

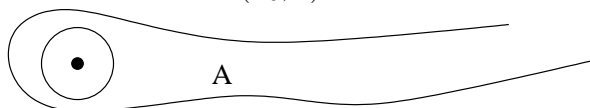
Chapter 3 Higher order derivatives

You certainly realize from single-variable calculus how very important it is to use derivatives of orders greater than one. The same is of course true for multivariable calculus. In particular, we shall definitely want a “second derivative test” for critical points.

A. Partial derivatives

First we need to clarify just what sort of domains we wish to consider for our functions. So we give a couple of definitions.

DEFINITION. Suppose $x_0 \in A \subset \mathbb{R}^n$. We say that x_0 is an *interior point* of A if there exists $r > 0$ such that $B(x_0, r) \subset A$.



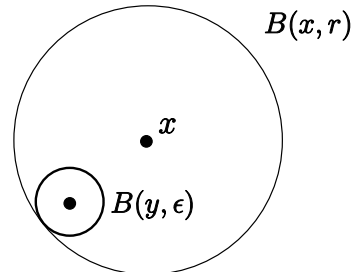
DEFINITION. A set $A \subset \mathbb{R}^n$ is said to be *open* if every point of A is an interior point of A .

EXAMPLE. An open ball is an open set. (So we are fortunate in our terminology!) To see this suppose $A = B(x, r)$. Then if $y \in A$, $\|x - y\| < r$ and we can let

$$\epsilon = r - \|x - y\|.$$

Then $B(y, \epsilon) \subset A$. For if $z \in B(y, \epsilon)$, then the triangle inequality implies

$$\begin{aligned} \|z - x\| &\leq \|z - y\| + \|y - x\| \\ &< \epsilon + \|y - x\| \\ &= r, \end{aligned}$$



so that $z \in A$. Thus $B(y, \epsilon) \subset B(x, r)$ (cf. Problem 1–34). This proves that y is an interior point of A . As y is an arbitrary point of A , we conclude that A is open.

PROBLEM 3–1. Explain why the empty set is an open set and why this is actually a problem about the logic of language rather than a problem about mathematics.

PROBLEM 3–2. Prove that if A and B are open subsets of \mathbb{R}^n , then their union $A \cup B$ and their intersection $A \cap B$ are also open.

PROBLEM 3–3. Prove that the closed ball $\overline{B}(x, r)$ is not open.

PROBLEM 3–4. Prove that an *interval* in \mathbb{R} is an open subset \iff it is of the form (a, b) , where $-\infty \leq a < b \leq \infty$. Of course, $(a, b) = \{x \in \mathbb{R} \mid a < x < b\}$.

PROBLEM 3–5. Prove that an interval in \mathbb{R} is never an open subset of \mathbb{R}^2 . Specifically, this means that the x -axis is not an open subset of the $x - y$ plane. More generally, prove that if we regard \mathbb{R}^k as a subset of \mathbb{R}^n , for $1 \leq k \leq n - 1$, by identifying \mathbb{R}^k with the Cartesian product $\mathbb{R}^k \times \{0\}$, where 0 is the origin in \mathbb{R}^{n-k} , then no point of \mathbb{R}^k is an interior point relative to the “universe” \mathbb{R}^n .

NOTATION. Suppose $A \subset \mathbb{R}^n$ is an open set and $A \xrightarrow{f} \mathbb{R}$ is differentiable at every point $x \in A$. Then we say that f is *differentiable on* A . Of course, we know then that in particular the partial derivatives $\partial f / \partial x_j$ all exist. It may happen that $\partial f / \partial x_j$ itself is a differentiable function on A . Then we know that its partial derivatives also exist. The notation we shall use for the latter partial derivatives is

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i} \left(\frac{\partial f}{\partial x_j} \right).$$

Notice the care we take to denote the *order* in which these differentiations are performed. In case $i = j$ we also write

$$\frac{\partial^2 f}{\partial x_i^2} = \frac{\partial^2 f}{\partial x_i \partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{\partial f}{\partial x_i} \right).$$

By the way, if we employ the subscript notation (p. 2–20) for derivatives, then we have logically

$$f_{x_j x_i} = (f_{x_j})_{x_i} = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Partial derivatives of higher orders are given a similar notation. For instance, we have

$$\frac{\partial^7 f}{\partial x^3 \partial y \partial z^2 \partial x} = f_{xzzzyxxx}.$$

PROBLEM 3–6. Let $(0, \infty) \times \mathbb{R} \xrightarrow{f} \mathbb{R}$ be defined as

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

Compute f_{xx} , f_{xy} , f_{yx} , f_{yy} .

PROBLEM 3–7. Let $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ be defined for $x \neq 0$ as $f(x) = \log \|x\|$. Show that

$$\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} = 0.$$

This is called *Laplace's equation* on \mathbb{R}^2 .

PROBLEM 3–8. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be defined for $x \neq 0$ as $f(x) = \|x\|^{2-n}$. Show that

$$\sum_{j=1}^n \frac{\partial^2 f}{\partial x_j^2} = 0.$$

(Laplace's equation on \mathbb{R}^n .)

PROBLEM 3–9. Let $\mathbb{R} \xrightarrow{\varphi} \mathbb{R}$ be a differentiable function whose derivative φ' is also differentiable. Define $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ as

$$f(x, y) = \varphi(x + y).$$

Show that

$$\frac{\partial^2 f}{\partial x^2} - \frac{\partial^2 f}{\partial y^2} = 0.$$

Do the same for $g(x, y) = \varphi(x - y)$.

Now suppose $A \subset \mathbb{R}^n$ is open and $A \xrightarrow{f} \mathbb{R}$ is differentiable at every point $x \in A$. Then we say f is *differentiable on A* . The most important instance of this is the case in which all the partial derivatives $\partial f / \partial x_j$ exist and are continuous functions on A — recall the important

corollary on p. 2–32. We say in this case that f is of *class* C^1 and write

$$f \in C^1(A).$$

(In case f is a continuous function on A , we write

$$f \in C^0(A).)$$

Likewise, suppose $f \in C^1(A)$ and each $\partial f/\partial x_j \in C^1(A)$; then we say

$$f \in C^2(A).$$

In the same way we define recursively $C^k(A)$, and we note of course the string of inclusions

$$\cdots \subset C^3(A) \subset C^2(A) \subset C^1(A) \subset C^0(A).$$

The fact is that most functions encountered in calculus have the property that they have continuous partial derivatives of all orders. That is, they belong to the class $C^k(A)$ for all k . We say that these functions are *infinitely (often) differentiable*, and we write the collection of such functions as $C^\infty(A)$. Thus we have

$$C^\infty(A) = \bigcap_{k=0}^{\infty} C^k(A)$$

and

$$C^\infty(A) \subset \cdots \subset C^2(A) \subset C^1(A) \subset C^0(A).$$

PROBLEM 3–10. Show that even in single-variable calculus all the above inclusions are strict. That is, for every k show that there exists $f \in C^k(\mathbb{R})$ for which $f \notin C^{k+1}(\mathbb{R})$.

We now turn to an understanding of the basic properties of $C^2(A)$. If $f \in C^2(A)$, then we have defined the “*pure*” partial derivatives

$$\frac{\partial^2 f}{\partial x_i \partial x_i} = \frac{\partial}{\partial x_i} \left(\frac{\partial f}{\partial x_i} \right)$$

as well as the “*mixed*” partial derivatives

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i} \left(\frac{\partial f}{\partial x_j} \right).$$

for $i \neq j$. Our next task is the proof that if $f \in C^2(A)$, then

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

(“the mixed partial derivatives are equal”). This result will clearly render calculations involving higher order derivatives much easier; we’ll no longer have to keep track of the order of computing partial derivatives. Not only that, there are fewer that must be computed:

PROBLEM 3–11. If $f \in C^2(\mathbb{R}^2)$, then only three second order partial derivatives of f need to be computed in order to know all four of its second order partial derivatives. Likewise, if $f \in C^2(\mathbb{R}^3)$, then only six need to be computed in order to know all nine. What is the result for the general case $f \in C^2(\mathbb{R}^n)$?

What is essentially involved in proving a result like this is showing that the two limiting processes involved can be done in either order. You will probably not be surprised, then, that there is pathology to be overcome. The following example is found in just about every text on vector calculus. It is sort of a canonical example.

PROBLEM 3–12. Define $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ by

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

Prove that

- a. $f \in C^1(\mathbb{R}^2)$,
- b. $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial x \partial y}$, $\frac{\partial^2 f}{\partial y \partial x}$, $\frac{\partial^2 f}{\partial y^2}$ exist on \mathbb{R}^2 ,
- c. $\frac{\partial^2 f}{\partial x \partial y}(0, 0) \neq \frac{\partial^2 f}{\partial y \partial x}(0, 0)$.

All pathology of this nature can be easily eliminated by a reasonable assumption on the function, as we now demonstrate.

THEOREM. *Let $A \subset \mathbb{R}^n$ be an open set and let $f \in C^2(A)$. Then*

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

PROOF. Since we need only consider a fixed pair i, j in the proof, we may as well assume $i = 1, j = 2$. And since x_3, \dots, x_n remain fixed in all our deliberations, we may also assume that $n = 2$, so that $A \subset \mathbb{R}^2$.

Let $x \in A$ be fixed, and let $\delta > 0$ and $\epsilon > 0$ be arbitrary but small enough that the points considered below belong to A (remember, x is an interior point of A).

