Chapter 5
Inverses

Abstract Inverses help us solve equations: if $5 = x^3$, then $x = \sqrt[3]{5}$. Equations also imply relations between their variables. For example, if $x^2 + y^2 - 1 = 0$, then we can "solve for y" to get either $y = +\sqrt{1-x^2}$ or $y = -\sqrt{1-x^2}$. We soon learn that a formula for an inverse or for an implicitly defined function is seldom available. Usually, the most we can expect to know is that such a function exists. As we show, even this apparently limited knowledge can simplify and clarify our view of a problem, the same way that changing coordinates can simplify an integration. In this chapter, we look only briefly at explicit formulas. We give the bulk of our attention to the way inverses give us a powerful tool for understanding maps, and to the conditions that guarantee their existence. The next chapter does the same for implicitly defined functions.

5.1 Solving equations

Early examples

The first inverse operations we learn are subtraction and division; after all, $x = y/m$ is the inverse of $y = mx$. And division is the first place where we see that an inverse may not exist: "You cannot divide by zero" is the way we say that $y = 0 \times x$ has no inverse. We use subtraction and division to solve equations, at the start, just linear equations of the form $y = mx + b$. After this come polynomial equations and the square root $x = \sqrt{y}$, introduced as the inverse of $y = x^2$. The square root function shows us that an inverse may have a restricted domain of definition ($y \geq 0$ in this case) and a restricted range (we need $x = -\sqrt{y}$ along with $x = +\sqrt{y}$).

Inverse of the hyperbolic cosine

For each new function in calculus, an inverse is introduced with it; the exponential and logarithm functions provide a good example. The immediate use of inverses is in solving equations, including even those that give alternate formulas for inverses themselves. For example, the hyperbolic cosine function $y = \cosh x$ has an inverse that is written simply $x = \operatorname{arccosh} y$ (or $x = \cosh^{-1} y$). But we can get a different—and possibly more useful—expression for the inverse by solving the defining equation

J.J. Callahan, *Advanced Calculus: A Geometric View*, Undergraduate Texts in Mathematics,
DOI 10.1007/978-1-4419-7332-0_5,

$$y = \cosh x = \frac{e^x + e^{-x}}{2}$$

for x algebraically. Some simple computations give

$$2ye^x = e^{2x} + 1 \text{ and then } e^{2x} - 2ye^x + 1 = 0.$$

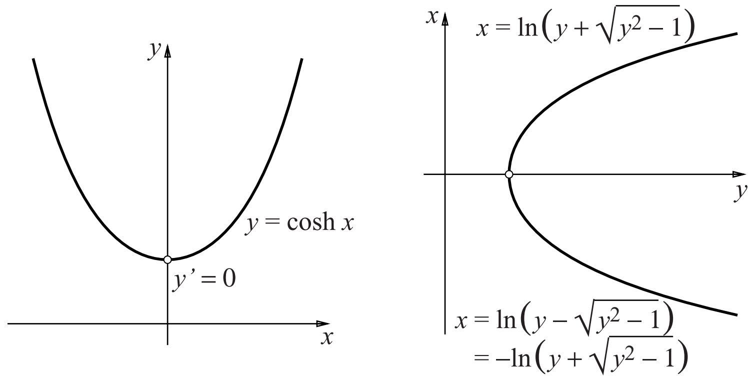
Notice that this is an ordinary quadratic equation in e^x; the quadratic formula (an inverse!) gives

$$e^x = \frac{2y \pm \sqrt{4y^2 - 4}}{2} = y \pm \sqrt{y^2 - 1}.$$

Branches of the inverse

We can finally solve for x itself by using the logarithm (yet another inverse):

$$x = \ln\left(y \pm \sqrt{y^2 - 1}\right) = \operatorname{arccosh} y.$$



The "$\pm$" in the formula for x means that the inverse splits into two parts, or **branches**, with a separate formula for each. The graph of $x = \operatorname{arccosh} y$ on the right, above, helps us see why. It is the reflection of the graph of $y = \cosh x$ across the line $y = x$. It splits into two halves at the point ($x = 0$) where $y' = 0$:

$$\begin{aligned} \text{upper, } x \geq 0: &\quad x = \ln\left(y + \sqrt{y^2 - 1}\right), \\ \text{lower, } x \leq 0: &\quad x = \ln\left(y - \sqrt{y^2 - 1}\right). \end{aligned}$$

The two branches imply that we should think of the inverse as a 1–2 map: for each $y > 1$, the inverse gives two x-values.

There is more to say here: the graphs of those two branches are symmetric across the y-axis, implying that the two corresponding x-values must be negatives of each other. In other words, the equation of the lower half should be

$$x = -\ln\left(y + \sqrt{y^2 - 1}\right).$$

There is no conflict, however. Note that

$$\left(y - \sqrt{y^2 - 1}\right)\left(y + \sqrt{y^2 - 1}\right) = y^2 - (y^2 - 1) = 1,$$

so

$$\ln\left(y-\sqrt{y^2-1}\right)=\ln\left(\frac{1}{y+\sqrt{y^2-1}}\right)=-\ln\left(y+\sqrt{y^2-1}\right).$$

Finally, notice that the term y^2-1 under the radical implies the inverse is defined only for $y \geq 1$, a fact borne out by the graph.

Inverses of other hyperbolic functions

In a similar way, you can show that $\operatorname{arcsinh} y=\ln\left(y+\sqrt{y^2+1}\right)$ (there is no "±" ambiguity here) and use this formula with the pullback substitution $y=\sinh x$ to show that

$$\int \frac{dy}{\sqrt{1+y^2}}=\ln\left(y+\sqrt{y^2+1}\right).$$

See the exercises for this and other questions involving the hyperbolic functions and their inverses.

Inverses without formulas

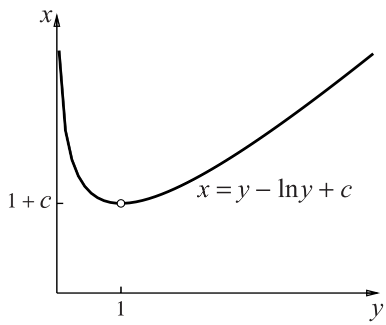
Inverses play a crucial role in solving problems even when there is no formula or explicit expression for the inverse in terms of elementary functions. For example, consider the differential equation

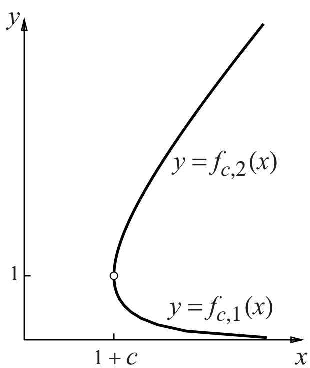
$$\frac{dy}{dx}=\frac{y}{y-1}.$$

This equation, as written, indicates that y changes with x, so x is the independent variable. Thus, we are looking for a function $y=f(x)$ for which the equation

$$f'(x)=\frac{f(x)}{f(x)-1}$$

is an identity in x, at least for all x in some interval.





Separating variables

A solution is shown above, at the right. We can obtain this solution and others by using the method of separation of variables. The method begins by rewriting the original differential equation as

$$dx=\frac{y-1}{y}\,dy=\left(1-\frac{1}{y}\right)dy$$

(a *differential equation* was originally "an equation involving differentials"). Integrating this, we get an expression involving an arbitrary constant c,

$$x = y - \ln(y) + c, \quad y > 0,$$

whose graph is shown on the left, above. Of course, the function $y = f(x)$ we seek is the inverse, whose graph is shown on the right.

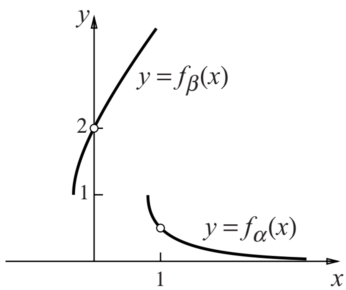
Branches of the solution

We put a hole in the graph of $x = y - \ln y + c$ at $(x,y) = (1+c,1)$ because the original differential equation is undefined when $y = 1$. The inverse hence has two branches; we call them $y = f_{c,1}(x)$ and $y = f_{c,2}(x)$. The branches have different ranges, $0 < f_{c,1}(x) < 1$, $1 < f_{c,2}(x)$, but the same domain, $x > 1 + c$. We can describe the two functions by their graphs or by the words "the two branches of the inverse of $x = y - \ln y + c$." They have no formulas.

Initial-value problems

Separate branches here are welcomed, because they provide the flexibility needed to solve different initial-value problems. For example, sketched below are the two particular solutions f_α and f_β to the differential equation that satisfy the different initial conditions

$$f_\alpha(1) = \tfrac{1}{2}, \qquad f_\beta(0) = 2.$$



These examples raise an obvious question: what is the solution if the initial value is not positive? We can extend our formula for x to $y < 0$ by

$$x = y - \ln(-y) + k,$$

where k is a constant unconnected to c. The inverse $y = f_{k,3}(x)$ here is the branch we need; its range is $y < 0$. See the exercises.

Solving two equations in two unknowns

Our rather ad hoc way of solving equations can, with some luck, be carried over to several functions of several variables, for example, to produce formulas for the inverse of a map. Consider the quadratic map

$$\mathbf{f} : \begin{cases} x = u^2 - v^2, \\ y = 2uv, \end{cases}$$

from Chapter 4. The inverse of $\mathbf{f}$ expresses u and v in terms of x and y. We do this—that is, we solve for u and v—by isolating each of these variables in its own separate equation. The key is to notice that

$$x^2+y^2=u^4-2u^2v^2+v^4+4u^2v^2=\left(u^2+v^2\right)^2.$$

We are then able to isolate u and v by adding and then by subtracting the pair of equations

$$\begin{aligned} u^2+v^2&=\sqrt{x^2+y^2},\\ u^2-v^2&=x; \end{aligned}$$

this gives us the components of $\mathbf{f}^{-1}$:

$$u=\pm\sqrt{\frac{\sqrt{x^2+y^2}+x}{2}},\quad v=\pm\sqrt{\frac{\sqrt{x^2+y^2}-x}{2}}.$$

(The expressions for u and v are real because $\sqrt{x^2+y^2}=u^2+v^2\geq 0$.)

Choosing signs for $\mathbf{f}^{-1}$

The "$\pm$" signs put the image point in the four different quadrants of the (u,v)-plane. To decide which signs to use, recall that the original map $\mathbf{f}:(u,v)\mapsto(x,y)$ "doubled angles." In particular, it mapped the first quadrant of the (u,v)-plane to the upper half-plane $y\geq 0$ and the second quadrant to the lower half-plane $y\leq 0$. Thus $\mathbf{f}^{-1}$ maps $y\geq 0$ to the first quadrant,

$$u=+\sqrt{\frac{\sqrt{x^2+y^2}+x}{2}},\quad v=+\sqrt{\frac{\sqrt{x^2+y^2}-x}{2}},\quad \text{if } y\geq 0,$$

and $y\leq 0$ to the second quadrant,

$$u=-\sqrt{\frac{\sqrt{x^2+y^2}+x}{2}},\quad v=+\sqrt{\frac{\sqrt{x^2+y^2}-x}{2}},\quad \text{if } y\leq 0.$$

These are the formulas for $\mathbf{g}_+$ (Exercise 4.13) and $\mathbf{g}_-$ (Exercise 4.17) on pages 144ff.

Do the formulas agree on the overlap?

What happens on the overlap $y=0$? If $x<0$ and $y=0$, then

$$v=\sqrt{\frac{\sqrt{x^2}-x}{2}}=\sqrt{\frac{|x|-x}{2}}=\sqrt{\frac{-x-x}{2}}=\sqrt{-x}=\sqrt{|x|}>0,$$

and

$$u=\pm\sqrt{\frac{|x|+x}{2}}=\pm\sqrt{\frac{-x+x}{2}}=0.$$

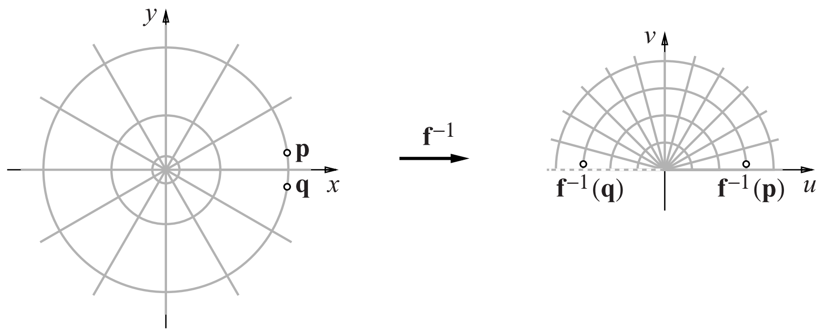
The two pairs of formulas agree: for both, the image of the negative x-axis is the positive v-axis. On the other hand, if $x>0$ and $y=0$, then $|x|=x$, so

$$v=\sqrt{\frac{|x|-x}{2}}=0,\ \text{ and }\ u=\pm\sqrt{\frac{|x|+x}{2}}=\pm\sqrt{|x|}.$$

Here there is a conflict: the first pair of formulas (where the sign of u is "+") maps the positive x-axis to the positive u-axis, but the second pair maps it to the negative u-axis.

