \neq 0$. Let us call this latter difference w. Of course, w depends on y and in fact $\|w\| \leq C\|y\|$. In this situation

$$\begin{aligned}(*) &= \frac{\|g_2(f(x)+w) - g_2(f(x))\|}{\|w\|} \cdot \frac{\|w\|}{\|y\|} \\ &\leq \frac{\|g_2(f(x)+w) - g_2(f(x))\|}{\|w\|} \cdot C.\end{aligned}$$

This last quantity has limit zero as $y \to 0$, since also $w \to 0$ and we are given that $Dg_2(f(x)) = 0$.

Finally we establish the Lipschitz condition for f. We first notice that the entries of the fixed matrix A are just fixed numbers, and thus there is a number C_0 such that we have the inequality for norms,

$$\|Ay\| \leq C_0\|y\| \qquad \text{for all } \; y \in \mathbb{R}^n.$$

Next, the triangle inequality implies

$$\begin{aligned}\frac{\|f(x+y) - f(x)\|}{\|y\|} &\leq \frac{\|f(x+y) - f(x) - Ay\|}{\|y\|} + \frac{\|Ay\|}{\|y\|} \\ &\leq \frac{\|f(x+y) - f(x) - Ay\|}{\|y\|} + C_0;\end{aligned}$$

this sum has limit C_0 as $y \to 0$, because $A = Df(x)$. Thus it is no larger than, say, $1 + C_0 = C$ for all sufficiently small y (say, $\|y\| < \delta$).

QED

PROBLEM 2–88. In a general situation $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ in which f is differentiable at a fixed point x, define the affine approximation to be the function $\mathrm{Aff}(f, x)$, where

$$\mathrm{Aff}(f,x)(u) = f(x) + Df(x)(u - x) \qquad \text{for all } \; u \in \mathbb{R}^n.$$

(Why?) Then prove as a result of the chain rule

$$\mathrm{Aff}(g \circ f, x) = \mathrm{Aff}(g, f(x)) \circ \mathrm{Aff}(f, x).$$

ILLUSTRATIONS.

1. Look again at the polar coordinate example on pp. 2–55 through 2–58. There we have a situation where f and g are inverses of one another, so that $f \circ g =$ the identity function from $\mathbb{R}^2$ to $\mathbb{R}^2$. Thus $D(f \circ g)(x, y) = I$ by Problem 2–82. The chain rule thus gives

$$Df\,(g(x, y))\,Dg(x, y) = I,$$

just as we observed by explicit calculation on p. 2–58.

2. More generally, any time we have a situation $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^n \xrightarrow{g} \mathbb{R}^n$ in which f and g are *inverses* in the sense that $g \circ f =$ the identity function, then

$$Dg\,(f(x))\,Df(x) = I.$$

(We say that the corresponding Jacobian matrices are *inverses* of one another.) This comes as no surprise. After all, if f and g are inverses then their affine approximations should also be inverses, thanks to Problem 2–88.

3. The most important abstract situation to understand is the case $n = 1$, $\ell = 1$:

$$\mathbb{R} \xrightarrow{f} \mathbb{R}^m \xrightarrow{g} \mathbb{R}.$$

We shall in fact show in the two subsequent items that the generalization to arbitrary n and ℓ is then immediate. For this illustration we shall write generically $f = f(t)$ and $g = g(u) = g(u_1, \ldots, u_m)$. The chain rule then tells us immediately

$$D(g \circ f)(t) = Dg\,(f(t))\,Df(t).$$

That is,

$$\begin{aligned}\frac{d}{dt}g\,(f(t)) &= \left(\frac{\partial g}{\partial u_1}\,(f(t)) \quad . \quad . \quad . \quad \frac{\partial g}{\partial u_m}\,(f(t))\right)\begin{pmatrix} df_1/dt \\ \vdots \\ df_m/dt \end{pmatrix} \\ &= \sum_{k=1}^{m} \frac{\partial g}{\partial u_k}\,(f(t))\,\frac{df_k}{dt}.\end{aligned}$$

It seems to help in remembering this formula to abuse the notation by writing $f(t)$ as $u(t)$. The idea is that the *independent variables* u_k have been replaced by *functions* $u_k(t)$ and the resulting calculus formula is

$$\frac{d}{dt}\Big(g\,(u_1(t), \ldots, u_m(t))\Big) = \sum_{k=1}^{m} \frac{\partial g}{\partial u_k}\,\frac{du_k}{dt}.$$

So you can really just use the single-variable chain rule as a pattern and just keep differentiating "as long as you see a t."