The first step in our proof is the writing of a combination of four values of f near the point x . Then we shall perform a limiting procedure to obtain the desired result. Before defining the crucial expression we present the following guideline to our reasoning. Thus we hope that approximately

$$\begin{aligned} \frac{\partial f}{\partial x_1}(x_1, x_2) &\approx \frac{f(x_1 + \delta, x_2) - f(x_1, x_2)}{\delta}, \\ \frac{\partial f}{\partial x_1}(x_1, x_2 + \epsilon) &\approx \frac{f(x_1 + \delta, x_2 + \epsilon) - f(x_1, x_2 + \epsilon)}{\delta}. \end{aligned}$$

Now subtract and divide by ϵ :

$$\begin{aligned} \frac{\partial^2 f}{\partial x_2 \partial x_1}(x_1, x_2) &\approx \frac{\frac{\partial f}{\partial x_1}(x_1, x_2 + \epsilon) - \frac{\partial f}{\partial x_1}(x_1, x_2)}{\epsilon} \\ &\approx \frac{\frac{f(x_1 + \delta, x_2 + \epsilon) - f(x_1, x_2 + \epsilon)}{\delta} - \frac{f(x_1 + \delta, x_2) - f(x_1, x_2)}{\delta}}{\epsilon}. \end{aligned}$$

Thus we hope that

$$\frac{\partial^2 f}{\partial x_2 \partial x_1} \approx \delta^{-1} \epsilon^{-1} [f(x_1 + \delta, x_2 + \epsilon) - f(x_1, x_2 + \epsilon) - f(x_1 + \delta, x_2) + f(x_1, x_2)].$$

Now we come to the actual proof. Let

$$D(\delta, \epsilon) = f(x_1 + \delta, x_2 + \epsilon) - f(x_1, x_2 + \epsilon) - f(x_1 + \delta, x_2) + f(x_1, x_2).$$

Then define the auxiliary function (this is a trick!) of one real variable,

$$g(y) = f(x_1 + \delta, y) - f(x_1, y).$$

By the mean value theorem of single-variable calculus,

$$g(x_2 + \epsilon) - g(x_2) = \epsilon g'(t)$$

for some $x_2 < t < x_2 + \epsilon$. That is,

$$D(\delta, \epsilon) = \epsilon \left[\frac{\partial f}{\partial x_2}(x_1 + \delta, t) - \frac{\partial f}{\partial x_2}(x_1, t) \right].$$

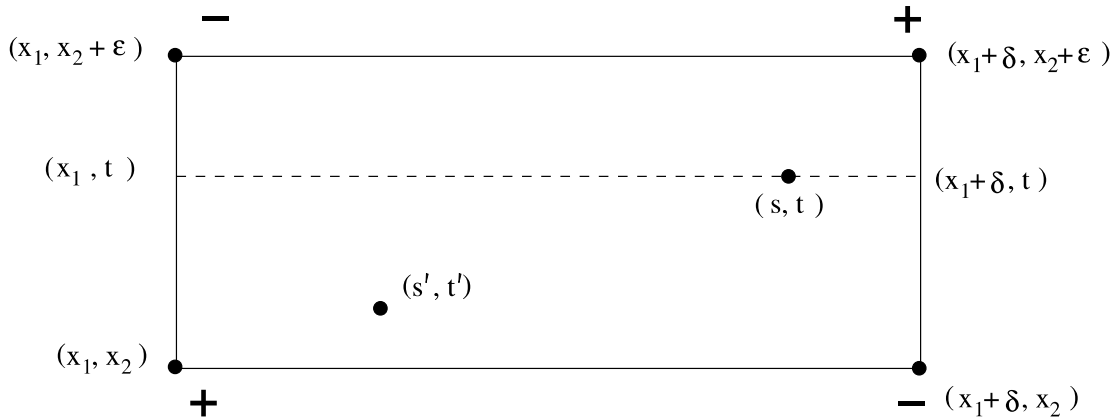
Now we use the mean value theorem once again, this time for the term in brackets:

$$\frac{\partial f}{\partial x_2}(x_1 + \delta, t) - \frac{\partial f}{\partial x_2}(x_1, t) = \delta \frac{\partial^2 f}{\partial x_1 \partial x_2}(s, t)$$

for some $x_1 < s < x_1 + \delta$. The result is

$$D(\delta, \epsilon) = \delta \epsilon \frac{\partial^2 f}{\partial x_1 \partial x_2}(s, t).$$

Here is a diagram illustrating this situation:



Now carry out the very same procedure, except with the roles of x_1 and x_2 interchanged. The result can be immediately written down as

$$D(\delta, \epsilon) = \delta \epsilon \frac{\partial^2 f}{\partial x_2 \partial x_1}(s', t').$$

Comparing the two equations we have thus obtained, we find after dividing by $\delta\epsilon$,

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(s, t) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(s', t').$$

Finally, we let $\epsilon \rightarrow 0$ and at this time use the continuity of the two mixed partial derivatives to obtain

$$\frac{\partial^2 f}{\partial x_1 \partial x_2}(x) = \frac{\partial^2 f}{\partial x_2 \partial x_1}(x).$$

QED

REMARK. The proof we have presented certainly does not require the full hypothesis that $f \in C^2(D)$. It requires only that each of $\partial^2 f / \partial x_i \partial x_j$ and $\partial^2 f / \partial x_j \partial x_i$ be continuous at the point in question. As I have never seen an application of this observation, I have chosen to state the theorem in its weaker version. This will provide all we shall need in practice. However, the weakened hypothesis does indeed give an interesting result. Actually, the same proof outline provides even better results with very little additional effort:

PROBLEM 3–13*. Assume only that $\partial f / \partial x_i$ and $\partial f / \partial x_j$ exist in A and that $\partial^2 f / \partial x_i \partial x_j$ exists in A and is continuous at x . Prove that $\partial^2 f / \partial x_j \partial x_i$ exists at x and that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \quad \text{at } x.$$

(HINT: start with the equation derived above,

$$\delta^{-1}\epsilon^{-1}D(\delta, \epsilon) = \frac{\partial^2 f}{\partial x_1 \partial x_2}(s, t).$$

The right side of this equation has a limit as δ and ϵ tend to zero. Therefore so does the left side. Perform this limit in the iterated fashion

$$\lim_{\epsilon \rightarrow 0} \left(\lim_{\delta \rightarrow 0} (\text{left hand side}) \right)$$

to obtain the result.)

PROBLEM 3–14*. Assume only that $\partial f/\partial x_i$ and $\partial f/\partial x_j$ exist in A and are themselves differentiable at x . Prove that

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \quad \text{at } x.$$

(Hint: start with the expression for $D(\delta, \epsilon)$ on p. 3–6. Then use the hypothesis that $\partial f/\partial x_2$ is differentiable at x to write

$$\begin{aligned} \frac{\partial f}{\partial x_2}(x_1 + \delta, t) &= \frac{\partial f}{\partial x_2}(x_1, x_2) + \frac{\partial^2 f}{\partial x_1 \partial x_2}(x_1, x_2)\delta + \frac{\partial^2 f}{\partial x_2^2}(x_1, x_2)(t - x_2) \\ &\quad + \text{small multiple of } (\delta + (t - x_2)) \\ \frac{\partial f}{\partial x_2}(x_1, t) &= \frac{\partial f}{\partial x_2}(x_1, x_2) + \frac{\partial^2 f}{\partial x_2^2}(x_1, x_2)(t - x_2) \\ &\quad + \text{small multiple of } (t - x_2). \end{aligned}$$

Conclude that

$$D(\delta, \epsilon) = \epsilon \delta \frac{\partial^2 f}{\partial x_1 \partial x_2} + \epsilon \text{ (small multiple of } (\delta + \epsilon))$$

and carefully analyze how small the remainder is. Conclude that the limit of $\epsilon^{-2}D(\epsilon, \epsilon)$ as ϵ tends to zero is the value of $\partial^2 f/\partial x_1 \partial x_2$ at (x_1, x_2) .

PROBLEM 3–15. Prove that if f is of class C^k on an open set A , then any mixed partial derivative of f of order k can be computed without regard to the order of performing the various differentiations.

B. Taylor's theorem (Brook Taylor, 1712)

We are now preparing for our study of critical points of functions on \mathbb{R}^n , and we specifically want to devise a “second derivative test” for the type of critical point. The efficient theoretical way to study this question is to approximate the given function by an expression (the Taylor polynomial) which is a *polynomial* of degree 2. The device for doing this is called *Taylor's theorem*.