$\mathbf{f}^{-1}$ is discontinuous

One way to eliminate the conflict is to remove the positive x-axis from the domain of the second pair of formulas. Then $\mathbf{f}^{-1}$ is well defined on the whole plane $\mathbb{R}^2$. However, there is a cost: along the positive x-axis, $\mathbf{f}^{-1}$ is *discontinuous*. For example, as the points $\mathbf{p}$ and $\mathbf{q}$, below, become arbitrarily close, their images do not.



Give $\mathbf{f}^{-1}$ a second branch

There is a radical way to solve this problem that builds on the fact $\mathbf{f}$ is a 2–1 map (because diametrically opposite points in the (u,v)-plane have the same image under $\mathbf{f}$). This suggests that $\mathbf{f}^{-1}$ is, more properly, a 1–2 map and therefore has a second branch. (Consider, for a moment, the 1-dimensional analogue of $\mathbf{f}$: $x = f(u) = u^2$; f^{-1} has the two familiar branches $u = +\sqrt{x}$ and $u = -\sqrt{x}$. Not coincidentally, the two u-values are opposites.) If $\mathbf{f}^{-1}$ already assigns to the point $\mathbf{x}$ the point $\mathbf{u}$ in the upper half-plane, then we can easily get the second branch by having $\mathbf{f}^{-1}$ assign to $\mathbf{x}$ also the diametrically opposite point $-\mathbf{u}$ in the lower half-plane:

$$\mathbf{f}^{-1}(\mathbf{x}) = \pm\mathbf{u}.$$

Here is a set of formulas that expresses both branches in terms of components. The two branches are distinguished from each other by the "±"signs:

$$(u,v) = \mathbf{f}^{-1}(x,y) = \begin{cases} \pm\left(\sqrt{\dfrac{\sqrt{x^2+y^2}+x}{2}}, \sqrt{\dfrac{\sqrt{x^2+y^2}-x}{2}}\right), & y \geq 0, \\ \pm\left(-\sqrt{\dfrac{\sqrt{x^2+y^2}+x}{2}}, \sqrt{\dfrac{\sqrt{x^2+y^2}-x}{2}}\right), & y \leq 0. \end{cases}$$

The second branch eliminates the discontinuity along the positive x-axis. For example, $\mathbf{f}^{-1}(\mathbf{q})$ in the figure above now becomes the pair of points $\pm\mathbf{f}^{-1}(\mathbf{q})$, and similarly $\mathbf{f}^{-1}(\mathbf{p})$ branches into $\pm\mathbf{f}^{-1}(\mathbf{p})$. Then, although $\pm\mathbf{f}^{-1}(\mathbf{q})$ is not close to $\pm\mathbf{f}^{-1}(\mathbf{p})$, it *is* close to $\mp\mathbf{f}^{-1}(\mathbf{p})$.

Solving equations by finding fixed points

There is another important method we can use to solve equations: find the fixed points of a suitably chosen map by iterating the map. We show immediately below how this gives us a valuable computational tool; in Chapter 5.3, we show that it provides the theoretical key to the proof of the inverse function theorem.

Definition 5.1 *Suppose $g : X \to X$ is a map of a set X to itself; then $\widehat{x}$ is a* ***fixed point*** *of g if $g(\widehat{x}) = \widehat{x}$.*

Suppose, now, we must solve the numerical equation $y = f(x)$ for x when y is given. Let $g(x) = f(x) - y + x$; then the following chain of implications shows that every fixed point of g is a solution of $f(x) = y$, and conversely:

$$g(\widehat{x}) = \widehat{x} \iff f(\widehat{x}) - y + \widehat{x} = \widehat{x} \iff f(\widehat{x}) = y.$$

The formula $g(x) = f(x) - y + x$ is just one way to construct a function whose fixed points are the solutions to $y = f(x)$; there are others. One familiar example is provided by the *Newton–Raphson method* for finding roots of $f(x) = 0$:

$$g(x) = x - \frac{f(x)}{f'(x)}.$$

Another is provided by the ancient *Babylonian algorithm* for finding square roots. We look at this in detail in order to see how well the fixed-point approach lends itself to computation. Given $a > 0$, our goal is to find $\widehat{x} > 0$ so that $\widehat{x}^2 = a$. We have

The Babylonian algorithm

$$\widehat{x} = a/\widehat{x} \text{ so } 2\widehat{x} = \widehat{x} + a/\widehat{x} \text{ and } \widehat{x} = \frac{\widehat{x} + a/\widehat{x}}{2}.$$

In other words, $\widehat{x} = \sqrt{a}$ is a fixed point of

$$g(x) = \frac{x + a/x}{2}.$$

But g is just the function; the algorithm itself tells us how to find $\widehat{x}$: pick x_0 arbitrarily (but reasonably close to $\sqrt{a}$), and then set

$$x_1 = g(x_0), \quad x_2 = g(x_1), \quad x_3 = g(x_2),$$

and so on. The sequence x_0, x_1, x_2, ... converges to the fixed point $\widehat{x} = \sqrt{a}$. An example makes it clear how rapid this convergence can be. Take $a = 6$ and let $x_0 = 2$. Then

n	x_n	x_n^2
1	2.5	6.25
2	2.45	6.0025
3	2.449 489 795 918 367	6.000 000 260 308 205
4	2.449 489 742 783 179	6.000 000 000 000 004
5	2.449 489 742 783 178	5.999 999 999 999 999

To fifteen decimal places, $\widehat{x} = 2.449\,489\,742\,783\,178$. The convergence here is especially rapid: the number of correct digits roughly doubles with each iteration.

Fixed points by iteration

The Babylonian algorithm suggests the following general procedure for finding a fixed point. Take a point x_0 and construct its *iterates* under g: $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$. If the sequence has a limit, let $\widehat{x}$ be that limit. Then

$$g(\widehat{x}) = g\left(\lim_{n\to\infty} x_n\right) = \lim_{n\to\infty} g(x_n) = \lim_{n\to\infty} x_{n+1} = \widehat{x}$$

if g is continuous. (Continuity is needed to be sure that the limit can be taken either before or after g is evaluated.) Thus $\widehat{x}$ is a fixed point of g. The Newton–Raphson method is implemented by the same kind of iteration.

Contraction mappings

Of course, in order to use this procedure, we have to make certain that the iterates have a limit, and the map g is continuous. For a *contraction mapping* (Definition 5.3, p. 167), these conditions are satisfied, and the *contraction mapping principle* (Theorem 5.1, p. 167) then guarantees the existence of a unique fixed point.

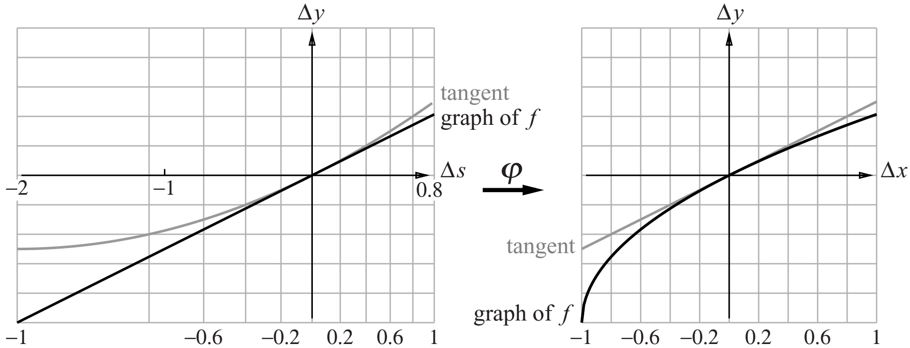
5.2 Coordinate changes

We already use coordinate changes in integration, to simplify an integrand or to convert it into a more recognizable form. In this section we put coordinate changes to larger use, to simplify the geometry of a map. For instance, we saw in Chapter 4 that a map frequently "looked like" its derivative near a point. The derivative, being linear, was essentially simple; the resemblance between the map and its derivative meant that the map itself was simple, too, at least near that point. Our goal in this section is to explain what it means for one map to *look like* a second; in fact, it means that, when the first map is expressed using appropriate new coordinates, it will be *identical to* the second one. To see how coordinate changes play this vital role, we consider several examples.

Example 1: $f(x) = \sqrt{x}$ at $x = 1$

At the point $x = 1$, the tangent to the graph of $y = f(x) = \sqrt{x}$ is a straight line of slope 1/2. Let us analyze f in a window centered at $(x,y) = (1,1)$, first using coordinates $\Delta x = x - 1$, $\Delta y = y - 1$ based at the center of the window. Then

$$\Delta y = y - 1 = \sqrt{x} - 1 = \sqrt{1+\Delta x} - 1,$$



so the formula for f in the window coordinates is $\Delta y = -1+\sqrt{1+\Delta x}$. The graph is the familar one shown in black on the right, above. With it, in gray, is the graph of the derivative, $\Delta y = \mathrm{d}f_1(\Delta x) = \frac{1}{2}\Delta x$. The black and gray graphs "share ink" near $\Delta x = 0$, ample evidence that the square root map "looks like" its derivative there. But we can do even more: with the proper coordinate change $\Delta x = \varphi(\Delta s)$, we can make the formula for f become $\Delta y = \frac{1}{2}\Delta s$. In the new $(\Delta s, \Delta y)$ window, the graph of f will be straight.

A pullback to simplify the formula for f

How can we find φ? Because our goal is to simplify the formula for f, and because that formula involves $\sqrt{1+\Delta x}$, a reasonable approach is to make $1+\Delta x$ a perfect square. Thus, let

$$1+\Delta x = 1+\Delta s+\frac{(\Delta s)^2}{4} = \left(1+\frac{\Delta s}{2}\right)^2.$$

Then

$$\Delta x = \Delta s + \frac{(\Delta s)^2}{4} = \varphi(\Delta s)$$

is a pullback substitution that does what we want:

$$\begin{aligned} f : \Delta y &= -1+\sqrt{1+\Delta x} = -1+\sqrt{1+\varphi(\Delta s)} \\ &= -1+\sqrt{1+\Delta s+\frac{(\Delta s)^2}{4}} = -1+\left(1+\frac{\Delta s}{2}\right) = \tfrac{1}{2}\Delta s. \end{aligned}$$

Thus the formula for f in the $(\Delta s, \Delta y)$ window is identical to the formula for $\mathrm{d}f_1$ in the $(\Delta x, \Delta y)$ window.

The pullback map

Let us extend our pullback to a map $\boldsymbol{\varphi} : (\Delta s, \Delta y) \mapsto (\Delta x, \Delta y)$ of one window to the other:

$$\boldsymbol{\varphi} : \begin{cases} \Delta x = \varphi(\Delta s), \\ \Delta y = \Delta y. \end{cases}$$

We see the effect of $\boldsymbol{\varphi}$ in the figure above, on the left. For a start, $\boldsymbol{\varphi}$ pulls back the uniform grid to the nonuniform one shown. The numbers at the bottom of the vertical grid lines are the Δx-values in both cases. Pick a vertical line with the same Δx value in each of the windows; you should check that, at a point where the black graphs cross those lines, the Δy coordinates agree. This means that the black line in the $(\Delta s, \Delta y)$ window is the graph of the same function—namely f—as the black curved line in the $(\Delta x, \Delta y)$ window.

Nonlinearity of the pullback

The pullback gradually stretches the grid on the left and compresses it on the right. This is just a geometric manifestation of the nonlinearity of the map $\boldsymbol{\varphi}$. Near the origin, there is virtually no distortion in the grid. In other words, the "coordinate change" does not change anything there. (This is a consequence of $\varphi'(0) = 1$.) The nonlinearity of $\boldsymbol{\varphi}$ makes it possible to straighten the curved graph of f. Of course, the same nonlinearity causes the straight tangent line to bend into a curve, in this case the parabolic curve

$$\Delta y = \frac{1}{2}\left(\Delta s + \frac{(\Delta s)^2}{4}\right).$$

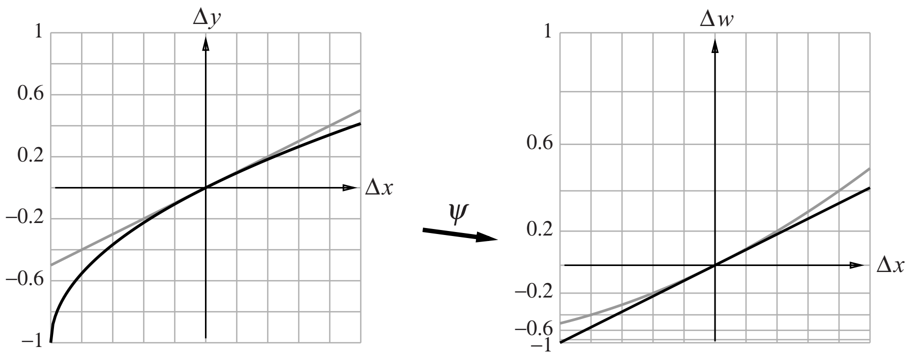
In terms of its own coordinate, the $(\Delta s, \Delta y)$ window covers the horizontal range $-2 \le \Delta s \le 2(-1+\sqrt{2}) \approx 0.8$. The points $\Delta s = -2$, -1, and 0.8 are marked on the Δs-axis; compare them to nearby values of Δx.