4. The generalization
$$\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m \xrightarrow{g} \mathbb{R}$$
is immediate, as the act of computing $\partial/\partial x_i$ is a matter of letting the other coordinates be fixed and applying number 3:
$$\frac{\partial}{\partial x_i}\Big(g\left(u_1(x),\ldots,u_m(x)\right)\Big) = \sum_{k=1}^{m} \frac{\partial g}{\partial u_k}\ \frac{\partial u_k}{\partial x_i},$$
an equation that is valid for each i between 1 and n.

5. The full generalization
$$\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m \xrightarrow{g} \mathbb{R}^\ell$$
now follows simply by applying number 4 to each component of g, one at a time:
$$\frac{\partial}{\partial x_i}\Big(g_j\left(u_1(x),\ldots,u_m(x)\right)\Big) = \sum_{k=1}^{m} \frac{\partial g_j}{\partial u_k}\ \frac{\partial u_k}{\partial x_i}$$
valid for $1 \le i \le n$ and $1 \le j \le \ell$.

6. The special case
$$\mathbb{R}^n \xrightarrow{f} \mathbb{R} \xrightarrow{g} \mathbb{R}$$
is often quite useful. We have
$$\frac{\partial}{\partial x_i} g\left(f(x)\right) = g'\left(f(x)\right) \frac{\partial f}{\partial x_i}.$$
In terms of the *gradient* notation,
$$\nabla(g \circ f) = g'\left(f(x)\right) \nabla f(x).$$
For instance,
$$\nabla(e^f) = e^f \nabla f,$$
$$\nabla(g(\|x\|)) = g'(\|x\|)\frac{x}{\|x\|}.$$

PROBLEM 2–89. Here is another proof that the pathological function of Problem 2–43 is indeed not differentiable at the origin. Calculate for that function that

$$\frac{d}{dt} f(t, t^2) \Big|_{t=0} = \frac{1}{2}.$$

Why does this show that f is not differentiable at 0?

PROBLEM 2–90. Define $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ by

$$f(x,y) = \begin{cases} \frac{x^2|y|^{5/4}}{x^4+y^2} & \text{for } (x,y) \neq (0,0), \\ 0 & \text{for } (x,y) = (0,0). \end{cases}$$

Show that f is continuous on $\mathbb{R}^2$ and that all directional derivatives $Df(0;h) = 0$. Then prove that f is not differentiable at the origin by consideration of $f(t, t^2)$.

PROBLEM 2–91. This is a rather standard situation that frequently arises in thermodynamics and other applications. Suppose $\mathbb{R}^3 \xrightarrow{F} \mathbb{R}$ is a differentiable function whose first order partial derivatives are never zero. Furthermore, suppose that the equation

$$F(x,y,z) = 0$$

can be solved for each of the three "unknowns" as functions of the other two variables. For example, in this way we can regard x as a function of y and z, so with abuse of notation

$$F(x,y,z) = 0 \text{ produces } x = x(y,z).$$

It then makes sense to define $\partial x/\partial y$, the partial derivative of $x(y,z)$ with z held fixed. Prove that

$$\frac{\partial x}{\partial y} \cdot \frac{\partial y}{\partial z} \cdot \frac{\partial z}{\partial x} = -1.$$

PROBLEM 2–92. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^2$ be defined by

$$f(x_1, x_2) = (x_1^2 - x_2^2, 2x_1x_2).$$

Calculate the 2×2 Jacobian matrix $Df(x)$.

PROBLEM 2–93. The function of the preceding problem actually comes from the corresponding *complex* function $(x_1 + ix_2)^2$. Since every complex number other than 0 has two square roots, the equation $f(x) = y$ should have two distinct solutions for each $y \neq 0$ in $\mathbb{R}^2$. Find them explicitly. (Part of the answer is

$$x_1 = \pm\sqrt{\frac{\|y\| + y_1}{2}}.)$$

PROBLEM 2–94. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^2$ be defined by

$$f(x_1, x_2) = (x_1^3 - 3x_1x_2^2,\ 3x_1^2x_2 - x_2^3),$$

and calculate $Df(x)$. What complex function does this resemble?

PROBLEM 2–95. The complex exponential function e^z produces through Euler's formula the function $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^2$ given as

$$f(x_1, x_2) = (e^{x_1} \cos x_2, e^{x_1} \sin x_2).$$

Calculate the Jacobian matrix $Df(x_1, x_2)$.