We begin with a quick review of the single-variable calculus version. Suppose that g is a

function from \mathbb{R} to \mathbb{R} which is defined and of class C^2 near the origin. Then the fundamental theorem of calculus implies

$$g(t) = g(0) + \int_0^t g'(s)ds.$$

Integrate partially after a slight adjustment:

$$\begin{aligned} g(t) &= g(0) - \int_0^t g'(s)d(t-s) \quad (t \text{ is fixed}) \\ &= g(0) - g'(s)(t-s) \Big|_{s=0}^{s=t} + \int_0^t g''(s)(t-s)ds \\ &= g(0) + g'(0)t + \int_0^t g''(s)(t-s)ds. \end{aligned}$$

Now we slightly adjust the remaining integral by writing

$$g''(s) = g''(0) + (g''(s) - g''(0))$$

and doing the explicit integration with the $g''(0)$ term:

$$\begin{aligned} \int_0^t g''(s)(t-s)ds &= g''(0) \int_0^t (t-s)ds + R \\ &= g''(0) \frac{(t-s)^2}{-2} \Big|_{s=0}^{s=t} + R \\ &= \frac{1}{2}g''(0)t^2 + R. \end{aligned}$$

Here R , the “remainder,” is of course

$$\int_0^t (g''(s) - g''(0))(t-s)ds.$$

Suppose

$$|g''(s) - g''(0)| \leq Q \quad \text{for } s \text{ between } 0 \text{ and } t.$$

Then

$$|R| \leq \left| \int_0^t Q(t-s)ds \right| = \frac{1}{2}Qt^2.$$

(The absolute value signs around the integral are necessary if $t < 0$.) So the version of Taylor’s theorem we obtain is

$$g(t) = g(0) + g'(0)t + \frac{1}{2}g''(0)t^2 + R,$$

where

$$|R| \leq \frac{1}{2}Qt^2.$$

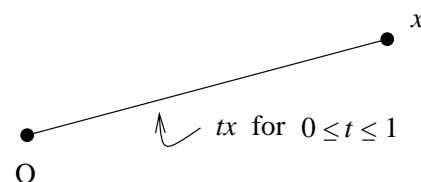
(By the way, if g is of class C^3 , then $|g''(s) - g''(0)| \leq C|s|$ for some constant C (Lipschitz condition), so that

$$|R| \leq \left| \int_0^t Cs(t-s)ds \right| = \frac{1}{6}C|t|^3.$$

Thus the remainder term in the Taylor series tends to zero one degree faster than the quadratic terms in the Taylor polynomial. This is the usual way of thinking of Taylor approximations.)

Now we turn to the case $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$, and we work first with the Taylor series centered at the origin. (We can easily extend to other points later.) We assume f is of class C^2 near 0. If $x \in \mathbb{R}^n$ is close to 0, we use the familiar ploy of moving between 0 and x along a straight line: we thus define the composite function

$$g(t) = f(tx),$$



regarding x as fixed but small. We then have as above, with $t = 1$,

$$g(1) = g(0) + g'(0) + \frac{1}{2}g''(0) + R,$$

where

$$|R| \leq \frac{1}{2}Q,$$

where

$$|g''(s) - g''(0)| \leq Q \quad \text{for } 0 \leq s \leq 1.$$

Now we apply the chain rule:

$$g'(t) = \sum_{i=1}^n (D_i f)(tx)x_i$$

($D_i f = \partial f / \partial x_i$) and again

$$g''(t) = \sum_{i=1}^n \sum_{j=1}^n (D_j D_i f)(tx)x_j x_i.$$

In particular,

$$g'(0) = \sum_{i=1}^n D_i f(0) x_i,$$

$$g''(0) = \sum_{i,j=1}^n D_i D_j f(0) x_i x_j.$$

We now see what the statement of theorem should be:

THEOREM. Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be of class C^2 near 0. Let $\epsilon > 0$. Then there exists $\delta > 0$ such that if $\|x\| \leq \delta$, then

$$f(x) = f(0) + \sum_{i=1}^n D_i f(0) x_i + \frac{1}{2} \sum_{i,j=1}^n D_i D_j f(0) x_i x_j + R,$$

where

$$|R| \leq \epsilon \|x\|^2.$$

PROOF. We continue with the above development. Because f is of class C^2 , its second order partial derivatives are all continuous. Hence given $\epsilon > 0$, we may choose $\delta > 0$ such that for all $\|y\| \leq \delta$,

$$|D_i D_j f(y) - D_i D_j f(0)| \leq \epsilon.$$

Then if $\|x\| \leq \delta$, we have a pretty crude estimate valid for $0 \leq s \leq 1$:

$$\begin{aligned} |g''(s) - g''(0)| &= \left| \sum_{i,j=1}^n (D_i D_j f(sx) - D_i D_j f(0)) x_i x_j \right| \\ &\leq \sum_{i,j=1}^n \epsilon |x_i| |x_j| \\ &= \epsilon \left(\sum_{i=1}^n |x_i| \right)^2 \\ &\leq \epsilon \sum_{i=1}^n 1 \cdot \sum_{i=1}^n x_i^2 \quad (\text{Schwarz inequality}) \\ &= \epsilon n \|x\|^2. \end{aligned}$$

The remainder term above therefore satisfies

$$|R| \leq \frac{1}{2}\epsilon n \|x\|^2.$$

To obtain the final result, first replace ϵ by $2\epsilon/n$.

QED

B'. Taylor's theorem *bis*

There is actually a much different proof of Taylor's theorem which enables us to weaken the hypothesis significantly and still obtain the same conclusion. As far as applications to calculus are concerned, the C^2 version we have proved above is certainly adequate, but the improved version presented here is nevertheless quite interesting for its minimal hypothesis.

THEOREM. *Let $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ be differentiable in a neighborhood of 0, and assume each partial derivative $\partial f/\partial x_i$ is differentiable at 0. Let $\epsilon > 0$. Then there exists $\delta > 0$ such that if $\|x\| \leq \delta$, then*

$$f(x) = f(0) + \sum_{i=1}^n D_i f(0)x_i + \frac{1}{2} \sum_{i,j=1}^n D_i D_j f(0)x_i x_j + R,$$

where

$$|R| \leq \epsilon \|x\|^2.$$

PROOF. We still use the function $g(t) = f(tx)$, and we still have the chain rule

$$g'(t) = \sum_{i=1}^n D_i f(tx)x_i.$$

However, now we use the hypothesis that $D_i f$ is differentiable at 0 as follows: for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$D_i f(y) = D_i f(0) + \sum_{j=1}^n D_j D_i f(0)y_j + R_i(y),$$

where

$$|R_i(y)| \leq \epsilon \|y\| \quad \text{for all } \|y\| \leq \delta.$$

Therefore, if $\|x\| \leq \delta$ and $0 \leq t \leq 1$, we obtain

$$g'(t) = \sum_{i=1}^n D_i f(0)x_i + \sum_{i=1}^n \sum_{j=1}^n D_j D_i f(0)tx_j x_i + R(t, x),$$

where

$$\begin{aligned} |R(t, x)| &= \left| \sum_{i=1}^n x_i R_i(tx) \right| \\ &\leq \sum_{i=1}^n |x_i| \epsilon t \|x\| \\ &\leq \epsilon t \sqrt{n} \|x\|^2. \end{aligned}$$

At this point we would like to apply a theorem from single-variable calculus known as the Cauchy mean value theorem. This is used in most proofs of l'Hôpital's rule. However, instead of appealing to this theorem we insert its proof in this context. Define

$$\varphi(t) = g(t) - g(0) - g'(0)t - [g(1) - g(0) - g'(0)]t^2,$$

noting that $\varphi(0) = \varphi(1) = 0$ and that φ is differentiable for $0 \leq t \leq 1$. Thus the mean value theorem implies $\varphi'(t) = 0$ for some $0 < t < 1$. That is,

$$\begin{aligned} [g(1) - g(0) - g'(0)]2t &= g'(t) - g'(0) \\ &= \sum_{i,j=1}^n D_j D_i f(0)tx_j x_i + R(t, x). \end{aligned}$$

Dividing by $2t$ gives

$$f(x) - f(0) - \sum_{i=1}^n D_i f(0)x_i = \frac{1}{2} \sum_{i,j=1}^n D_j D_i f(0)x_j x_i + \frac{R(t, x)}{2t},$$

where

$$\left| \frac{R(t, x)}{2t} \right| \leq \frac{\epsilon}{2} \sqrt{n} \|x\|^2.$$

QED

PROBLEM 3–16. Here’s a simplification of the proof of the theorem in case $n = 1$. Calculate

$$\lim_{x \rightarrow 0} \frac{f(x) - f(0) - f'(0)x}{x^2}$$

and then deduce the theorem.

The hypothesis of this version of Taylor’s theorem is rather minimal. It is strong enough to guarantee the equality of the mixed second order partial derivatives of f at 0, as we realize from Problem 3–14 (although we did not require this fact in the proof). The following problem shows that there is little hope of further weakening the hypothesis.

PROBLEM 3–17. Let f be the “canonical” pathological function of Problem 3–12. This function is of class C^1 on \mathbb{R}^2 and has the origin as a critical point. Show that there do not exist constants c_1, c_2, c_3 such that

$$f(x, y) = c_1x^2 + c_2xy + c_3y^2 + R(x, y)$$

and

$$\lim_{(x,y) \rightarrow (0,0)} \frac{R(x, y)}{x^2 + y^2} = 0.$$

C. The second derivative test for \mathbb{R}^2

To get started with this analysis, here is a problem which you very much need to investigate carefully.

PROBLEM 3–18. Let A, B, C be real numbers such that

$$A > 0, \quad C > 0, \quad AC - B^2 > 0.$$

a. Prove that a number $\lambda > 0$ exists such that for all $(x, y) \in \mathbb{R}^2$,

$$Ax^2 + 2Bxy + Cy^2 \geq \lambda(x^2 + y^2).$$

b. Calculate the largest possible λ .

(Answer: $\frac{A+C}{2} - \sqrt{(\frac{A+C}{2})^2 - (AC - B^2)}$)

Now we are ready to give the second derivative test. We suppose that we are dealing with a function $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ which is of class C^2 in a neighborhood of the origin (we shall later immediately generalize to any point). Using coordinates (x, y) for \mathbb{R}^2 , we obtain from Taylor's theorem in Section B

$$\begin{aligned} f(x, y) = & \underbrace{f(0, 0)}_{0 \text{ order term}} + \underbrace{D_1 f(0, 0)x + D_2 f(0, 0)y}_{1^{\text{st}} \text{ order terms}} \\ & + \underbrace{\frac{1}{2} (D_{11} f(0, 0)x^2 + 2D_{12} f(0, 0)xy + D_{22} f(0, 0)y^2)}_{2^{\text{nd}} \text{ order terms}} + \underbrace{R(x, y)}_{\text{remainder term}}, \end{aligned}$$

and the remainder term satisfies the condition that for any $\epsilon > 0$ there exists $\delta > 0$ such that

$$x^2 + y^2 \leq \delta^2 \Rightarrow |R(x, y)| \leq \epsilon(x^2 + y^2).$$

(Here we have continued to employ the notation $D_{11}f = \partial^2 f / \partial x^2$ etc.)