Changing y instead of x

We just converted f into its derivative by changing the source variable x. We can accomplish the same thing in a different way by making an appropriate change in the target variable y. In the exercises you are asked to find an explicit formula for a push-forward substitution $\Delta w = \psi(\Delta y)$ that converts

$$f : \Delta y = -1 + \sqrt{1+\Delta x}$$

into $\Delta w = \frac{1}{2}\Delta x$. The figure below shows the form that the coordinate change takes: to straighten the graph of f, the bottom of the $(\Delta x, \Delta y)$ window must be compressed (quite severely near $\Delta y = -1$), and the top stretched.



Example 2: semi-log paper

The two maps $\boldsymbol{\varphi}$ and $\boldsymbol{\psi}$ suggest a general principle: to convert a curved graph to a straight one, plot it on a suitable nonuniform grid. Perhaps the most familiar example of this is semi-log graph paper, on which an exponential function plots as a straight line. We take this now as our second example of a coordinate change that simplifies the geometry of functions.

To be concrete, consider the function $g(x) = 3 \times 10^{0.1x}$. We use base 10 here because the usual semi-log paper is geared to it (rather than to base e, for example). On the left, below, is the graph of g; it has, of course, the familar shape of an exponential curve. The coordinate change $Y = \log_{10} y$ (a push-forward substitution) gives

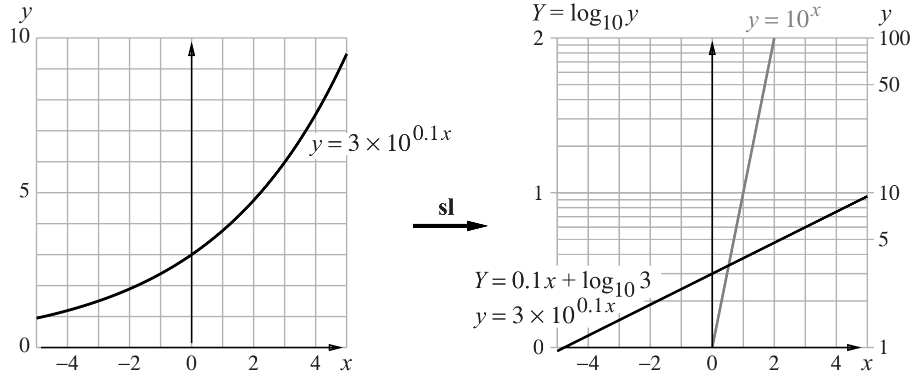
$$Y = \log_{10} y = \log_{10}\left(3 \times 10^{0.1x}\right) = 0.1x + \log_{10} 3,$$

making Y a linear function of x. Its graph is the straight line shown in black, on the right. For comparison, the graph of $y = 10^x$ is shown in gray. It is also a straight line, with a slope 10 times steeper than the black graph.

The semi-log map and exponentials

You can verify that the **semi-log** map,

$$\mathbf{sl}: \begin{cases} x = x, \\ Y = \log_{10} y, \end{cases}$$



shown above, converts any exponential function $y = Ba^{kx}$ into a linear one:

$$Y = (k \log_{10} a)\, x + \log_{10} B;$$

see the exercises. It compresses the image of a uniform grid more and more in the vertical direction. In particular, notice that the vertical spacing is $\Delta y = 1$ on the lower half of the image grid but $\Delta y = 10$ on the upper. (Although the grid immediately below $Y = 0$ is not shown, you will find it repeats the same nonuniform pattern but with a spacing of $\Delta y = 0.1$.) Semi-log paper has two virtues that are commonly exploited. First, it allows data values that vary over several orders of magnitude to be plotted in a small space. Second, it makes exponential growth or decline easier to discern and to quantify, by plotting it on a straight line. You can explore these features, and the related log–log map, in the exercises.

Example 3: the quadratic map

For our third example we move to a 2-dimensional source and target, and to the quadratic map (Chapter 4.2)

$$\mathbf{f}: \begin{cases} x = u^2 - v^2, \\ y = 2uv, \end{cases} \qquad d\mathbf{f}_{(a,b)}: \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} = \begin{pmatrix} 2a & -2b \\ 2b & 2a \end{pmatrix} \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix}.$$

Let us see how a suitable coordinate change near an arbitrary point $(u,v) = (a,b)$ can convert $\mathbf{f}$ into its derivative $d\mathbf{f}_{(a,b)}$. As usual, we set

$$\begin{aligned} \Delta u &= u - a, & \Delta x &= x - (a^2 - b^2), \\ \Delta v &= v - b, & \Delta y &= y - 2ab, \end{aligned}$$

to get coordinates $(\Delta u, \Delta v)$ in a window centered at $(u,v) = (a,b)$ and coordinates $(\Delta x, \Delta y)$ in a window centered at the image point $(x,y) = \mathbf{f}(a,b) = (a^2 - b^2, 2ab)$.

In window coordinates, we represent the map $\mathbf{f}$ by the **window map** $\Delta\mathbf{f}$ (see below, p. 172) defined by

$$(\Delta x, \Delta y) = \Delta\mathbf{f}(\Delta u, \Delta v) = \mathbf{f}(a+\Delta u, b+\Delta v) - \mathbf{f}(a,b).$$

The formula for $\Delta\mathbf{f}$ is therefore

$$\begin{aligned} \Delta x &= (a+\Delta u)^2 - (b+\Delta v)^2 - (a^2 - b^2) & \Delta y &= 2(a+\Delta u)(b+\Delta v) - 2ab \\ &= 2a\,\Delta u - 2b\,\Delta v + (\Delta u)^2 - (\Delta v)^2, & &= 2b\,\Delta u + 2a\,\Delta v + 2\,\Delta u\,\Delta v. \end{aligned}$$

Our goal is to change coordinates in the source window,

$$\mathbf{h} : \begin{cases} \Delta s = h(\Delta u, \Delta v), \\ \Delta t = k(\Delta u, \Delta v), \end{cases}$$

so that, in terms of the new coordinates, the formula for $\Delta\mathbf{f}$ becomes the formula for $d\mathbf{f}_{(a,b)}$. That is, $\Delta\mathbf{f}$ expresses Δx and Δy as the linear functions

$$\Delta\mathbf{f} : \begin{cases} \Delta x = 2a\,\Delta s - 2b\,\Delta t, \\ \Delta y = 2b\,\Delta s + 2a\,\Delta t. \end{cases}$$

Solving for Δs and Δt

Note that we now have two expressions for Δx (and, likewise, for Δy), one involving Δu and Δv, the other Δs and Δt. To find the functions h and k that connect Δs and Δt with Δu and Δv, we can begin by equating those expressions (in matrix form):

$$\begin{pmatrix} 2a & -2b \\ 2b & 2a \end{pmatrix} \begin{pmatrix} \Delta s \\ \Delta t \end{pmatrix} = \begin{pmatrix} 2a & -2b \\ 2b & 2a \end{pmatrix} \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} + \begin{pmatrix} (\Delta u)^2 - (\Delta v)^2 \\ 2\,\Delta u\,\Delta v \end{pmatrix}.$$

Then, to solve for Δs and Δt, we need only multiply by the appropriate inverse matrix:

$$\begin{pmatrix} \Delta s \\ \Delta t \end{pmatrix} = \begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} + \frac{1}{2(a^2+b^2)} \begin{pmatrix} a & b \\ -b & a \end{pmatrix} \begin{pmatrix} (\Delta u)^2 - (\Delta v)^2 \\ 2\,\Delta u\,\Delta v \end{pmatrix}.$$

This is the coordinate change we seek; in effect, $\mathbf{h} = (d\mathbf{f}_{\mathbf{a}})^{-1} \circ \Delta\mathbf{f}$. The individual components of $\mathbf{h}$ are

$$\mathbf{h} : \begin{cases} \Delta s = h(\Delta u, \Delta v) = \Delta u + \dfrac{a(\Delta u)^2 + 2b\,\Delta u\,\Delta v - a(\Delta v)^2}{2(a^2+b^2)}, \\[2ex] \Delta t = k(\Delta u, \Delta v) = \Delta v + \dfrac{-b(\Delta u)^2 + 2a\,\Delta u\,\Delta v + b(\Delta v)^2}{2(a^2+b^2)}. \end{cases}$$

Incidentally, it is not yet evident that $\mathbf{h}$ is a coordinate change: that is, that the map $\mathbf{h}(\Delta u, \Delta v)$ has an inverse defined in some neighborhood W of $(\Delta s, \Delta t) = (0,0)$. In fact, there is such a local inverse, but rather than go through a proof in this particular case, we simply appeal to the inverse function theorem, proven later in this chapter. (In particular, see Corollary 5.4, page 176. It says $\mathbf{h}$ will have a local inverse at

$(\Delta s, \Delta t) = (0,0)$ if its derivative is continuous near $(0,0)$ and invertible at $(0,0)$. All these conditions are satisfied; in particular, $d\mathbf{h}_{(0,0)} = I$, the identity map.)

Why $\mathbf{f}$ "looks like" $d\mathbf{f}_{\mathbf{a}}$

By using the vectors

$$\Delta\mathbf{u} = (\Delta u, \Delta v), \quad \Delta\mathbf{s} = (\Delta s, \Delta t), \quad \Delta\mathbf{x} = (\Delta x, \Delta y), \quad \mathbf{a} = (a, b),$$

we can write the formula that connects the maps $\mathbf{f}$, $\mathbf{h}$, and $d\mathbf{f}_{\mathbf{a}}$ as

$$\mathbf{f}(\mathbf{a} + \Delta\mathbf{u}) - \mathbf{f}(\mathbf{a}) = \Delta\mathbf{f}(\Delta\mathbf{u}) = \Delta\mathbf{x} = d\mathbf{f}_{\mathbf{a}}(\mathbf{h}(\Delta\mathbf{u})) = d\mathbf{f}_{\mathbf{a}}(\Delta\mathbf{s}).$$

Think of the formula this way. Each point $\mathbf{p}$ in the window centered at $\mathbf{a}$ has two different coordinate labels, $\Delta\mathbf{u}$ and $\Delta\mathbf{s}$. The map $\mathbf{h}$ connects those labels. The image of $\mathbf{p}$ under the action of $\mathbf{f}$ (i.e., $\Delta\mathbf{f}$) has coordinate $\Delta\mathbf{x}_1 = \Delta\mathbf{f}(\Delta\mathbf{u})$. The image of $\mathbf{p}$ under the action of $d\mathbf{f}_{\mathbf{a}}$ has coordinate $\Delta\mathbf{x}_2 = d\mathbf{f}_{\mathbf{a}}(\Delta\mathbf{s})$. But $\Delta\mathbf{x}_1 = \Delta\mathbf{x}_2$; these are the coordinates of the same point $\mathbf{q}$. Thus, $\mathbf{f}$ (written as $\Delta\mathbf{f}$ in the window W) and $d\mathbf{f}_{\mathbf{a}}$ both map $\mathbf{p}$ to $\mathbf{q}$. That is why $\mathbf{f}$ "looks like" $d\mathbf{f}_{\mathbf{a}}$; they are just different coordinate descriptions of the same map. All of this is diagrammed on the left, below, and summarized more briefly on the right.

$$\begin{tikzcd}
 & \Delta\mathbf{s} \arrow[r, mapsto, "d\mathbf{f}_{\mathbf{a}}"] & \Delta\mathbf{x}_2 \arrow[dr, leftrightarrow] \arrow[dd, equal] & \\
\mathbf{p} \arrow[ur, leftrightarrow] \arrow[dr, leftrightarrow] & & & \mathbf{q} \\
 & \Delta\mathbf{u} \arrow[uu, mapsto, "\mathbf{h}"] \arrow[r, mapsto, "\Delta\mathbf{f}"'] & \Delta\mathbf{x}_1 \arrow[ur, leftrightarrow] &
\end{tikzcd}
\qquad
\begin{tikzcd}
 & \Delta\mathbf{s} \arrow[dr, mapsto, "d\mathbf{f}_{\mathbf{a}}"] & \\
\Delta\mathbf{u} \arrow[ur, mapsto, "\mathbf{h}"] \arrow[rr, mapsto, "\Delta\mathbf{f}"'] & & \Delta\mathbf{x}
\end{tikzcd}$$

$\mathbf{f}$ *factors through* $\mathbf{h}$

Thus we have $\Delta\mathbf{f} = d\mathbf{f}_{\mathbf{a}} \circ \mathbf{h}$. If we think of composition of maps as a kind of product, then we can say $\Delta\mathbf{f}$ *factors into* $\mathbf{h}$ and $d\mathbf{f}_{\mathbf{a}}$. In effect, we constructed the coordinate change map $\mathbf{h}$ so that, in a small window centered at $\mathbf{a}$, $\Delta\mathbf{f}$ **factors through h**.