PROBLEM 2–96. Given $y \neq 0$ in $\mathbb{R}^2$ one can define a *complex logarithm* of $y_1 + iy_2$ to be any complex number $x_1 + ix_2$ such that $e^{x_1+ix_2} = y_1 + iy_2$. In terms of $\mathbb{R}^2$ and the function of the preceding exercise, this means that $f(x) = y$. Find all such points x. Roughly speaking, the answer is

$$x = \left(\log \|y\|, \arctan \frac{y_2}{y_1} \right).$$

PROBLEM 2–97. Pretend that the formula of Problem 2–96 gives a (single-valued) function $x = g(y)$. Verify directly from that and Problem 2–95 the relation

$$Dg(f(x))Df(x) = I.$$

PROBLEM 2–98. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}^4$ be defined by

$$f(\alpha, \beta) = (\cos\alpha, \sin\alpha, \cos\beta, \sin\beta).$$

Calculate the 4×2 Jacobian matrix $Df(\alpha, \beta)$. Show that the two columns of this matrix are unit vectors orthogonal to one another.

L. **Confession**

"Confess your faults to one another"
James 5^{16}

Although we have given a correct definition of the differentiability at x of a function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$, together with the Jacobian matrix $Df(x)$, that is not quite the whole truth. The reason is that we have relied on the identification of linear functions with their corresponding matrices.

The better definition ignores matrices entirely and just mentions the linear functions. Thus we can equivalently define f to be differentiable at x if there exists a *linear function* $\mathbb{R}^n \xrightarrow{L} \mathbb{R}^m$ such that

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

The corresponding terminology might be this: the linear function L is called the *differential* of f at x, and is denoted

$$(df)(x) = L.$$

The correspondence with the Jacobian matrix $Df(x)$ is that

$$\underset{m\times 1}{(df)(x)(y)} = \underset{m\times n}{Df(x)}\ \ \underset{n\times 1}{y} \qquad \text{for all } \ y \in \mathbb{R}^n.$$

Thus for example we have

$$(df)(x)(\hat{e}_j) = \frac{\partial f}{\partial x_j}(x).$$

PROBLEM 2–99.
a. If $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ is linear, show that $df(x) = f$.
b. If $\mathbb{R}^n \xrightarrow{f} \mathbb{R}^m$ is affine, show that $df(x) = f - f(0)$.

PROBLEM 2–100. Use the notation of the statement of the chain rule as given in Section K. Show that

$$d(g \circ f)(x) = dg(f(x)) \circ df(x).$$

What is the point of this shift in emphasis? It is that we gain in geometric insight by stressing the geometric object $df(x)$ instead of the algebraic object $Df(x)$. But not only that. We often have situations in which the $\hat{e}_j$'s are not the natural basic vectors to be using and the given coordinates are somehow unnatural. Then the actual $m \times n$ matrix $Df(x)$ may be of little interest, and we might prefer using a different $m \times n$ matrix.

Notice also that the statement of the chain rule is more elegant in the new formulation, as both sides of the formula involve composition of functions. The algebra involved in matrix multiplication does not appear.

In summary, we might say that the *linear function* $df(x)$ *is represented by the Jacobian matrix* $Df(x)$ in the usual coordinate systems we are using.

M. **Homogeneous functions and Euler's formula**

In this section we are concerned with functions $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ which are defined on all of $\mathbb{R}^n$ except the origin. Let a be a fixed real number.

DEFINITION. The function is *homogeneous of degree a* if

$$f(tx) = t^a f(x) \quad \text{for all} \quad 0 < t < \infty \quad \text{and all} \quad x \in \mathbb{R}^n - \{0\}. \tag{$*$}$$

Assume from now on that f is of class C^1 on the set $\mathbb{R}^n - \{0\}$.

PROBLEM 2–101. Prove that if f is homogeneous of degree a, then the partial derivatives $\partial f/\partial x_j$ are homogeneous of degree $a - 1$.

PROBLEM 2–102. For fixed $x \neq 0$ differentiate the equation $(*)$ with respect to t. Then set $t = 1$ and conclude that the *Euler equation* is satisfied:

$$\sum_{j=1}^{n} x_j \frac{\partial f}{\partial x_j} = af. \tag{$**$}$$

PROBLEM 2–103. Conversely, assume that the Euler equation $(**)$ is satisfied and then prove that f is homogeneous of degree a. (HINT: $\frac{d}{dt}(t^{-a} f(tx))$.)

PROBLEM 2–104. Assume f is a *polynomial* which is homogeneous of degree a, where a is of course a nonnegative integer. Establish the Euler equation for f by explicitly calculating what happens for the individual monomials

$$x_1^{a_1} x_2^{a_2} \dots x_n^{a_n}, \quad a_1 + a_2 + \cdots + a_n = a.$$

PROBLEM 2–105. Suppose f is homogeneous of degree 1, and define $f(0) = 0$. Assume that f is differentiable at 0. Prove that

$$f(x) = \nabla f(0) \bullet x.$$

PROBLEM 2–106. Give an example of a function which is homogeneous of degree 1 and continuous on $\mathbb{R}^n$ and not differentiable at 0.