Now we suppose that $(0, 0)$ is a *critical point* for f . That is, by definition,

$$D_1 f(0, 0) = D_2 f(0, 0) = 0.$$

Let us now denote

$$\begin{aligned} A &= D_{11}f(0, 0) = \partial^2 f / \partial x^2(0, 0), \\ B &= D_{12}f(0, 0) = \partial^2 f / \partial x \partial y(0, 0), \\ C &= D_{22}f(0, 0) = \partial^2 f / \partial y^2(0, 0). \end{aligned}$$

Then the Taylor expression for f takes the form

$$f(x, y) = f(0, 0) + \frac{1}{2}(Ax^2 + 2Bxy + Cy^2) + R(x, y).$$

The underlying rationale for the second derivative test is that in the Taylor expression for f the first order terms are absent, thanks to the fact that we have a critical point; therefore we expect that the second order terms will determine the behavior of the function near the critical point. In fact, this usually happens, as we now show.

Let us first assume that the numbers A, B, C satisfy the inequalities

$$\begin{aligned} A > 0 \quad \text{and} \quad C > 0, \\ AC - B^2 > 0. \end{aligned}$$

We know from Problem 3–18 that there exists a positive number λ (depending only on A, B, C) such that

$$Ax^2 + 2Bxy + Cy^2 \geq \lambda(x^2 + y^2) \quad \text{for all } (x, y) \in \mathbb{R}^2.$$

Therefore, $x^2 + y^2 \leq \delta^2 \Rightarrow$

$$\begin{aligned} f(x, y) &\geq f(0, 0) + \frac{1}{2}\lambda(x^2 + y^2) - |R(x, y)| \\ &\geq f(0, 0) + \frac{1}{2}\lambda(x^2 + y^2) - \epsilon(x^2 + y^2) \\ &= f(0, 0) + \left(\frac{1}{2}\lambda - \epsilon\right)(x^2 + y^2). \end{aligned}$$

This inequality is exactly what we need! For we may choose $0 < \epsilon < \frac{1}{2}\lambda$ and then conclude with the corresponding δ that

$$0 < x^2 + y^2 \leq \delta^2 \Rightarrow f(x, y) > f(0, 0).$$

The conclusion: **at the critical point 0, f has a strict local minimum.** (“Strict” in this context means that near $(0, 0)$ the equality $f(x, y) = f(0, 0)$ holds only if $(x, y) = (0, 0)$.)

If we assume instead that

$$\begin{aligned} A < 0 \quad \text{and} \quad C < 0, \\ AC - B^2 > 0, \end{aligned}$$

then we obtain the corresponding conclusion: **at the critical point 0, f has a strict local maximum.** The proof is immediate: we could either repeat the above type of analysis, or we could simply replace f by $-f$ and use the previous result. For replacing f by $-f$ also replaces A, B, C by their negatives, so that the assumed condition for f becomes the previous condition for $-f$.

Finally, assume

$$AC - B^2 < 0.$$

The quadratic form in question, $Ax^2 + 2Bxy + Cy^2$, can now be positive at some points, negative at others. We check this quickly. If $A \neq 0$, then the equation $Ax^2 + 2Bxy + Cy^2 = 0$ has distinct roots

$$\frac{x}{y} = \frac{-B \pm \sqrt{B^2 - AC}}{A}$$

and therefore changes sign. The reason is clear. Renaming x/y as t , the function $At^2 + 2Bt + C$ has a graph which is a parabola (since $A \neq 0$). It hits the t axis at two distinct points, so that its position is that of a parabola that genuinely crosses the t -axis. Likewise if $C \neq 0$. If both A and C are 0, then we are dealing with $2Bxy$, which has opposite signs at (say) $(1, 1)$ and $(1, -1)$. (Notice that our assumption implies $B \neq 0$ in this case.)

Now suppose (x_0, y_0) satisfies

$$Ax_0^2 + 2Bx_0y_0 + Cy_0^2 = \alpha < 0.$$

Then for small $|t|$ we have

$$\begin{aligned} f(tx_0, ty_0) &= f(0, 0) + \frac{1}{2}\alpha t^2 + R(tx_0, ty_0) \\ &\leq f(0, 0) + \frac{1}{2}\alpha t^2 + \epsilon t^2(x_0^2 + y_0^2) \\ &= f(0, 0) + t^2 \left(\frac{1}{2}\alpha + \epsilon(x_0^2 + y_0^2) \right). \end{aligned}$$

Thus if ϵ is small enough that $\frac{1}{2}\alpha + \epsilon(x_0^2 + y_0^2) < 0$ and $|t|$ is sufficiently small and not zero,

$$f(tx_0, ty_0) < f(0, 0).$$

Thus in any neighborhood of $(0, 0)$ there are points where $f < f(0, 0)$.

In the same way we show that in any neighborhood of $(0, 0)$ there are points where $f > f(0, 0)$. The conclusion: **at the critical point 0, f has neither a local minimum nor a local maximum value.**

To recapitulate, all the above analysis is based on the same observation: at a critical point $(0, 0)$, the behavior of f is governed by the *quadratic* terms in its Taylor expansion, at least in the three cases we have considered. These three cases are all subcategories of the assumption

$$AC - B^2 \neq 0.$$

For if $AC - B^2 > 0$, then $AC > 0$ and we see that A and C are either both positive or both negative. If $AC - B^2 < 0$, that is the third case.

The conclusion is that, if $AC - B^2 \neq 0$, then the quadratic terms in the Taylor expansion of f are determinative. If $AC - B^2 = 0$, we have no criterion.

For example, the function

$$f(x, y) = x^2 + y^4$$

has the origin as a strict local minimum; its negative has the origin as a strict local maximum. And the function

$$f(x, y) = x^2 - y^4$$

has a critical point at the origin, which is neither a local minimum nor a local maximum. All these examples satisfy $AC - B^2 = 0$.

PROBLEM 3–19. Find three functions $f(x, y)$ which have $(0, 0)$ as a critical point and have $AC = B = 1$ (so that $AC - B^2 = 0$) such that the critical point $(0, 0)$ is

- a strict local minimum for one of them,
- a strict local maximum for another,
- neither a local minimum nor a local maximum for the other.

D. The nature of critical points

This section contains mostly definitions. We suppose that $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ and that f is of class C^1 near a critical point x_0 . Thus $\nabla f(x_0) = 0$. There are just three types of behavior for f near x_0 :

f has a local minimum at x_0 :

$$f(x) \geq f(x_0) \text{ for all } x \in B(x_0, \epsilon) \text{ for some } \epsilon > 0.$$

f has a local maximum at x_0 :

$$f(x) \leq f(x_0) \text{ for all } x \in B(x_0, \epsilon) \text{ for some } \epsilon > 0.$$

f has a saddle point at x_0 :

neither of the above two conditions holds — in other words, for all $\epsilon > 0$ there exist $x, x' \in B(x_0, \epsilon)$ such that $f(x) < f(x_0) < f(x')$.

There is a further refinement in the first two cases:

f has a strict local minimum at x_0 :

$f(x) > f(x_0)$ for all $0 < \|x - x_0\| < \epsilon$ for some $\epsilon > 0$.
 f has a strict local maximum at x_0 :
 $f(x) < f(x_0)$ for all $0 < \|x - x_0\| < \epsilon$ for some $\epsilon > 0$.

In Section C we analyzed the \mathbb{R}^2 case when the critical point was the *origin*. That is really no loss of generality in view of the correspondence between the function f and its translated version g given by

$$g(x) = f(x_0 + x), \quad x \in \mathbb{R}^n.$$

For f has a critical point at $x_0 \iff g$ has a critical point at 0. Moreover, the nature of the critical points is the same. Moreover,

$$D_i D_j g(0) = D_i D_j f(x_0),$$

so the information coming from the second derivatives is the same for f and g .

SUMMARY. We take this opportunity to summarize the \mathbb{R}^2 case. Assume

$\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ is of class C^2 near x_0 , and x_0 is a critical point of f .

Define

$$\begin{aligned}
 A &= D_1 D_1 f(x_0), \\
 B &= D_1 D_2 f(x_0), \\
 C &= D_2 D_2 f(x_0).
 \end{aligned}$$

Assume

$$AC - B^2 \neq 0.$$

Then we have the results:

$$A > 0 \quad \text{and} \quad C > 0,$$

$\implies f$ has a strict local minimum at x_0 .

$$AC - B^2 > 0$$

$$A < 0 \quad \text{and} \quad C < 0,$$

$\implies f$ has a strict local maximum at x_0 .

$$AC - B^2 > 0$$

$$AC - B^2 < 0$$

$\implies f$ has a saddle point at x_0 .