Converting $\Delta\mathbf{f}$ to $d\mathbf{f}$ at $(\sqrt{3}/2, 1/2)$

We can get a better idea how the coordinate change $\mathbf{h}$ converts $\Delta\mathbf{f}$ (or $\mathbf{f}$) into $d\mathbf{f}_{(a,b)}$ by focusing on a specific point. In the figure below, we have taken $(a,b) = (\sqrt{3}/2, 1/2)$ and used windows that measure 1 unit on a side. Thus, the square grid in the $(\Delta s, \Delta t)$-window at the top has a spacing of 0.1 unit. The same grid appears in the lower-left window, "pulled back" by $\mathbf{h}$; it becomes curved there because $\mathbf{h}$ is nonlinear. The lower-left window therefore demonstrates concretely what we said above: that each point in the source has two sets of coordinates. The curved grid provides $(\Delta s, \Delta t)$ whereas the "native" coordinates (whose square grid is not drawn) are $(\Delta u, \Delta v)$. Thus, for example,

$$(-0.3, -0.1) = (\Delta s, \Delta t) \ \leftrightarrow\ \mathbf{p} \ \leftrightarrow\ (\Delta u, \Delta v) \approx (-0.3754, -0.0996).$$

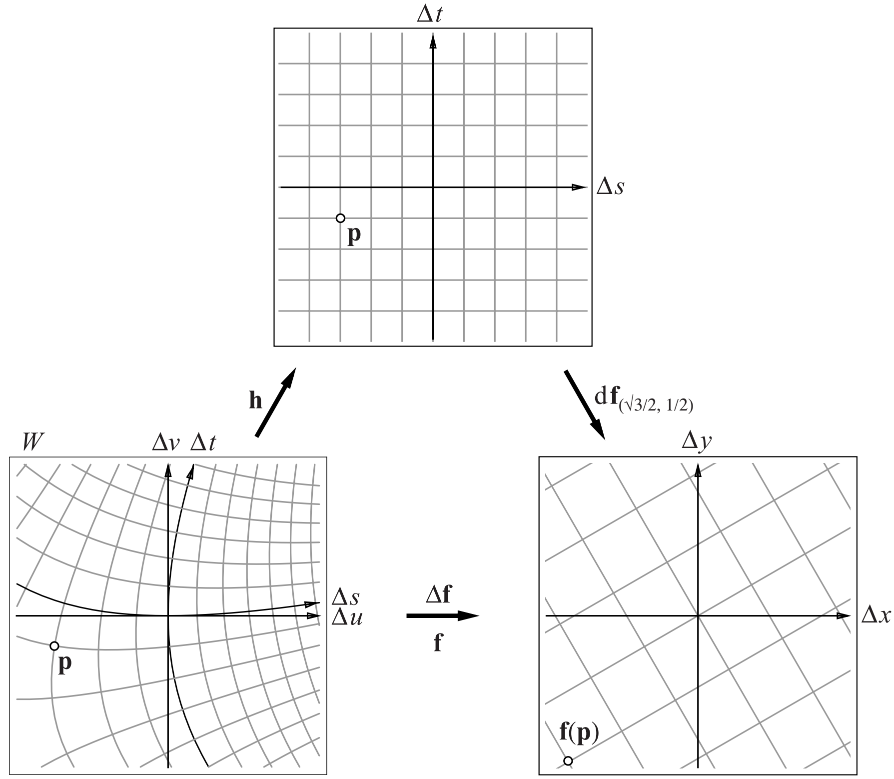
Mapping the curved grid

The curved grid is the key to visualizing both the action of $\mathbf{f}$ (as $\Delta\mathbf{f}$) and the connection between $\mathbf{f}$ and its derivative. First, follow $\mathbf{f}$ as it maps the source on the left directly to the target on the right; it sends the curved grid to the grid of large squares, straightening all grid lines in the process. Second, follow $\mathbf{h}$ and $d\mathbf{f}_{(\sqrt{3}/2,1/2)}$ into and out of the upper window. This time $\mathbf{h}$ itself straightens the curved grid, mapping it to the "native" grid in the $(\Delta s, \Delta t)$-window. The linear map $d\mathbf{f}_{(\sqrt{3}/2,1/2)}$

then carries the $(\Delta s, \Delta t)$ grid to the grid of large squares in the target. Now we already know that

$$\mathbf{df}_{(\sqrt{3}/2,1/2)} = 2R_{\pi/6}$$

(rotation by $\pi/6$ radians with all lengths doubled; see p. 118), so the large squares in the $(\Delta x, \Delta y)$-window are 0.2 units on a side and make an angle of 30° with the horizontal.



Details of the $(\Delta s, \Delta t)$ coordinates

The figure has even more to say about the map $\mathbf{h}$ that pulls back the $(\Delta s, \Delta t)$ coordinates to the $(\Delta u, \Delta v)$ window. The curves that make up the new grid are, of course, just contour lines of the two functions

$$\Delta s = h(\Delta u, \Delta v), \quad \Delta t = k(\Delta u, \Delta v)$$

defined in the window. Contours of h give the roughly vertical curves; contours of k give the roughly horizontal ones. The following Mathematica 5 code generates the curved grid in the $(\Delta u, \Delta v)$-window.

```
scon = ContourPlot[u + Sqrt[3](u^2 - v^2)/4 + u v/2,
  {u, -.5, .5}, {v, -.5, .5}, Contours ->
  {-.6, -.5, -.4, -.3, -.2, -.1, 0, .1, .2, .3, .4, .5, .6},
  ContourShading -> False, FrameTicks -> None]
```

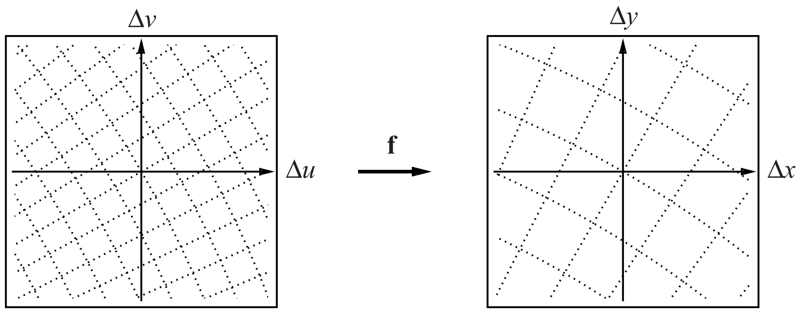
```
tcon = ContourPlot[v + Sqrt[3]u v/2 + (v^2 - u^2)/4 ,
  {u, -.5, .5}, {v, -.5, .5}, Contours ->
  {-.6, -.5, -.4, -.3, -.2, -.1, 0, .1, .2, .3, .4, .5, .6},
  ContourShading -> False, FrameTicks -> None]

Show[scon, tcon]
```

The two sets of curves are everywhere orthogonal. This is not automatic. It happens because the map $\mathbf{h}$ is conformal (p. 118); see Exercise 5.19. Note furthermore how the axes are pulled back: the Δs-axis is tangent to the Δu-axis, and the Δt-axis to the Δv-axis. Moreover, the grid squares around the origin undergo almost no distortion: the pullbacks are nearly the same size and shape as the original. This is a consequence of $d\mathbf{h}_\mathbf{0} = I$.

Curvilinear coordinates

Because the coordinate lines $\Delta s =$ constant and $\Delta t =$ constant become curved when they appear in the $(\Delta u, \Delta v)$-window, we say that Δs and Δt are **curvilinear coordinates** there. As we have just seen, curvilinear coordinates can simplify our view of a map. This is a trade-off, of course: to simplify the map, we complicate the coordinates. But this is a cost we have already accepted when, for example, we plot exponential functions on semi-log paper. We have also accepted it when we use polar coordinates: it was a curved polar coordinate grid that first clarified the action of the quadratic map $\mathbf{f}$. Here is our earlier view (p. 117) of the local action of $\mathbf{f}$ in a window centered at $(\sqrt{3}/2, 1/2)$:



Compare this figure now with the new one (p. 164) that uses the curvilinear $(\Delta s, \Delta t)$ coordinates in the $(\Delta u, \Delta v)$-window.

Summary: Under a suitable coordinate change, a complicated situation can often be made simpler; for example, it may be possible to convert a map locally into its derivative.

5.3 The inverse function theorem

As we have seen, coordinate changes give us a powerful tool to simplify the description of a map. But a coordinate change must be invertible, a condition that is often difficult to verify directly. In this section we state and prove the inverse function theorem. Simply put, the theorem says that if the derivative of a map is continuous and invertible at a point, the map itself is invertible locally near that point. The proof uses several tools, beginning with the *contraction mapping principle*, which can be nicely illustrated by the model village found in Bourton-on-the-Water.

The nested models of Bourton-on-the-Water

Bourton-on-the-Water is in the English Cotswold hills, near Oxford. Filling the back garden of one of its houses is a scale model of the whole village. Now every point in the model village corresponds to a point in the actual village, so some point in the model must correspond exactly to *itself*. Which one? Because the model contains a copy of everything in the village, you would expect it to contain a copy of the model itself. It does; you can see it in the foreground of the photo above. The model of the model is small, of course; it covers only a few square meters. And that smaller model likewise contains a copy of everything, so it has a still smaller copy of itself. In theory, the nested copies could go on forever, getting smaller and smaller and converging, ultimately, to a single point. However, the third iteration was the last that was practical to build. Now return to our question about the point in the model that corresponds to itself. It is in the first model, by definition, but it must be in the second model, too, and the third, and so on. The point that corresponds to itself—the point that is left fixed by the model—must therefore be the limit point of the nested sequence of models.

A more formal view

A little more formally, the model defines a map $\mathbf{m} : V \to V$ of the village, V, to itself, and the point in the model that corresponds to itself is the fixed point of that map (Definition 5.1, p. 157). The model is built at the scale $\sigma = 1/9$. Thus, if $\mathbf{x}$ and

$\mathbf{y}$ are any points in V, and $\mathbf{m}(\mathbf{x})$ and $\mathbf{m}(\mathbf{y})$ are their copies in the model, then

$$\|\mathbf{m}(\mathbf{x}) - \mathbf{m}(\mathbf{y})\| = \sigma \|\mathbf{x} - \mathbf{y}\|.$$

Now $\mathbf{x}$ appears in the model of the model at the point

$$\mathbf{m}(\mathbf{m}(\mathbf{x})) = (\mathbf{m} \circ \mathbf{m})(\mathbf{x}) = \mathbf{m}^2(\mathbf{x});$$

it appears in the model of the model of the model at $\mathbf{m}^3(\mathbf{x})$, and so forth. For the kth iterate of the model, the scale factor would be σ^k:

$$\|\mathbf{m}^k(\mathbf{x}) - \mathbf{m}^k(\mathbf{y})\| = \sigma^k \|\mathbf{x} - \mathbf{y}\|.$$

Because $\sigma < 1$, $\sigma^k \to 0$ as $k \to \infty$; the size of the kth iterate shrinks to zero. Intuitively, this forces the nested models to converge to a single point, say $\mathbf{p}$. Now $\mathbf{p}$ is in every iterate, so it must be the fixed point of the model: $\mathbf{m}(\mathbf{p}) = \mathbf{p}$.

Although the contraction mapping principle can be stated quite generally, we need only the special circumstance of maps defined on a *ball* in $\mathbb{R}^n$.

Contraction mapping principle

Definition 5.2 *The* ***ball*** B_r *of radius* r *centered at the origin in* $\mathbb{R}^n$ *is the set of all points* $\mathbf{x}$ *in* $\mathbb{R}^n$ *for which* $\|\mathbf{x}\| \le r$.

Definition 5.3 *A* ***contraction mapping*** *on* B_r *is a map* $\mathbf{m} : B_r \to B_r$ *for which*

$$\|\mathbf{m}(\mathbf{x}) - \mathbf{m}(\mathbf{y})\| \le \sigma \|\mathbf{x} - \mathbf{y}\|$$

for some $\sigma < 1$ *and for all* $\mathbf{x}$, $\mathbf{y}$ *in* B_r.

A contraction mapping is thus somewhat more general than a "scale model" map, where all distances are contracted by exactly the same factor. Here the factor can vary, as long as it is bounded by a fixed $\sigma < 1$. The additional generality does not weaken the contraction mapping principle, nor does it make the proof more difficult.