In case $AC - B^2 = 0$, it is *impossible* to tell from A, B, C what the nature of the critical point is. Here are some easy examples for the critical point $(0, 0)$ in the $x - y$ plane:

1. $f(x, y) = x^2$ local minimum
2. $f(x, y) = -x^2$ local maximum (these are even global extrema)
3. $f(x, y) = x^2 + y^4$ strict local minimum
4. $f(x, y) = -x^2 - y^4$ strict local maximum
5. $f(x, y) = x^2 - y^4$ saddle point

In each of the next five problems determine the nature of each critical point of the indicated function on \mathbb{R}^2 .

PROBLEM 3–20. $f(x, y) = \frac{1}{3}x^3 + \frac{1}{2}y^2 + 2xy + 5x + y.$
(Answer: one local minimum, one saddle)

PROBLEM 3–21. $f(x, y) = x^2y^2 - 5x^2 - 8xy - 5y^2.$
(Answer: one local maximum, four saddles)

PROBLEM 3–22. $f(x, y) = xy(12 - 3x - 4y).$

PROBLEM 3–23. $f(x, y) = x^4 + y^4 - 2x^2 + 4xy - 2y^2.$

PROBLEM 3–24. $f(x, y) = (ax^2 + by^2)e^{-x^2-y^2}$ ($a > b > 0$ are constants).

PROBLEM 3–25. The function of Problem 2–56 has just the critical point $(1, 1)$. Use the above test to determine its local nature.
(Answer: local minimum)

PROBLEM 3–26. The function of Problem 2–64 has just the critical point $(1, 0)$, and it was shown to be a local maximum. Use the above test to determine its local nature.

PROBLEM 3–27. Here’s a function given me by Richard Stong:

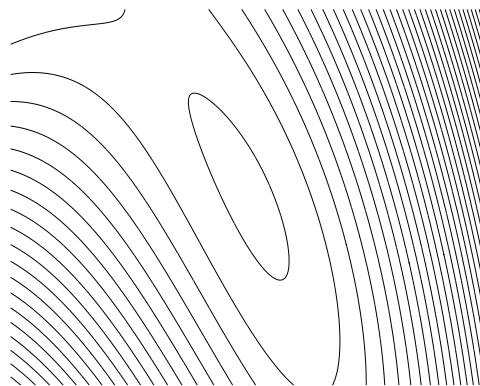
$$f(x, y) = 24x + 6x^2y^2 + 6x^2y - x^2 + x^3y^3.$$

Show that it has just one critical point, which is a strict local maximum but not a global maximum. (Stong’s point: this provides the same *moral* as does Problem 2–64, but with a *polynomial* function.)

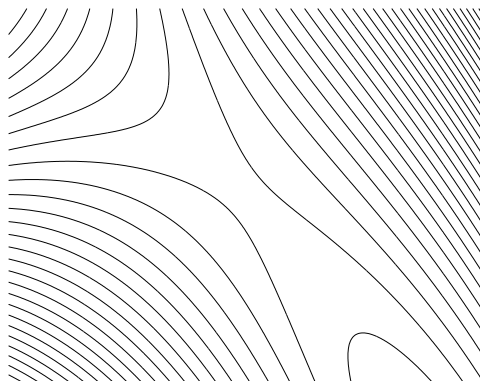
There is a beautiful geometric way to think of these criteria, in terms of level sets. For instance, suppose $\mathbb{R}^2 \xrightarrow{f} \mathbb{R}$ has a critical point at $(0, 0)$, and that $A = 2$, $B = 0$, $C = 1$, so we know the critical point is a minimum. If f were just a quadratic polynomial, then

$$f(x, y) = f(0, 0) + x^2 + \frac{1}{2}y^2.$$

The level set description of f then is concentric ellipses. In case f is not a quadratic polynomial, then near the critical point the level sets will still resemble ellipses, but with some distortion. The distortion decreases as we approach the critical point.



Likewise, the level sets of a quadratic polynomial which is a saddle point are concentric hyperbolas, but in general they will resemble distorted hyperbolas.



E. The Hessian matrix

Now we return to the general n -dimensional case, so we shall be assuming throughout this section that

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

is a function of class C^2 near its critical point x_0 . The Taylor expansion from Section B therefore becomes

$$f(x_0 + y) = f(x_0) + \frac{1}{2} \sum_{i,j=1}^n D_i D_j f(x_0) y_i y_j + R(y),$$

where for any $\epsilon > 0$ the remainder satisfies

$$|R(y)| \leq \epsilon \|y\|^2 \quad \text{if } \|y\| \text{ is sufficiently small.}$$

We are going to try to analyze this situation just as we did for the case $n = 2$, hoping that the quadratic terms in the Taylor expansion are determinative. In order to begin our deliberations, we single out the essence of the quadratic terms by giving the following terminology.

DEFINITION. The *Hessian matrix* of f at its critical point x_0 is the $n \times n$ matrix

$$H = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} (x_0) \right).$$

In case $n = 2$ and in the notation of the preceding sections,

$$H = \begin{pmatrix} A & B \\ B & C \end{pmatrix}.$$

In case $n = 1$ we have simply

$$H = f''(x_0).$$

Notice that in all cases the fact that f is of class C^2 implies that $D_i D_j f(x_0) = D_j D_i f(x_0)$, in other words that H is a *symmetric* matrix.

Here is a crucial concept:

DEFINITION. In the above situation the critical point x_0 is *nondegenerate* if $\det H \neq 0$. And it is *degenerate* if $\det H = 0$.

Our goal is to establish a “**second derivative test**” to determine the local nature of a critical point. This will prove to be possible only if the critical point is nondegenerate. For $n = 1$ this means $f''(x_0) \neq 0$, and standard single-variable calculus gives a strict local minimum if $f''(x_0) > 0$ and a strict local maximum if $f''(x_0) < 0$ (and no conclusion if $f''(x_0) = 0$). For $n = 2$ the nondegeneracy means $AC - B^2 \neq 0$, and the results are given on p. 3–20.

To repeat: if the critical point is degenerate, no test involving only the second order partial derivatives of f can detect the nature of a critical point. We shall therefore restrict our analysis to nondegenerate critical points.

REMARK. If $n = 1$ there is no such thing as a nondegenerate saddle point. Thus the structure of critical point behavior is much richer in multivariable calculus than in single-variable calculus. It's a good idea to keep in mind easy examples of nondegenerate saddle points for $n = 2$, such as the origin for functions such as:

$$f(x, y) = xy,$$

or

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

or

$$f(x, y) = x^2 + 5xy + y^2.$$

Here is a simple but very interesting example which illustrates the sort of thing that can happen at degenerate critical points.

PROBLEM 3–28. Define

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

- Show that $(0, 0)$ is the only critical point.
- Show that it is degenerate.
- Show that when f is restricted to any straight line through $(0, 0)$, the resulting function has a strict local minimum at $(0, 0)$.
- Show that $(0, 0)$ is a saddle point.

We are going to discover that quadratic polynomials are much more complicated algebraically for dimensions greater than two than for dimension two.

PROBLEM 3–29. The quadratic polynomial below corresponds to a Hessian matrix

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

Use whatever *ad hoc* method you can devise to show that

$$x^2 + 3y^2 + 2z^2 - 2xy + 2xz > 0$$

except at the origin.

REMARK. Since the terms in the above expression are purely quadratic, the homogeneity argument you may have used for Problem 3–18 shows that there exists a positive number λ such that

$$x^2 + 3y^2 + 2z^2 - 2xy + 2xz \geq \lambda(x^2 + y^2 + z^2)$$

for all $(x, y, z) \in \mathbb{R}^3$. However, the problem of finding the largest such λ is a difficult algebra problem, equivalent to finding a root of a certain *cubic* equation. We are very soon going to discuss this algebra. (The maximal λ is approximately 0.1206.)

We have just defined nondegenerate critical points by invoking the *determinant* of the associated Hessian matrix. Before we proceed further, we need to pause to define determinants and develop their properties.

F. Determinants

Every $n \times n$ matrix (every *square* matrix) has associated with it a number which is called its *determinant*. We are going to denote the determinant of A as $\det A$. This section is concerned with the definition and properties of this extremely important function.

If $n = 1$, a 1×1 matrix is just a real number, so we define $\det a = a$.

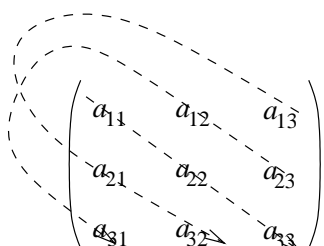
If $n = 2$, then we rely on our experience to define

$$\det A = \det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Some students at this stage also have experience with 3×3 matrices, and know that the correct definition in this case is

$$\begin{aligned} \det A &= \det \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \\ &= a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}. \end{aligned}$$

Often this definition for $n = 3$ is presented in the following scheme for forming the six possible products of the a_{ij} 's, one from each row and column:



the three indicated receive a plus sign, the three formed with the opposite slant a minus sign

No such procedure is available for the correct definition for $n > 3$. However, meditation on the formulas $n = 2$ and $n = 3$ reveals certain basic properties, which we can turn into axioms for the general case. These properties are as follows:

MULTILINEAR: $\det A$ is a linear function of each column of A ;

ALTERNATING: $\det A$ changes sign if two adjacent columns are interchanged;

NORMALIZATION: $\det I = 1$.

(In the first two of these properties we could replace “column” by “row,” but we choose not to do that now. More on this later.) Incidentally, a function satisfying the first condition is said to be *multilinear*, not linear. A *linear* function would satisfy $\det(A + B) = \det A + \det B$, definitely not true. In order to work effectively with these axioms it is convenient to rewrite a matrix A in a notation which displays its columns. Namely, we denote the columns of A as A_1, A_2, \dots, A_n , respectively. In other words,

$$A_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}.$$

Then we present A in the form

$$A = (A_1 \ A_2 \ \dots \ A_n).$$

Here is the major theorem of this section.

THEOREM. *Let $n \geq 2$. Then there exists one and only one real-valued function \det defined on all $n \times n$ matrices which satisfies*

MULTILINEAR: $\det(A_1 \ A_2 \ \dots \ A_n)$ is a linear function of each A_j ;

ALTERNATING: $\det(A_1 \ \dots \ A_j A_{j+1} \ \dots \ A_n) = -\det(A_1 \ \dots \ A_{j+1} A_j \ \dots \ A_n)$
for each $1 \leq j \leq n - 1$;

NORMALIZATION: $\det I = 1$.

Rather than present a proof in our usual style, we are going to present a discussion of the ramifications of these axioms. Not only will we prove the theorem, we shall also in the process present useful techniques for evaluation of the determinant. Here's an example.

PROBLEM 3–30. Prove that if \det satisfies the alternating condition of the theorem, then interchanging *any* two columns (not just adjacent ones) of A produces a matrix whose determinant equals $-\det A$.

PROBLEM 3–31. Prove that if two columns of A are equal, then $\det A = 0$.

PROBLEM 3–32. Conversely, prove that if \det is known to be multilinear and satisfies $\det A = 0$ whenever two columns of A are equal, then A is alternating.

(HINT: expand

$$\det(A_1 \cdots (A_j + A_{j+1})(A_j + A_{j+1})A_{j+2} \cdots A_n)$$

as a sum of four terms.)

Now we can almost immediately prove the uniqueness. It just takes a little patience with the notation. Rewrite the column A_j in terms of the unit coordinate (column) vectors as

$$A_j = \sum_{i=1}^n a_{ij} \hat{e}_i.$$

For each column we require a distinct summation index, so change i to i_j to get

$$A_j = \sum_{i_j=1}^n a_{i_j j} \hat{e}_{i_j}.$$

Now the linearity of \det as a function of each column produces a huge multi-indexed sum:

$$\begin{aligned} \det A &= \det(A_1 \ A_2 \ \dots \ A_n) \\ &= \sum_{i_1, \dots, i_n} a_{i_1 1} \dots a_{i_n n} \det(\hat{e}_{i_1} \dots \hat{e}_{i_n}). \end{aligned}$$

In this huge sum each i_j runs from 1 to n , giving us n^n terms. Now look at each individual term in this sum. If the indices i_1, \dots, i_n are not all distinct, then at least two columns of the matrix

$$(\hat{e}_{i_1} \dots \hat{e}_{i_n})$$

are equal, so that Problem 3–31 implies the determinant is 0. Thus we may restrict the sum to *distinct* i_1, \dots, i_n . Thus, the indices i_1, \dots, i_n are just the integers $1, \dots, n$ written in some other order. Thus the matrix

$$(\hat{e}_{i_1} \dots \hat{e}_{i_n})$$

can be transformed to the matrix

$$(\hat{e}_1 \dots \hat{e}_n) = I$$

by the procedure of repeatedly interchanging columns. Thus Problem 3–30 shows that

$$\det(\hat{e}_{i_1} \dots \hat{e}_{i_n}) = \pm 1,$$

and the *sign* does not depend on \det at all, but only on the interchanges needed. Thus

$$(*) \quad \det A = \sum_{i_1, \dots, i_n} \pm a_{i_1 1} \dots a_{i_n n},$$

and the signs depend on how i_1, \dots, i_n arise from $1, \dots, n$. This finishes the proof of *uniqueness* of \det !

Now we turn to the harder proof of existence. A very logical thing to do would be to use the formula we have just obtained to *define* $\det A$; then we would “just” have to verify the three required properties. However, it turns out that there is an alternate procedure that gives a different formula and one that is very useful for computations. We now describe this formula.

Temporarily denote by A'_j the column A_j with first entry replaced by 0:

$$A_j = \begin{pmatrix} a_{1j} \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix} = a_{1j} \hat{e}_1 + A'_j.$$

The multilinearity of \det , supposing it exists, gives

$$\begin{aligned} \det A &= \det(a_{11} \hat{e}_1 + A'_1 \dots a_{1n} \hat{e}_1 + A'_n) \\ &= \text{a sum of } 2^n \text{ terms,} \end{aligned}$$

where each term is a determinant arising from $\det A$ by replacing various columns of A by $a_{1j} \hat{e}_1$'s. But if as many as two columns are so replaced the resulting determinant is 0, thanks to its multilinearity and Problem 3–31. Thus (if it exists)

$$\det A = \sum_{j=1}^n a_{1j} \det(A'_1 \dots \underset{\substack{\uparrow \\ \text{column } j}}{\hat{e}_1} \dots A'_n) + \det(A'_1 \dots A'_n).$$

The last of these determinants is 0, as the first row of the matrix consists of 0's and (*) gives 0. By interchanging the required $j - 1$ columns in the other matrices, we find

$$\det A = \sum_{j=1}^n (-1)^{j-1} a_{1j} \det(\hat{e}_1 A'_1 \dots A'_{j-1} A'_{j+1} \dots A'_n).$$

The matrix that appears in the j^{th} summand is

$$\begin{pmatrix} 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & a_{21} & & a_{2,j-1} & a_{2,j+1} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0 & a_{n1} & & a_{n,j-1} & a_{n,j+1} & \cdots & a_{nn} \end{pmatrix}.$$

It's not hard to see that its determinant must be the same as that of the $(n-1) \times (n-1)$ matrix that lies in rows and columns 2 through n . However, we don't need to prove that at this stage, as we can just use the formula to give the definition of \det by induction on n .

Before doing that it is convenient to introduce a bit of notation. Given an $n \times n$ matrix A and two indices i and j , we obtain a new matrix from A by deleting its i^{th} row and j^{th} column. We denote this matrix by A_{ij} , and call it the *minor* for the i^{th} row and j^{th} column. It is of course an $(n-1) \times (n-1)$ matrix. Now here is the crucial result.

THEOREM. *The uniqueness proof of the preceding theorem guarantees that only one determinant function for $n \times n$ matrices can exist. It does exist, and can be defined by induction on n from the starting case $n = 1$ ($\det a = a$) and the inductive formula for $n \geq 2$,*

$$\det A = \sum_{j=1}^n (-1)^{i+j} a_{ij} \det A_{ij}.$$

In this formula the row index i is fixed, and can be any of the integers $1, \dots, n$ (the sum is independent of i).

The important formula in the theorem has a name: it is called **expansion of $\det A$ by the minors along the i^{th} row**.

PROOF. What we must do is for any $1 \leq i \leq n$ prove that $\det A$ as defined above satisfies the three properties listed in the preceding theorem. In doing this we are of course allowed to use the same properties for the determinants of $(n-1) \times (n-1)$ matrices. We consign the proofs of the multilinearity and normalization to the problems following this discussion, as they are relatively straightforward, and content ourselves with the alternating property about switching two adjacent columns. Let us then assume that B is the matrix which comes from A by switching columns k and $k+1$. Then if $j \neq k$ or $k+1$ the minors B_{ij} and A_{ij} also differ from one another by a column switch, so that the inductive hypothesis implies $\det B_{ij} = -\det A_{ij}$.

Thus only $j = k$ and $k + 1$ matter:

$$\begin{aligned} \det B + \det A &= (-1)^{i+k} b_{ik} \det B_{ik} + (-1)^{i+k+1} b_{i,k+1} \det B_{i,k+1} \\ &\quad + (-1)^{i+k} a_{ik} \det A_{ik} + (-1)^{i+k+1} a_{i,k+1} \det A_{i,k+1} \\ &= (-1)^{i+k} a_{i,k+1} [\det B_{ik} - \det A_{i,k+1}] \\ &\quad + (-1)^{i+k} a_{ik} [-\det B_{i,k+1} + \det A_{ik}]. \end{aligned}$$

Aha! We have $B_{ik} = A_{i,k+1}$ and $B_{i,k+1} = A_{ik}$. Thus $\det B + \det A = 0$.

QED

PROBLEM 3–33. Prove by induction that $\det A$ as defined inductively is a linear function of each column of A .

PROBLEM 3–34. Prove by induction that $\det A$ as defined inductively satisfies $\det I = 1$.

PROBLEM 3–35. Let A be “upper triangular” in the sense that $a_{ij} = 0$ for all $i > j$. Prove that $\det A = a_{11}a_{22} \dots a_{nn}$. That is,

$$\det \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & & a_{nn} \end{pmatrix} = a_{11}a_{22} \dots a_{nn}.$$

Do the same for lower triangular.

One more property is needed in order to be able to compute determinants efficiently. This property states that all the properties involving column vectors and expansion along a row are valid also for the row vectors of the matrix and expansion along columns. Because of the uniqueness of the determinant, it will suffice to prove the following result.

THEOREM. *The determinant as defined above satisfies*

ROW MULTILINEAR: $\det A$ is a linear function of each row of A .

ROW ALTERNATING: $\det A$ changes sign if two rows of A are interchanged.

PROOF. The row multilinearity is an immediate consequence of the formula (*) or of the formula for expansion of $\det A$ by the minors along the i^{th} row. The latter formula actually displays $\det A$ as a linear function of the row vector $(a_{i1}, a_{i2}, \dots, a_{in})$.