Theorem 5.1 (Contraction mapping principle). *A contraction mapping* $\mathbf{m}$ *on* B_r *has a unique fixed point* $\widehat{\mathbf{x}}$ *in* B_r. *Moreover, for any* $\mathbf{x}$ *in* B_r,

$$\widehat{\mathbf{x}} = \lim_{k \to \infty} \mathbf{m}^k(\mathbf{x}).$$

Proof. Pick $\mathbf{x}_0$ arbitrarily in B_r, and let $\mathbf{x}_k = \mathbf{m}^k(\mathbf{x}_0)$. For the "telescoping" sum

$$\mathbf{x}_k - \mathbf{x}_{k+l} = \mathbf{x}_k - \mathbf{x}_{k+1} + \mathbf{x}_{k+1} - \mathbf{x}_{k+2} + \cdots + \mathbf{x}_{k+l-1} - \mathbf{x}_{k+l},$$

we have

$$\|\mathbf{x}_k - \mathbf{x}_{k+l}\| \le \|\mathbf{x}_k - \mathbf{x}_{k+1}\| + \|\mathbf{x}_{k+1} - \mathbf{x}_{k+2}\| + \cdots + \|\mathbf{x}_{k+l-1} - \mathbf{x}_{k+l}\|.$$

Now, for any $i \ge 0$,

$$\|\mathbf{x}_{k+i} - \mathbf{x}_{k+i+1}\| = \|\mathbf{m}^{k+i}(\mathbf{x}_0) - \mathbf{m}^{k+i}(\mathbf{x}_1)\| \le \sigma^{k+i} \|\mathbf{x}_0 - \mathbf{x}_1\|;$$

therefore the previous inequality implies

$$\begin{aligned}\|\mathbf{x}_k - \mathbf{x}_{k+l}\| &\le \sigma^k \|\mathbf{x}_0 - \mathbf{x}_1\| + \sigma^{k+1} \|\mathbf{x}_0 - \mathbf{x}_1\| + \cdots + \sigma^{k+l-1} \|\mathbf{x}_0 - \mathbf{x}_1\| \\ &\le \sigma^k \|\mathbf{x}_0 - \mathbf{x}_1\| \left(1 + \sigma + \sigma^2 + \cdots \sigma^{l-1}\right) \\ &= \sigma^k \|\mathbf{x}_0 - \mathbf{x}_1\| \frac{1-\sigma^l}{1-\sigma}.\end{aligned}$$

Because $\sigma^k \to 0$ as $k \to \infty$, the last inequality implies that

$$\lim_{k\to\infty} \|\mathbf{x}_k - \mathbf{x}_{k+l}\| = 0 \text{ for any integer } l > 0.$$

Let $x_k^{(j)}$ be the jth coordinate of $\mathbf{x}_k$; then

$$\left|x_k^{(j)} - x_{k+l}^{(j)}\right|^2 \le \left|x_k^{(1)} - x_{k+l}^{(1)}\right|^2 + \cdots + \left|x_k^{(n)} - x_{k+l}^{(n)}\right|^2 = \|\mathbf{x}_k - \mathbf{x}_{k+l}\|^2.$$

Thus, for each fixed $l \ge 0$ and for every $j = 1, 2, \ldots, n$,

$$\lim_{k\to\infty} \left|x_k^{(j)} - x_{k+l}^{(j)}\right| = 0.$$

Lemma 5.1. *If y_k is a sequence of real numbers for which $|y_k - y_{k+l}| \to 0$ as $k \to \infty$, (for any positive integer l), then y_k has a limiting value, $\widehat{y}$, as $k \to \infty$.*

Proof. See an analysis text for a proof of this basic fact ("Every Cauchy sequence of real numbers has a limit"). □

The lemma permits us to define

$$\widehat{x}^{(j)} = \lim_{k\to\infty} x_k^{(j)}, \quad j = 1, 2, \ldots, n, \text{ and then } \widehat{\mathbf{x}} = (\widehat{x}^{(1)}, \ldots, \widehat{x}^{(n)}).$$

In other words, $\widehat{\mathbf{x}} = \lim_{k\to\infty} \mathbf{x}_k$.

Lemma 5.2. *The contraction map $\mathbf{m} : B_r \to B_r$ is continuous at every point of B_r.*

Proof. We must show $\mathbf{m}(\mathbf{x}_n) \to \mathbf{m}(\mathbf{x})$ when $\mathbf{x}_n \to \mathbf{x}$. But we have

$$\|\mathbf{m}(\mathbf{x}_n) - \mathbf{m}(\mathbf{x})\| \le \sigma \|\mathbf{x}_n - \mathbf{x}\| \to 0.$$

(In fact, the inequality implies that $\mathbf{m}$ is *uniformly* continuous, even if $\sigma \ge 1$; see an analysis text.) □

Because $\mathbf{m}$ is continuous (i.e., it commutes with limit processes),

$$\mathbf{m}(\widehat{\mathbf{x}}) = \mathbf{m}\left(\lim_{k\to\infty} \mathbf{x}_k\right) = \lim_{k\to\infty} \mathbf{m}(\mathbf{x}_k) = \lim_{k\to\infty} \mathbf{x}_{k+1} = \widehat{\mathbf{x}},$$

so $\widehat{\mathbf{x}}$ is a fixed point of $\mathbf{m}$. Here is a different argument, which does not depend explicitly on the continuity of $\mathbf{m}$. It begins with the "telescoping" sum

$$\widehat{\mathbf{x}} - \mathbf{m}(\widehat{\mathbf{x}}) = \widehat{\mathbf{x}} - \mathbf{x}_{k+1} + \mathbf{m}(\mathbf{x}_k) - \mathbf{m}(\widehat{\mathbf{x}}),$$

which implies

$$\|\widehat{\mathbf{x}} - \mathbf{m}(\widehat{\mathbf{x}})\| \le \|\widehat{\mathbf{x}} - \mathbf{x}_{k+1}\| + \|\mathbf{m}(\mathbf{x}_k) - \mathbf{m}(\widehat{\mathbf{x}})\| \le \|\widehat{\mathbf{x}} - \mathbf{x}_{k+1}\| + \sigma\|\mathbf{x}_k - \widehat{\mathbf{x}}\|,$$

an inequality that is true for all k. But because $\mathbf{x}_k \to \widehat{\mathbf{x}}$ as $k \to \infty$, the right-hand side vanishes as $k \to \infty$, leaving $\|\widehat{\mathbf{x}} - \mathbf{m}(\widehat{\mathbf{x}})\| = 0$, or $\widehat{\mathbf{x}} = \mathbf{m}(\widehat{\mathbf{x}})$.

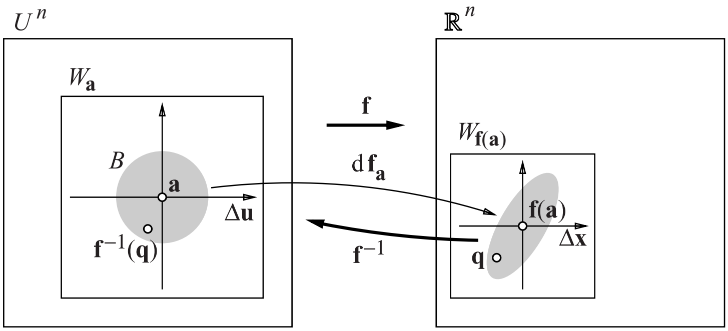
It remains to verify that the fixed point is unique. If $\widehat{\mathbf{y}}$ is also a fixed point, then

$$\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\| = \|\mathbf{m}(\widehat{\mathbf{x}}) - \mathbf{m}(\widehat{\mathbf{x}})\| \le \sigma \|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\|.$$

If $\|\widehat{\mathbf{x}} - \widehat{\mathbf{y}}\| \neq 0$, we can divide both sides of this inequality and get $1 \le \sigma$, contradicting our assumption that $\sigma < 1$. This forces $\widehat{\mathbf{x}} = \widehat{\mathbf{y}}$. □

Theorem 5.2 (Inverse function theorem). *Suppose* $\mathbf{f} : U^n \to \mathbb{R}^n$ *is continuously differentiable on* U^n*, and its derivative is invertible at the point* $\mathbf{a}$ *in* U^n*. Then* $\mathbf{f}$ *itself is invertible on the image* $d\mathbf{f}_{\mathbf{a}}(B)$ *in the target of some ball* B *of positive radius centered at the point* $\mathbf{a}$*. The inverse is continuously differentiable on its domain, and* $d(\mathbf{f}^{-1})_{\mathbf{q}} = (d\mathbf{f}_{\mathbf{f}^{-1}(\mathbf{q})})^{-1}$ *for all* $\mathbf{q}$ *in* $d\mathbf{f}_{\mathbf{a}}(B)$.

Proof. Our proof expands an argument found in Lang [10, 11].. The proof is long; therefore we split it into a number of steps.



The inverse is a *local* object, that is, one defined essentially in some window centered at the image point $\mathbf{f}(\mathbf{a})$. Therefore, we begin by setting up windows and introducing window coordinates. Let $W_{\mathbf{a}}$ be a window centered at $\mathbf{a}$ in the source, with coordinate $\Delta\mathbf{u} = \mathbf{u} - \mathbf{a}$. Similarly, let $W_{\mathbf{f(a)}}$ be a window in the target centered at $\mathbf{f}(\mathbf{a})$; its coordinate is $\Delta\mathbf{x} = \mathbf{x} - \mathbf{f}(\mathbf{a})$. Finding $\mathbf{f}^{-1}$ means solving the equation Step 1

$$\mathbf{f}(\mathbf{a} + \Delta\mathbf{u}) - \mathbf{f}(\mathbf{a}) = \Delta\mathbf{x}$$

for $\Delta\mathbf{u}$ in $W_{\mathbf{a}}$, given $\Delta\mathbf{x}$ in some suitable region in $W_{\mathbf{f(a)}}$. In fact, the $\Delta\mathbf{u}$ we seek is a fixed point of the map

$$\mathbf{g}(\Delta\mathbf{u}) = \Delta\mathbf{u} + (d\mathbf{f}_{\mathbf{a}})^{-1}[\Delta\mathbf{x} - (\mathbf{f}(\mathbf{a} + \Delta\mathbf{u}) - \mathbf{f}(\mathbf{a}))].$$

The bulk of the proof of invertibility involves showing that $\mathbf{g}$ is a contraction mapping on a suitable ball B_r, implying that $\Delta\mathbf{u}$ exists. The map $(\mathrm{d}\mathbf{f}_\mathbf{a})^{-1}$ is needed here to bring the point $\Delta\mathbf{x} - (\mathbf{f}(\mathbf{a}+\Delta\mathbf{u}) - \mathbf{f}(\mathbf{a}))$ from $W_{\mathbf{f}(\mathbf{a})}$ back to $W_\mathbf{a}$, where it can be added to $\Delta\mathbf{u}$.

Step 2

Our analysis of $\mathbf{g}$ begins with the portion

$$\boldsymbol{\varphi}(\Delta\mathbf{u}) = (\mathrm{d}\mathbf{f}_\mathbf{a})^{-1}\big(\mathbf{f}(\mathbf{a}+\Delta\mathbf{u}) - \mathbf{f}(\mathbf{a})\big),$$

$$\begin{tikzcd} & W_{\mathbf{f}(\mathbf{a})} \arrow[dr, "(\mathrm{d}\mathbf{f}_\mathbf{a})^{-1}"] & \\ W_\mathbf{a} \arrow[ur, "\mathbf{f}"] \arrow[rr, "\varphi"'] & & W_\mathbf{a} \end{tikzcd}$$

which is a map $\boldsymbol{\varphi} : W_\mathbf{a} \to W_\mathbf{a}$. By the chain rule,

$$\mathrm{d}\boldsymbol{\varphi}_{\Delta\mathbf{v}} = (\mathrm{d}\mathbf{f}_\mathbf{a})^{-1} \circ \mathrm{d}\mathbf{f}_{\mathbf{a}+\Delta\mathbf{v}}$$

for every $\Delta\mathbf{v}$ in $W_\mathbf{a}$. By hypothesis, $\mathrm{d}\mathbf{f}_{\mathbf{a}+\Delta\mathbf{v}}$ depends continuously on $\Delta\mathbf{v}$; therefore, $\mathrm{d}\boldsymbol{\varphi}_{\Delta\mathbf{v}}$ depends continuously on $\Delta\mathbf{v}$, as well. Furthermore, $\mathrm{d}\boldsymbol{\varphi}_\mathbf{0} = (\mathrm{d}\mathbf{f}_\mathbf{a})^{-1} \circ \mathrm{d}\mathbf{f}_\mathbf{a} = I$.

Step 3

Define $\mathbf{h} : W_\mathbf{a} \to W_\mathbf{a}$ by

$$\mathbf{h}(\Delta\mathbf{u}) = \Delta\mathbf{u} - \boldsymbol{\varphi}(\Delta\mathbf{u});$$

then $\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}} = I - \mathrm{d}\boldsymbol{\varphi}_{\Delta\mathbf{v}}$ is likewise continuous as a function of $\Delta\mathbf{v}$. It follows that the real-valued function

$$N(\Delta\mathbf{v}) = \max_{\|\Delta\mathbf{u}\|=1} \|\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{u})\|$$

is a continuous function of $\Delta\mathbf{v}$, as well. Because $\mathrm{d}\mathbf{h}_\mathbf{0} = \mathbf{0}$, the zero linear map, we have $N(\mathbf{0}) = 0$, and continuity implies that $N(\Delta\mathbf{v})$ will be small when $\Delta\mathbf{v}$ is small. Specifically, choose $r > 0$ so that

$$\|\Delta\mathbf{v}\| < 2r \implies |N(\Delta\mathbf{v})| < \tfrac{1}{2}.$$

Step 4

The value of r now set in Step 3 allows us to specify both the $\Delta\mathbf{x}$-values that we allow (for the domain of $\mathbf{f}^{-1}$) and the domain of $\mathbf{g}$ itself. First, we require that $\Delta\mathbf{x}$ be in the image $\mathrm{d}\mathbf{f}_\mathbf{a}(B_{r/2})$ ($B_{r/2}$ is the ball of radius $r/2$ at the center of $W_\mathbf{a}$); this means

$$\|(\mathrm{d}\mathbf{f}_\mathbf{a})^{-1}(\Delta\mathbf{x})\| \le r/2.$$

Second, we require that the domain of $\mathbf{g}$ be restricted to B_r; this means

$$\|\Delta\mathbf{u}\| \le r.$$

The next three steps show that $\mathbf{g}$ maps B_r to itself and is a contraction mapping there.

Step 5

Because $\mathbf{g} = \mathbf{h} + (\mathrm{d}\mathbf{f}_\mathbf{a})^{-1}(\Delta\mathbf{x})$, that is, $\mathbf{g}$ and $\mathbf{h}$ just "differ by a constant," most of what we still need to prove about $\mathbf{g}$ can be done by working with $\mathbf{h}$.

Lemma 5.3. *If* $\|\Delta\mathbf{v}\| < 2r$*, then* $\|\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{w})\| < \frac{1}{2}\|\Delta\mathbf{w}\|$ *for all* $\Delta\mathbf{w}$.

Proof. We can assume $\Delta\mathbf{w}$ is nonzero; then $\Delta\mathbf{u} = \Delta\mathbf{w}/\|\Delta\mathbf{w}\|$ is a unit vector and $\Delta\mathbf{w} = \|\Delta\mathbf{w}\|\,\Delta\mathbf{u}$. Because $\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}$ is linear, we have

$$\|\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{w})\| = \|\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}(\|\Delta\mathbf{w}\|\Delta\mathbf{u})\| = \|\Delta\mathbf{w}\|\;\|\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{u})\| \le \|\Delta\mathbf{w}\|\,N(\Delta\mathbf{v}) < \tfrac{1}{2}\|\Delta\mathbf{w}\|.$$

We have used the fact that $\|\mathrm{d}\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{u})\| \le N(\Delta\mathbf{v})$ for any $\Delta\mathbf{v}$ and for all unit vectors $\Delta\mathbf{u}$, by the definition of N (Step 2). $\square$

Suppose $\Delta\mathbf{u}_1$ and $\Delta\mathbf{u}_2$ are in the ball of radius r, then the entire line segment $\Delta\mathbf{v} = \Delta\mathbf{u}_1 + t(\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1)$ $(0 \le t \le 1)$ is likewise, and we have Step 6

$$\|\mathrm{d}\mathbf{h}_{\Delta\mathbf{u}_1+t(\Delta\mathbf{u}_2-\Delta\mathbf{u}_1)}(\Delta\mathbf{w})\| \le \tfrac{1}{2}\|\Delta\mathbf{w}\|$$

for all $\Delta\mathbf{w}$. This inequality and the continuous differentiability of $\mathbf{h}$ (Step 2) allow us to use the mean value theorem for maps (Theorem 4.15, p. 140) to conclude

$$\|\mathbf{h}(\Delta\mathbf{u}_2) - \mathbf{h}(\Delta\mathbf{u}_1)\| \le \tfrac{1}{2}\|\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1\|.$$

Moreover, if we set $\Delta\mathbf{u}_1 = \mathbf{0}$, then $\|\mathbf{h}(\Delta\mathbf{u}_2)\| \le \frac{1}{2}\|\Delta\mathbf{u}_2\| \le r/2$. In other words, $\mathbf{h}$ maps the ball of radius r into the ball of radius $r/2$.

We now move on to $\mathbf{g}$ itself. Step 7

Lemma 5.4. *For any* $\Delta\mathbf{x}$ *in* $\mathrm{d}\mathbf{f}_{\mathbf{a}}(B_{r/2})$, $\mathbf{g} : B_r \to B_r$.

Proof. Because $\Delta\mathbf{u}$ is in B_r, by hypothesis, we have $\|\mathbf{h}(\Delta\mathbf{u}\| \le r/2$ (Step 6). Also by hypothesis, $\|(\mathrm{d}\mathbf{f}_{\mathbf{a}})^{-1}(\Delta\mathbf{x})\| \le r/2$, so

$$\|\mathbf{g}(\Delta\mathbf{u})\| = \|\mathbf{h}(\Delta\mathbf{u}) + (\mathrm{d}\mathbf{f}_{\mathbf{a}})^{-1}(\Delta\mathbf{x})\| \le \|\mathbf{h}(\Delta\mathbf{u})\| + \|(\mathrm{d}\mathbf{f}_{\mathbf{a}})^{-1}(\Delta\mathbf{x})\| \le r. \qquad \square$$

Lemma 5.5. *For any* $\Delta\mathbf{x}$ *in* $\mathrm{d}\mathbf{f}_{\mathbf{a}}(B_{r/2})$, $\mathbf{g}$ *is a contraction mapping on* B_r.

Proof. Because $\mathbf{g}(\Delta\mathbf{u}) = \mathbf{h}(\Delta\mathbf{u}) + (\mathrm{d}\mathbf{f}_{\mathbf{a}})^{-1}(\Delta\mathbf{x})$, it follows that

$$\mathbf{g}(\Delta\mathbf{u}_2) - \mathbf{g}(\Delta\mathbf{u}_1) = \mathbf{h}(\Delta\mathbf{u}_2) - \mathbf{h}(\Delta\mathbf{u}_1);$$

therefore, by Step 6,

$$\|\mathbf{g}(\Delta\mathbf{u}_2) - \mathbf{g}(\Delta\mathbf{u}_1)\| = \|\mathbf{h}(\Delta\mathbf{u}_2) - \mathbf{h}(\Delta\mathbf{u}_1)\| \le \tfrac{1}{2}\|\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1\|. \qquad \square$$

Let $\widehat{\Delta\mathbf{x}}$ be an arbitrary point in $\mathrm{d}\mathbf{f}_{\mathbf{a}}(B_{r/2})$. This choice determines a specific map $\mathbf{g} : B_r \to B_r$, and $\mathbf{g}$ has a unique fixed point $\widehat{\Delta\mathbf{u}}$ in B_r, by the contraction mapping principle. Because $\widehat{\Delta\mathbf{x}}$ determines $\widehat{\Delta\mathbf{u}}$ uniquely, and Step 8

$$\mathbf{f}(\mathbf{a} + \widehat{\Delta\mathbf{u}}) - \mathbf{f}(\mathbf{a}) = \widehat{\Delta\mathbf{x}},$$

we now have the required inverse map $\mathbf{f}^{-1} : \mathrm{d}\mathbf{f}_{\mathbf{a}}(B_{r/2}) \to B_r : \widehat{\Delta\mathbf{x}} \mapsto \widehat{\Delta\mathbf{u}}$.

Before showing that $\mathbf{f}^{-1}$ is continuously differentiable, we pause to call attention to the relation between a map and the way we write it within a window. For example, in $W_{\mathbf{a}}$ and $W_{\mathbf{f}(\mathbf{a})}$, the equation $\mathbf{x} = \mathbf{f}(\mathbf{u})$ has the form Step 9

$$\mathbf{f}(\mathbf{a}) + \underline{\Delta\mathbf{x}} = \mathbf{x} = \mathbf{f}(\mathbf{u}) = \mathbf{f}(\mathbf{a} + \Delta\mathbf{u}) = \mathbf{f}(\mathbf{a}) + \underline{\mathbf{f}(\mathbf{a} + \Delta\mathbf{u}) - \mathbf{f}(\mathbf{a})}.$$

Window map and equation

The underlined elements are equal (this is the "window equation"), and they define the **window map** $\Delta\mathbf{f}$ for $\mathbf{f}$:

$$\Delta\mathbf{x} = \Delta\mathbf{f}(\Delta\mathbf{u}) = \mathbf{f}(\mathbf{a}+\Delta\mathbf{u}) - \mathbf{f}(\mathbf{a}).$$

Conversely, we can reconstruct the original formula $\mathbf{x} = \mathbf{f}(\mathbf{u})$ from its window equation $\Delta\mathbf{x} = \Delta\mathbf{f}(\Delta\mathbf{u})$. Furthermore, by solving the window equation for $\Delta\mathbf{u}$, we obtain the window equation of the inverse $\mathbf{u} = \mathbf{f}^{-1}(\mathbf{x})$ (at the point $\mathbf{b} = \mathbf{f}(\mathbf{a})$):

$$\Delta\mathbf{u} = \Delta\mathbf{f}^{-1}(\Delta\mathbf{x}) = \mathbf{f}^{-1}(\mathbf{b}+\Delta\mathbf{x}) - \mathbf{f}^{-1}(\mathbf{b}).$$

Step 10

In preparation for showing $\mathbf{f}^{-1}$ is differentiable, we first show that it is uniformly continuous, by working with the window map $\Delta\mathbf{f}^{-1}$.

Lemma 5.6. *There is a positive constant K such that, for any two points $\Delta\mathbf{x}_1$, $\Delta\mathbf{x}_2$ and their corresponding images $\Delta\mathbf{u}_1$, $\Delta\mathbf{u}_2$ under $\Delta\mathbf{f}^{-1}$,*

$$\|\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1\| \le K\|\Delta\mathbf{x}_2 - \Delta\mathbf{x}_1\|.$$

Proof. Recall the definition of the map $\mathbf{h}$ (now written using the window map for $\mathbf{f}$):

$$\mathbf{h}(\Delta\mathbf{u}) = \Delta\mathbf{u} - \boldsymbol{\varphi}(\Delta\mathbf{u}) = \Delta\mathbf{u} - (d\mathbf{f}_\mathbf{a})^{-1}(\Delta\mathbf{f}(\Delta\mathbf{u})) = \Delta\mathbf{u} - (d\mathbf{f}_\mathbf{a})^{-1}(\Delta\mathbf{x}).$$

If we now evaluate this equation at $\Delta\mathbf{u}_1$ and then at $\Delta\mathbf{u}_2$, subtract the first from the second, and use the linearity of $(d\mathbf{f}_\mathbf{a})^{-1}$, we get

$$\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1 = \mathbf{h}(\Delta\mathbf{u}_2) - \mathbf{h}(\Delta\mathbf{u}_1) + (d\mathbf{f}_\mathbf{a})^{-1}(\Delta\mathbf{x}_\mathbf{2} - \Delta\mathbf{x}_\mathbf{1}).$$

It follows that

$$\begin{aligned}\|\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1\| &\le \|\mathbf{h}(\Delta\mathbf{u}_2) - \mathbf{h}(\Delta\mathbf{u}_1)\| + \|(d\mathbf{f}_\mathbf{a})^{-1}(\Delta\mathbf{x}_\mathbf{2} - \Delta\mathbf{x}_\mathbf{1})\| \\ &\le \tfrac{1}{2}\|\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1\| + C\|\Delta\mathbf{x}_2 - \Delta\mathbf{x}_1\|.\end{aligned}$$

for some positive C (see Exercise 3.28, p. 104). The first term on the right side of the second inequality is a consequence of the contraction property of $\mathbf{h}$ (Step 6). A final subtraction gives

$$\tfrac{1}{2}\|\Delta\mathbf{u}_2 - \Delta\mathbf{u}_1\| \le C\|\Delta\mathbf{x}_2 - \Delta\mathbf{x}_1\|,$$

implying we can take $K = 2C$. □

The lemma establishes that $\Delta\mathbf{f}^{-1}$ is uniformly continuous (see the comment in the proof of Lemma 5.2, above).

Step 11

Because we claim the derivative of $\mathbf{f}^{-1}$ at the point $\mathbf{q}$ will be $(d\mathbf{f}_\mathbf{p})^{-1}$, where $\mathbf{p} = \mathbf{f}^{-1}(\mathbf{q})$ is a point in the ball B_r, we must first show $d\mathbf{f}_\mathbf{p}$ is invertible.