The row alternating property follows easily from the result of Problem 3–32, applied to rows instead of columns. Thus, assume that two rows of A are equal. If $n = 2$, $\det A = 0$ because of our explicit formula in this case. If $n > 2$, then expand $\det A$ by minors along a *different* row. The corresponding determinants of the minors A_{ij} are all zero (by induction on n) because each A_{ij} has two rows equal.

QED

The proofs of the following two corollaries are now immediate:

COROLLARY. $\det A^t = \det A$.

Here A^t is the transpose of A , as defined in Problem 2–83.

COROLLARY. *Expansion by minors along the j^{th} column:*

$$\det A = \sum_{i=1}^n (-1)^{i+j} a_{ij} \det A_{ij}.$$

If we had to work directly from the definition of determinants, calculating them would be utterly tedious for $n \geq 4$. We would essentially need to add together $n!$ terms, each a product of n matrix entries. Fortunately, the properties lend themselves to very efficient calculations. The goal is frequently to convert A to a matrix for which a row or column is mostly 0.

The major tool for doing this is the property that **if a scalar multiple of a column (or row) is added to a different column (or row), then the determinant is unchanged.** Such procedures are called *elementary column (or row) operations*.

For instance,

$$\begin{aligned}
 & \det \begin{pmatrix} 3 & 1 & 0 & 6 \\ 1 & -1 & -1 & 5 \\ 0 & 2 & 2 & 1 \\ 2 & 0 & -1 & 3 \end{pmatrix} \\
 &= \det \begin{pmatrix} 0 & 4 & 3 & -9 \\ 1 & -1 & -1 & 5 \\ 0 & 2 & 2 & 1 \\ 0 & 2 & 1 & -7 \end{pmatrix} && \text{(added multiples of row 2 to rows 1 and 4)} \\
 &= -\det \begin{pmatrix} 4 & 3 & -9 \\ 2 & 2 & 1 \\ 2 & 1 & -7 \end{pmatrix} && \text{(expanded by minors along column 1)} \\
 &= -\det \begin{pmatrix} 22 & 21 & -9 \\ 0 & 0 & 1 \\ 16 & 15 & -7 \end{pmatrix} && \text{(added multiples of column 3 to columns 1 and 2)} \\
 &= \det \begin{pmatrix} 22 & 21 \\ 16 & 15 \end{pmatrix} && \text{(expanded by minors along row 2)} \\
 &= 2 \cdot 3 \det \begin{pmatrix} 11 & 7 \\ 8 & 5 \end{pmatrix} && \text{(used linearity in columns 1 and 2)} \\
 &= 6(55 - 56) \\
 &= -6.
 \end{aligned}$$

PROBLEM 3–36. Find and prove an expression for the $n \times n$ determinant

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

PROBLEM 3–37. Given fixed numbers a, b, c , let x_n stand for the $n \times n$ “tridiagonal” determinant

$$x_n = \det \begin{pmatrix} a & b & & & \\ c & a & b & & \\ & c & a & b & \\ & & \cdot & \cdot & \cdot \\ & & & c & a \end{pmatrix}.$$

Find a formula for x_n in terms of x_{n-1} and x_{n-2} . (Blank spaces are all 0.)

PROBLEM 3–38. Express the $n \times n$ tridiagonal determinant

$$\det \begin{pmatrix} 1 & 1 & & & \\ -1 & 1 & 1 & & \\ & -1 & 1 & 1 & \\ & & \cdot & \cdot & \cdot \\ & & & -1 & 1 \end{pmatrix}$$

in Fibonacci terms. The Fibonacci numbers may be defined by $F_0 = 0$, $F_1 = 1$, $F_n = F_{n-1} + F_{n-2}$.

PROBLEM 3–39. Find and prove an expression for the $n \times n$ determinant

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

PROBLEM 3–40. Here is a famous determinant, called the *Vandermonde* determinant: prove that

$$\det \begin{pmatrix} 1 & a_1 & a_1^2 & \cdots & a_1^{n-1} \\ 1 & a_2 & a_2^2 & \cdots & a_2^{n-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & a_n & a_n^2 & \cdots & a_n^{n-1} \end{pmatrix} = \prod_{i < j} (a_j - a_i).$$

PROBLEM 3–41. Calculate

$$\det \begin{pmatrix} \lambda + a_1 b_1 & a_1 b_2 & a_1 b_3 & \cdots & a_1 b_n \\ a_2 b_1 & \lambda + a_2 b_2 & a_2 b_3 & \cdots & a_2 b_n \\ \vdots & \vdots & \vdots & & \vdots \\ a_n b_1 & a_n b_2 & a_n b_3 & \cdots & \lambda + a_n b_n \end{pmatrix}.$$

(Answer: $\lambda^{n-1}(\lambda + a \bullet b)$)

PROBLEM 3–42. When applied to $n \times n$ matrices, *determinant* can be regarded as a function from \mathbb{R}^{n^2} to \mathbb{R} . What is

$$\frac{\partial \det A}{\partial a_{ij}} \quad ?$$

G. Invertible matrices and Cramer's rule

Before entering into the main topic of this section, we give one more property of determinant.

THEOREM. $\det(AB) = \det A \det B$.

PROOF. Here of course A and B are square matrices of the same size, say $n \times n$. The $i - j$ entry of the product AB is

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

Thus the j^{th} column of AB is the column vector

$$\sum_{k=1}^n \begin{pmatrix} a_{1k}b_{kj} \\ a_{2k}b_{kj} \\ \vdots \\ a_{nk}b_{kj} \end{pmatrix} = \sum_{k=1}^n b_{kj}A_k,$$

where, as usual, A_k is our notation for the k^{th} column of A . We can now simply follow the procedure of p. 3–28. In the above formula rename the summation index $k = i_j$, so we have by the multilinearity of \det ,

$$\det AB = \sum_{i_1, \dots, i_n} b_{i_1 1} \dots b_{i_n n} \det(A_{i_1} \dots A_{i_n}).$$

Here the i_j 's run from 1 to n . If they are not distinct, then the corresponding determinant is 0. Thus the i_j 's are distinct in all nonzero terms in the sum, and by column interchanges we have

$$\begin{aligned} \det(A_{i_1} \dots A_{i_n}) &= \pm \det(A_1 \dots A_n) \\ &= \pm \det A. \end{aligned}$$

In fact, the sign is the same as

$$\det(\hat{e}_{i_1} \dots \hat{e}_{i_n}).$$

Therefore

$$\begin{aligned} \det AB &= \det A \sum_{i_1, \dots, i_n} b_{i_1 1} \dots b_{i_n n} \det(\hat{e}_{i_1} \dots \hat{e}_{i_n}) \\ &= \det A \det B. \end{aligned}$$

QED

REMARKS. This result is a rather astounding bit of multilinear algebra. To convince yourself of that, just write down what it says in the case $n = 2$. Amazingly, the identical technique leads to a vast generalization, the Cauchy-Binet determinant theorem, which we shall present in Section 11D. It is also amazing that there is an entirely different approach to proving this result, as we shall discover in our study of integration in Chapter 10.

DEFINITION. A square matrix A is *invertible* if there exists a matrix B such that $AB = BA = I$. The matrix B is called the *inverse* of A , and is written $B = A^{-1}$.

PROBLEM 3–43. For the above definition of inverse to make sense we need to know that the matrix B is unique. In fact, prove more: if A has a right inverse B ($AB = I$) and a left inverse C ($CA = I$), then $B = C$. (HINT: CAB .)

PROBLEM 3–44. Prove that if A and B are invertible, then AB is invertible and

$$(AB)^{-1} = B^{-1}A^{-1}.$$

PROBLEM 3–45. Prove that in general

$$(AB)^t = B^t A^t.$$

PROBLEM 3–46. Prove that A is invertible $\iff A^t$ is invertible. Also prove that

$$(A^{-1})^t = (A^t)^{-1}.$$

Now we are almost ready to state the major result of this section. There's one more concept we need, and that is the linear independence of points in \mathbb{R}^n .

DEFINITION. The vectors $x^{(1)}, x^{(2)}, \dots, x^{(k)}$ in \mathbb{R}^n are *linearly dependent* if there exist real numbers c_1, c_2, \dots, c_k , not all 0, such that

$$c_1 x^{(1)} + c_2 x^{(2)} + \dots + c_k x^{(k)} = 0.$$

In other words, the given vectors satisfy some nontrivial homogeneous linear relationship. The vectors are said to be *linearly independent* if they are not linearly dependent.

PROBLEM 3–47. Prove that $x^{(1)}, \dots, x^{(k)}$ are linearly dependent \iff one of them is equal to a linear combination of the others.

A good portion of elementary linear algebra is dedicated to proving that if $x^{(1)}, \dots, x^{(n)}$ are linearly independent vectors in \mathbb{R}^n — notice that it's the same n — then every vector in \mathbb{R}^n can be expressed as a unique linear combination of the given vectors. That is, for any $x \in \mathbb{R}^n$ there exist unique scalars c_1, \dots, c_n , such that

$$x = c_1x^{(1)} + \cdots + c_nx^{(n)}.$$

PROBLEM 3–48. Prove that the scalars c_1, \dots, c_n in the above equation are indeed unique.

PROBLEM 3–49. Prove that any $n + 1$ vectors in \mathbb{R}^n are linearly dependent.

PROBLEM 3–50. Prove that the column vectors of an upper triangular matrix (a_{ij}) are linearly independent \iff the diagonal entries a_{ii} are all nonzero.

DEFINITION. If $x^{(1)}, \dots, x^{(n)}$ are linearly independent vectors in \mathbb{R}^n , they are said to be a *basis* for \mathbb{R}^n .

Now for the result.

THEOREM. Let A be an $n \times n$ matrix. Then the following conditions are equivalent:

- (1) $\det A \neq 0$.
- (2) A is invertible.
- (3) A has a right inverse.
- (4) A has a left inverse.
- (5) The columns of A are linearly independent.
- (6) The rows of A are linearly independent.

PROOF. We first prove that $(3) \implies (1) \implies (5) \implies (3)$. First, $(3) \implies (1)$ because the matrix equation $AB = I$ implies $\det A \det B = \det I = 1$, so $\det A \neq 0$. To see that $(1) \implies (5)$, suppose the columns of A linearly dependent. Then one of them is a linear combination of the

others, as we know from Problem 3–47. Inserting the resulting equation into the corresponding column of A and using the linearity of $\det A$ with respect to that column leads to $\det A = 0$.

Now suppose (5) holds. Then every vector can be expressed as a linear combination of the columns A_1, \dots, A_n of A . In particular, each coordinate vector can be so expressed:

$$\hat{e}_j = \sum_{k=1}^n b_{kj} A_k$$

for certain unique scalars b_{1j}, \dots, b_{nj} . Now we define the matrix $B = (b_{ij})$, and we realize the equations for \hat{e}_j simply assert that

$$I = AB.$$

Thus (3) is valid.

So (1), (3), (5) are equivalent. Likewise, (1), (4), (6) are equivalent. We conclude that (1), (3), (4), (5), (6) are equivalent. Finally, if they all hold, then (3) and (4) imply A is invertible, thanks to Problem 3–43. The converse assertions that (2) \implies (3) and (4) are trivial.

QED

We also remark that in case A is invertible, the two matrices A and A^{-1} commute. Though this is rather obvious, still it is quite a wonderful fact since we do not generally expect two matrices to commute. Whenever it happens that $AB = BA$, that is really worth our attention.

This is certainly a terrific theorem! It displays in dramatic fashion the importance of the determinant. However, it does not show how we might go about computing A^{-1} . There is a theorem that does this, and it goes by the name of Cramer's rule.

Assume that $\det A \neq 0$ and that $B = A^{-1}$. Then we examine the equation $AB = I$. As we have seen, it can be written as the set of vector equations

$$\sum_{k=1}^n b_{kj} A_k = \hat{e}_j, \quad 1 \leq j \leq n.$$

Now replace the i^{th} column of A by \hat{e}_j and calculate the determinant of the resulting matrix. First, expansion by minors along the i^{th} column produces

$$\det(A_1 \dots A_{i-1} \hat{e}_j A_{i+1} \dots A_n) = (-1)^{i+j} \det A_{ji}.$$

On the other hand, the formula for \hat{e}_j yields

$$\begin{aligned} \det(A_1 \dots A_{i-1} \hat{e}_j A_{i+1} \dots A_n) &= \sum_{k=1}^n b_{kj} \det(A_1 \dots A_{i-1} A_k A_{i+1} \dots A_n) \\ &= b_{ij} \det(A_1 \dots A_{i-1} A_i A_{i+1} \dots A_n) \\ &= b_{ij} \det A. \end{aligned}$$

We conclude that

$$b_{ij} = (-1)^{i+j} \frac{\det A_{ji}}{\det A}.$$

Now that we have proved it, we record the statement:

CRAMER'S RULE. Assume $\det A \neq 0$. Then A^{-1} is given by the formula for its $i - j$ entry,

$$(-1)^{i+j} \frac{\det A_{ji}}{\det A}.$$

Elegant as it is, this formula is rarely used for calculating inverses, on account of the effort required to evaluate determinants. However, the 2×2 case is especially appealing and easy to remember:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

COROLLARY. Assume $\det A \neq 0$ and $b \in \mathbb{R}^n$ is a column vector. Then the unique column vector $x \in \mathbb{R}^n$ which satisfies $Ax = b$ is given by $x = A^{-1}b$, and the i^{th} entry of x equals

$$x_i = \frac{\det(A_1 \dots A_{i-1} b A_{i+1} \dots A_n)}{\det A}.$$

PROBLEM 3–51. Prove the corollary.

The equation $Ax = b$, with x as the “unknown,” is of course the general system of n linear equations in n unknowns, for if we display the coordinates we have

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1, \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n = b_n. \end{cases}$$

We can conclude from our results that four conditions concerning this system of equations are equivalent:

PROBLEM 3–52. Given an $n \times n$ matrix A , prove that the following four conditions are equivalent:

- (1) $\det A \neq 0$.
- (2) If $x \in \mathbb{R}^n$ is a column vector such that $Ax = 0$, then $x = 0$.
- (3) For every $b \in \mathbb{R}^n$, there exists $x \in \mathbb{R}^n$ such that $Ax = b$.
- (4) For every $b \in \mathbb{R}^n$, there exists one and only one $x \in \mathbb{R}^n$ such that $Ax = b$.

PROBLEM 3–53. The *classical adjoint* of a square matrix A is denoted $\text{adj}A$ and is defined by prescribing its $i - j$ entry to be

$$(-1)^{i+j} \det A_{ji}.$$

As we have shown if $\det A \neq 0$, then $A^{-1} = \text{adj}A / \det A$. Prove that in general

$$A \text{adj}A = (\det A)I.$$

(HINT: write down the definition of the $i - j$ entry of the product on the left side and use expansion by minors along the j^{th} row.

PROBLEM 3–54. Prove that A and $\text{adj}A$ commute.

PROBLEM 3–55. Prove that

$$\text{adj}(\text{adj}A) = (\det A)^{n-2}A.$$

PROBLEM 3–56. Prove that the classical adjoint of an upper triangular matrix is also upper triangular.

PROBLEM 3–57. Write out explicitly

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{pmatrix}^{-1}.$$

H. Recapitulation

As we have become involved in a great deal of mathematics in our analysis of critical points, it seems good to me that we pause for a look at the forest.

1. We are considering a function $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ of class C^2 .
2. We assume x_0 is a critical point for $f : \nabla f(x_0) = 0$.
3. We consider the Taylor expansion of $f(x_0 + y)$:

$$f(x_0 + y) = f(x_0) + \frac{1}{2} \sum_{i,j=1}^n D_i D_j f(x_0) y_i y_j + R.$$

4. We define the Hessian matrix

$$H = (D_i D_j f(x_0)).$$

5. We say the critical point is nondegenerate if $\det H \neq 0$.
6. Our goal is now to show that in the nondegenerate case, the Hessian matrix reveals the local nature of the critical point.
7. We have finished this program in dimension 2, where we learned that if the critical point is degenerate ($\det H = 0$), we cannot detect the local behavior through H .

What remains is the analysis of symmetric matrices like H and the associated behavior of the quadratic functions they generate. As this in itself is a fascinating and important topic, we devote the next chapter to this study.

PROBLEM 3–58. Although this problem is concerned with \mathbb{R}^n , you can analyze it directly without requiring the result of Chapter 4. This is a generalization of Problem 3–24 (see also Problem 2–60). Let $a_1 > a_2 > \cdots > a_n > 0$ be constants, and define $\mathbb{R}^n \xrightarrow{f} \mathbb{R}$ by

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

- Find all the critical points of f and determine the local nature of each one.
- Clearly, f attains a strict global minimum at $x = 0$. Show also that f attains a global maximum; is it strict?

I. A little matrix calculus

It is quite interesting to investigate the interplay between the algebra of matrices and calculus. Specifically we wish to analyze the operation of the inverse of a matrix.

First we observe that the space of all $n \times n$ real matrices may be viewed as \mathbb{R}^{n^2} , a fact we observed on p. 2–48. Since the determinant of A is a polynomial expression in the entries of A we conclude that $\det A$ is a C^∞ function on \mathbb{R}^{n^2} . The set of invertible matrices is the same as the set of matrices with nonzero determinant and is therefore an *open* subset of the set of all $n \times n$ matrices.

DEFINITION. The *general linear group* of $n \times n$ matrices is the set of all $n \times n$ real invertible matrices. It is commonly denoted

$$\mathrm{GL}(n).$$

This is to be regarded as an open subset of the space of all $n \times n$ real matrices.

There is of course a very interesting function defined on $\mathrm{GL}(n)$, namely the function which maps a matrix A to its inverse A^{-1} . Let's name this function *inv*. Thus

$$\mathrm{inv}(A) = A^{-1}.$$

Cramer's rule displays A^{-1} as a matrix whose entries are polynomial functions of A , divided by $\det A$. Thus A^{-1} is a C^∞ function of A . That is,

$$\mathrm{GL}(n) \xrightarrow{\mathrm{inv}} \mathrm{GL}(n)$$

is a C^∞ function.

PROBLEM 3–59. If the $n \times n$ matrix B is sufficiently small, $I + B \in \text{GL}(n)$. Prove that

$$(I + B)^{-1} = I - (I + B)^{-1}B.$$

Then prove that

$$(I + B)^{-1} = I - B + (I + B)^{-1}B^2.$$

PROBLEM 3–60. Use the preceding problem to show that the differential of inv at I is the linear mapping which sends B to $-B$. In terms of directional derivatives,

$$D\text{inv}(I; B) = -B.$$

PROBLEM 3–61. Let $A \in \text{GL}(n)$. Use the relation

$$A + B = A(I + A^{-1}B)$$

to prove that

$$D\text{inv}(A; B) = -A^{-1}BA^{-1}.$$

In other words,

$$\left. \frac{d}{dt}(A + tB)^{-1} \right|_{t=0} = -A^{-1}BA^{-1}.$$