Lemma 5.7. *Suppose $\mathbf{p} = \mathbf{a} + \Delta\mathbf{v}$ for some $\|\Delta\mathbf{v}\| \le r$ (i.e., $\mathbf{p}$ is in B_r); then $d\mathbf{f}_\mathbf{p}$ is invertible.*

Proof. From the definition

$$\mathbf{h}(\Delta\mathbf{u}) = \Delta\mathbf{u} + (d\mathbf{f_a})^{-1}(f(\mathbf{a}+\Delta\mathbf{u}) - \mathbf{f}(\mathbf{a}))$$

in Step 2, it follows that

$$d\mathbf{h}_{\Delta\mathbf{v}} = I - (d\mathbf{f_a})^{-1} \circ d\mathbf{f}_{\mathbf{a}+\Delta\mathbf{v}}, \quad \text{or} \quad I = d\mathbf{h}_{\Delta\mathbf{v}} + (d\mathbf{f_a})^{-1} \circ d\mathbf{f_p}.$$

Therefore, when the maps in the last equation are supplied with input $\Delta\mathbf{u}$, we get

$$\Delta\mathbf{u} = d\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{u}) + (d\mathbf{f_a})^{-1}(d\mathbf{f_p}(\Delta\mathbf{u})),$$

and hence

$$\begin{aligned}\|\Delta\mathbf{u}\| &\leq \|d\mathbf{h}_{\Delta\mathbf{v}}(\Delta\mathbf{u})\| + \|(d\mathbf{f_a})^{-1}(d\mathbf{f_p}(\Delta\mathbf{u}))\| \\ &\leq \tfrac{1}{2}\|\Delta\mathbf{u}\| + C\|d\mathbf{f_p}(\Delta\mathbf{u})\|\end{aligned}$$

for some $C > 0$, exactly as in Lemma 5.6. A bit of algebra now gives

$$\frac{1}{2C}\|\Delta\mathbf{u}\| \leq \|d\mathbf{f_p}(\Delta\mathbf{u})\|.$$

This inequality implies $d\mathbf{f_p}(\Delta\mathbf{u}) \neq \mathbf{0}$ when $\Delta\mathbf{u} \neq \mathbf{0}$. In other words, the kernel of $d\mathbf{f_p}$ contains only $\mathbf{0}$, so $d\mathbf{f_p}$ is invertible. □

We now show $\mathbf{f}^{-1}$ is differentiable at an arbitrary point $\mathbf{q}$ in the domain $d\mathbf{f_a}(B_{r/2})$, and its derivative is $(d\mathbf{f_p})^{-1}$ there $(\mathbf{p} = \mathbf{f}^{-1}(\mathbf{q}))$. We work in windows $W_\mathbf{q}$ and $W_\mathbf{p}$ with local coordinates $\Delta\mathbf{y}$ and $\Delta\mathbf{v}$, respectively, and with the window equations Step 12

$$\begin{aligned}\Delta\mathbf{v} &= \Delta\mathbf{f}^{-1}(\Delta\mathbf{y}) = \mathbf{f}^{-1}(\mathbf{q}+\Delta\mathbf{y}) - \mathbf{f}^{-1}(\mathbf{q}), \\ \Delta\mathbf{y} &= \Delta\mathbf{f}(\Delta\mathbf{v}) = \mathbf{f}(\mathbf{p}+\Delta\mathbf{v}) - \mathbf{f}^{-1}(\mathbf{p}).\end{aligned}$$

To prove $\mathbf{f}^{-1}$ is differentiable at $\mathbf{q}$, and has derivative $(d\mathbf{f_p})^{-1}$ there, it is necessary and sufficient to show $\mathbf{R}(\Delta\mathbf{y}) = \boldsymbol{o}(\Delta\mathbf{y})$, where

$$\Delta\mathbf{v} = \Delta\mathbf{f}^{-1}(\Delta\mathbf{y}) = (d\mathbf{f_p})^{-1}(\Delta\mathbf{y}) + \mathbf{R}(\Delta\mathbf{y}).$$

To analyze $\mathbf{R}$, apply $d\mathbf{f_p}$ to both sides of this equation,

$$d\mathbf{f_p}(\Delta\mathbf{v}) = \Delta\mathbf{y} + d\mathbf{f_p}(\mathbf{R}(\Delta\mathbf{y})),$$

and then solve for $\Delta\mathbf{y} = \Delta\mathbf{f}(\Delta\mathbf{v})$ to get

$$\Delta\mathbf{f}(\Delta\mathbf{v}) = d\mathbf{f_p}(\Delta\mathbf{v}) - \underbrace{d\mathbf{f_p}(\mathbf{R}(\Delta\mathbf{y}))}_{\boldsymbol{o}(\Delta\mathbf{v})}.$$

This equation expresses the differentiability of $\mathbf{f}$ itself at $\mathbf{p}$, so the last term must be $\boldsymbol{o}(\Delta\mathbf{v})$, as indicated. Because $L(\boldsymbol{o}(\mathbf{u})) = \boldsymbol{o}(\mathbf{u})$ for any linear map L (cf. the proof of Lemma 4.1, p. 133),

$$\mathbf{R}(\Delta\mathbf{y}) = (\mathrm{d}\mathbf{f}_{\mathbf{p}})^{-1}(\boldsymbol{o}(\Delta\mathbf{v})) = \boldsymbol{o}(\Delta\mathbf{v}).$$

However, we need $\boldsymbol{o}(\Delta\mathbf{y})$ on the right, not just $\boldsymbol{o}(\Delta\mathbf{v})$. The uniform continuity of $\Delta\mathbf{f}^{-1}$ (Lemma 5.6) in this setting implies $\|\Delta\mathbf{v}\| \le K\|\Delta\mathbf{y}\|$. Therefore, as $\Delta\mathbf{y} \to \mathbf{0}$, we also have $\Delta\mathbf{v} \to \mathbf{0}$ and

$$\frac{\|\mathbf{R}(\Delta\mathbf{y})\|}{\|\Delta\mathbf{y}\|} \le K\frac{\|\mathbf{R}(\Delta\mathbf{y})\|}{\|\Delta\mathbf{v}\|} \to 0.$$

(Lemma 4.2, p. 133 makes essentially the same point.) This proves that $\mathbf{R}(\Delta\mathbf{y}) = \boldsymbol{o}(\Delta\mathbf{y})$ and thus that $\mathbf{f}^{-1}$ is differentiable at $\mathbf{q}$.

Step 13

The last fact to prove is that the derivative $\mathrm{d}(\mathbf{f}^{-1})_{\mathbf{q}}$ depends continuously on $\mathbf{q}$. But $\mathrm{d}(\mathbf{f}^{-1})_{\mathbf{q}} = (\mathrm{d}\mathbf{f}_{\mathbf{f}^{-1}(\mathbf{q})})^{-1} = (\mathrm{d}\mathbf{f}_{\mathbf{p}})^{-1}$; therefore we can use the following chain of argument.

- The entries of the $n \times n$ matrix $(\mathrm{d}\mathbf{f}_{\mathbf{p}})^{-1}$ are polynomial functions of the entries of $\mathrm{d}\mathbf{f}_{\mathbf{p}}$, and hence depend continuously on them.
- $\mathrm{d}\mathbf{f}_{\mathbf{p}}$ depends continuously on $\mathbf{p}$.
- $\mathbf{p} = \mathbf{f}^{-1}(\mathbf{q})$ depends continuously on $\mathbf{q}$.

This completes the proof of the inverse function theorem. □

Corollary 5.3 *Suppose* $\mathbf{f} : U^n \to \mathbb{R}^n$ *satisfies the conditions of the inverse function theorem at the point* $\mathbf{a}$ *in* U^n. *Then the image* $\mathbf{f}(U^n)$ *contains a ball* $\widehat{B}$ *of positive radius centered at the point* $\mathbf{f}(\mathbf{a})$ *in the target of* $\mathbf{f}$.

Proof. The conditions of the inverse function theorem apply to the inverse map $\mathbf{f}^{-1}$ at $\mathbf{f}(\mathbf{a})$. For $\widehat{B}$ take the ball provided by the theorem. □

The role of *continuous* differentiability

The inverse function theorem assumes that the derivative is continuous. This hypothesis is invoked, for example, at the point in the proof where the mean value theorem of maps is used (Steps 6 and 7) to show that $\mathbf{g}$ was a contraction mapping. But is continuity necessary? Our proof needs it, but does the theorem itself? Can we find a better proof that dispenses with that hypothesis?

In fact, the hypothesis is indispensable: there are differentiable functions that have an invertible derivative at a point but are not themselves invertible on any open neighborhood of that point. One such example is

$$f(x) = \frac{x}{2} + \frac{x^2}{2}\sin\frac{\pi}{x} \quad \text{if } x \ne 0, \qquad f(0) = 0.$$

$f'(x)$ exists but is not continuous at $x = 0$

The function undergoes infinitely many oscillations near $x = 0$, but because the graph is squeezed between the parabolas $y = (x \pm x^2)/2$, it follows that $f'(0) = \frac{1}{2}$.

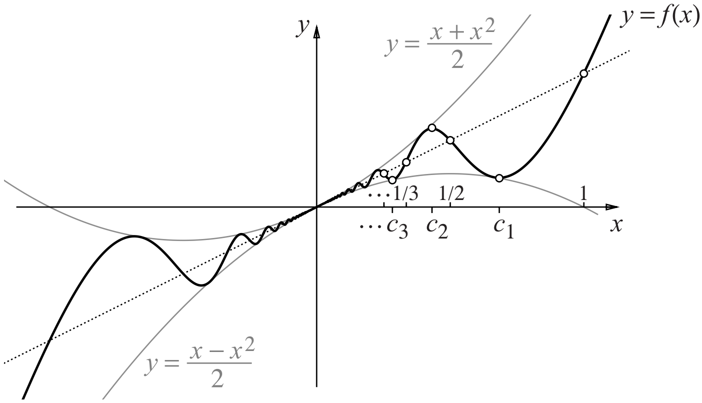
The derivative map $\mathrm{d}f_0(\Delta y) = \frac{1}{2}\Delta x$ is thus invertible. Now consider $f'(x)$ for $x \neq 0$; a direct computation gives

$$f'(x) = \frac{1}{2} + x\sin\frac{\pi}{x} - \frac{\pi}{2}\cos\frac{\pi}{x}.$$

At the points $1/n \to 0$, $f'(1/n)$ is alternately positive and negative. In particular, $f'(1/2n) = (1-\pi)/2$, so

$$\frac{1-\pi}{2} = \lim_{n\to\infty} f'\left(\frac{1}{2n}\right) \neq f'\left(\lim_{n\to\infty}\frac{1}{2n}\right) = f'(0) = \frac{1}{2}.$$

Thus, although f is differentiable everywhere, that derivative is not continuous at the origin.



However, f' is continuous away from the origin, so it must change sign at some point c_n between $1/(n+1)$ and $1/n$. That is, $f'(c_n) = 0$; in fact the c_n are alternately local maxima and minima of f. Because $1/n$ is an infinite sequence that converges to 0, the same must be true of the interlaced sequence c_n. Near each local extremum c_n, f is a 2–1 map. Inasmuch as $c_n \to 0$, there is no open interval around $x = 0$ on which f is 1–1. In other words, the oscillations make f noninvertible near the origin.

$f(x)$ is not 1–1 near $x = 0$

In our analysis of several examples of maps of the plane in Chapter 4.2, we found that a map usually "looked like" its derivative locally. When we returned to the quadratic map above (pp. 161–165), we actually converted the map into its derivative within a window by expressing the derivative in terms of appropriate curvilinear coordinates. The curvilinear coordinates were supplied by a map **h** for which

When will **f** "look like" $\mathrm{d}\mathbf{f_a}$?

$$\Delta\mathbf{f} = \mathrm{d}\mathbf{f_a} \circ \mathbf{h}.$$

We claimed that the new variables could indeed serve as coordinates; that is, we claimed, in effect, that **h** was invertible. Because $\mathrm{d}\mathbf{f_a}$ was obviously invertible, we defined **h** as

$$\mathbf{h} = (\mathrm{d}\mathbf{f_a})^{-1} \circ \Delta\mathbf{f}.$$

But this, in itself, does not prove $\mathbf{h}$ invertible. Now, however, we can settle the question: If $\mathbf{f}$ is continuously differentiable (and thus invertible in a neighborhood of $\mathbf{a}$, by the inverse function theorem), then $\mathbf{h}$ is invertible and

$$\mathbf{h}^{-1} = \Delta\mathbf{f}^{-1} \circ d\mathbf{f}_{\mathbf{a}}.$$

This discussion leads to the following corollary of the inverse function theorem.

Corollary 5.4 *If* $\mathbf{f} : U^n \to \mathbb{R}^n$ *is continuously differentiable on an open neighborhood* U^n *of* $\mathbf{a}$*, and* $d\mathbf{f}_{\mathbf{a}}$ *is invertible, then there is a coordinate change* $\mathbf{h} : V^n \to S^n$ *on some possibly smaller neighborhood* V^n *of* $\mathbf{a}$ *for which* $\Delta\mathbf{f} = d\mathbf{f}_{\mathbf{a}} \circ \mathbf{h}$. □

Comparison with Taylor's theorem

Before leaving the inverse function theorem, let us compare it with Taylor's theorem as a tool for understanding the geometric action of a map. Suppose we use Taylor's theorem to expand the map $\mathbf{f} : U^n \to \mathbb{R}^n$ at a point $\mathbf{a}$:

$$\Delta\mathbf{x} = \Delta\mathbf{f}(\Delta\mathbf{u}) = d\mathbf{f}_{\mathbf{a}}(\Delta\mathbf{u}) + \boldsymbol{O}(2).$$

This equation was the basis for our frequent observation, in Chapter 4, that the derivative $d\mathbf{f}_{\mathbf{a}}$ approximates $\mathbf{f}$ in a sufficiently small window centered at $\mathbf{a}$. The equation gives only an approximation because the difference $\boldsymbol{O}(2)$ between $\Delta\mathbf{f}$ and $d\mathbf{f}_{\mathbf{a}}$ is nonzero, in general. But the approximation is a good one in the sense that the difference vanishes like $\|\Delta\mathbf{u}\|^2$ as $\Delta\mathbf{u} \to \mathbf{0}$.

By contrast, suppose $d\mathbf{f}_{\mathbf{a}}$ is invertible. The inverse function theorem then says that new coordinates $\Delta\mathbf{s} = \mathbf{h}(\Delta\mathbf{u})$ can be found so that

$$\Delta\mathbf{x} = d\mathbf{f}_{\mathbf{a}}(\Delta\mathbf{s}).$$

In these circumstances, $d\mathbf{f}_{\mathbf{a}}$ equals $\Delta\mathbf{f}$ in a sufficiently small window (at least when $\Delta\mathbf{f}$ is expressed in terms of the proper curvilinear coordinates); the remainder $\boldsymbol{O}(2)$ is dispensed with. There are some minor technical differences, too. For Taylor's theorem, the components of $\mathbf{f}$ must have continuous second derivatives; for the inverse function theorem, only continuous first derivatives are needed.

The Taylor approximation goes a long way toward clarifying the action of $\mathbf{f}$; however, the inverse function theorem provides the ultimate simplification: it shows that $\mathbf{f}$ is essentially linear near $\mathbf{a}$. Perhaps most significantly, the inverse function theorem gives us a new tool to analyze maps: curvilinear coordinates and, more generally, alternative coordinate systems.

Exercises

5.1. Show that $\operatorname{arcsinh} y = \ln\left(y + \sqrt{y^2+1}\right)$; use this, the pullback substitution $y = \sinh x$, and other properties of hyperbolic functions to show

$$\int \frac{dy}{\sqrt{1+y^2}} = \ln\left(y + \sqrt{y^2+1}\right).$$

5.2. a. Show that $\operatorname{arctanh} y = \frac{1}{2}\ln\left(\dfrac{1+y}{1-y}\right)$.

b. Use the pullback $y = \tanh x$ (*not* partial fractions) to determine $\displaystyle\int \frac{dy}{1-y^2}$.

5.3. Use $y = \sinh x$ to show $\displaystyle\int \frac{dy}{(1+y^2)^{3/2}} = \frac{y}{\sqrt{1+y^2}}$.

5.4. Use suitable hyperbolic pullbacks to determine

$$\int \frac{dy}{(y^2-1)^{3/2}} \quad\text{and}\quad \int \frac{dy}{(1-y^2)^{3/2}}.$$

5.5. a. Sketch the graphs of $y = \operatorname{sech} x = 1/\cosh x$ and its inverse $x = \operatorname{arcsech} y$.

b. Show that $\operatorname{arcsech} y = \ln\left(\dfrac{1 \pm \sqrt{1-y^2}}{y}\right)$.

c. Use both the graph and the formula for $x = \operatorname{arcsech} y$ to explain why its domain is $0 < y \leq 1$.

d. Show that the two halves of the graph of $x = \operatorname{arcsech} y$ are negatives of each other by showing

$$-\ln\left(\frac{1+\sqrt{1-y^2}}{y}\right) = \ln\left(\frac{1-\sqrt{1-y^2}}{y}\right).$$

5.6. a. Sketch the graph of $x = y - \ln(-y) + k$, $y < 0$; use $k = 2$. Indicate the position of the “landmark” point $(x,y) = (-1+k,-1)$ on the graph. Determine the limiting value of x as $y \to -\infty$, and check that your sketch reflects this fact.

b. Sketch the inverse $y = f_{k,3}(x)$ of the function in part a. What are the domain and the range of $f_{k,3}$?

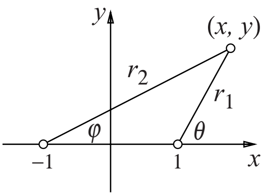
c. Show that the differential equation $y' = y/(y-1)$ has yet another solution $y = f_4(x) \equiv 0$.

d. Make a sketch of the (x,y)-plane that indicates there is precisely one solution of the differential equation $y' = y/(y-1)$ through each point (x,y) in which $y \neq 1$. This sketch exhibits, visually, the general solution to the differential equation.

5.7. Find the general solution of the differential equation

$$\frac{dy}{dx} = \frac{y}{y^2+1}.$$

Describe the solution in words and make a sketch that reflects its salient features.



5.8. a. Solve the following equations for x and y.

$$\theta = \arctan\left(\frac{y}{x-1}\right), \quad \varphi = \arctan\left(\frac{y}{x+1}\right).$$

Note: θ and φ are called the **biangular coordinates** for the plane.

b. Compute the Jacobians

$$\frac{\partial(x,y)}{\partial(\theta,\varphi)} \text{ and } \frac{\partial(\theta,\varphi)}{\partial(x,y)},$$

and show by direct computation that they are reciprocals.

5.9. a. Solve the following equations for x and y:

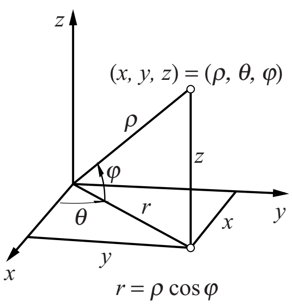
$$r_1 = \sqrt{(x-1)^2+y^2}, \quad r_2 = \sqrt{(x+1)^2+y^2}.$$

Note: r_1 and r_2 are called the **two-center bipolar coordinates** for the plane.

b. Compute the Jacobians

$$\frac{\partial(x,y)}{\partial(r_1,r_2)} \text{ and } \frac{\partial(r_1,r_2)}{\partial(x,y)},$$

and verify directly that they are reciprocals.



5.10. The following equations express Cartesian coordinates for space in terms of **spherical coordinates** ρ, θ, and φ (cf. Exercise 3.27, p. 104):

$$x = \rho\cos\theta\cos\varphi,$$
$$y = \rho\sin\theta\cos\varphi,$$
$$z = \rho\sin\varphi.$$

a. Solve the equations for the spherical coordinates ρ, θ, φ.

b. Compute the Jacobians

$$\frac{\partial(x,y,z)}{\partial(\rho,\theta,\varphi)} \text{ and } \frac{\partial(\rho,\theta,\varphi)}{\partial(x,y,z)},$$

and verify directly that they are reciprocals.

5.11. The following equations express Cartesian coordinates for space in terms of **cylindrical coordinates** r, θ, z:

$$x = r\cos\theta,$$
$$y = r\sin\theta,$$
$$z = z.$$

Show that

$$\frac{\partial(r,\theta,z)}{\partial(x,y,z)} = \frac{1}{\sqrt{x^2+y^2}} = \frac{1}{r},$$

and explain why this is already evident from your knowledge of polar coordinates in the plane.

5.12. a. Show that the formulas that express spherical coordinates (ρ,θ,φ) in terms of cylindrical coordinates *in the new order* (r,z,θ) are identical to the formulas that express cylindrical coordinates *in the same order* in terms of Cartesian coordinates:

$$\begin{aligned} \rho &= f(r,z,\theta), & r &= f(x,y,z),\\ \theta &= g(r,z,\theta), & z &= g(x,y,z),\\ \varphi &= h(r,z,\theta), & \theta &= h(x,y,z). \end{aligned}$$

It will be sufficient to determine f, g, and h.

b. Determine

$$\frac{\partial(\rho,\theta,\varphi)}{\partial(r,\theta,z)} \text{ and } \frac{\partial(r,z,\theta)}{\partial(x,y,z)}$$

and verify directly that

$$\frac{\partial(\rho,\theta,\varphi)}{\partial(x,y,z)} = \frac{\partial(\rho,\theta,\varphi)}{\partial(r,\theta,z)}\frac{\partial(r,\theta,z)}{\partial(x,y,z)}.$$

5.13. Use the Babylonian algorithm to determine $\sqrt{10}$ to 12 decimal places accuracy. Take $x_0 = 3$; how many iterations were required? Take $x_0 = 10$; how many iterations are required now?

5.14. a. To solve $x^2 = a$, the Babylonian algorithm first rewrites the equation as $x = a/x$ and then finds iterates of the average of the left- and right-hand sides: $g(x) = (x+a/x)/2$. This suggests solving $x^3 = a$ by iterating on the average of x and a/x^2: $g_1(x) = (x+a/x^2)/2$. Compute $\sqrt[3]{10}$ by finding the fixed point of g_1 with $x_0 = 2$. Convergence is relatively slow; how many iterates were needed to get 8 decimal places accuracy?

b. Convergence can be sped up by iterating on a weighted average of x and a/x^2: $g_2(x) = (2x+a/x^2)/3$. Compute $\sqrt[3]{10}$ by finding the fixed point of g_2 with $x_0 = 2$. How many iterates were needed to get 8 decimal places accuracy? How does this compare with convergence using g_1?

c. Devise an effective algorithm for solving $x^4 = a$ (that is not just the original Babylonian algorithm applied to the pair $y^2 = a$, $x^2 = y$). Use your algorithm to find $\sqrt[4]{120}$.

5.15. Use the Newton–Raphson method to find (to 6 decimal places) the three real roots of $f(x) = x^3 - 3x + 1$. Sketch the graphs of $y = x^3 - 3x$ and $y = f(x)$ to get the approximate locations of the roots to serve as initial values x_0 for the three iterations.

5.16. Consider the linear map $P : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$P : \begin{cases} s = \frac{1}{2}x - \frac{\sqrt{3}}{2}y, \\ t = \frac{\sqrt{3}}{2}x + \frac{1}{2}y. \end{cases}$$

a. The map $P : (x,y) \mapsto (s,t)$ is rotation by angle α; what is the value of α?
b. In the (s,t)-plane, sketch the images of the x- and y-axes. According to your sketch, does rotation by α turn the positive s-axis to the image of the positive x-axis?
c. Now let P pull back the variables s and t to provide a second coordinate system in the (x,y)-plane. Describe how the new (s,t) coordinate grid is related to the original (x,y) grid. In particular, describe the position of the positive s-axis in relation to the positive x-axis.
d. If $R_\theta : (x,y) \mapsto (s,t)$ is rotation by the angle θ, how does R_θ pull back the (s,t) coordinate grid to the (x,y)-plane? In particular, describe where the positive s-axis appears in this pullback.

5.17. This exercise uses the semi-log map **sl** (cf. page 161) and the fact that **sl** transforms the exponential function $y = Ba^{kx}$ into the linear function

$$Y = (k \log_{10} a)x + \log_{10} B.$$

a. Plot the US population census data for 1790–1900 on semi-log graph paper and verify that the points lie approximately on a straight line L. Let the horizontal coordinate x be years since 1790.
b. Estimate the slope and Y-intercept of the line L.
c. Use the estimates to obtain an exponential function $y = B10^{kx}$ that approximates the US cenus values, where x denotes years since 1790.

5.18. Define the log–log map (cf. page 161) $\mathbf{L} : (x,y) \to (X,Y)$ by

$$\mathbf{L} : \begin{cases} X = \log_{10} x, \\ Y = \log_{10} y; \end{cases} \qquad x,y > 0.$$

a. Show that the image of the graph of $y = ax^p$ under the map $\mathbf{L}$ is a straight line. Determine the equation of this line.
b. Using log–log paper in which the coordinates of the lower left hand corner are $(X,Y) = (0.1, 0.1)$, plot the graphs of $Y = (2/3)X + 4$ and $Y = -2X + 1$. Sketch the pullbacks of these graphs in the (x,y)-plane (using an ordinary uniform coordinate grid).

5.19. This exercise concerns the quadratic map $\mathbf{f}$ of Example 3 (p. 161) and the local coordinate change $\mathbf{h}$ that factors the window map $\Delta\mathbf{f}$ (cf. p. 163). Write $\mathbf{h}$ as $\mathbf{h_a}$ to reflect the fact that this coordinate change depends on the point $\mathbf{a} = (a,b)$ in the (u,v)-plane at which $\Delta\mathbf{f}$ is constructed, and then write the input $(\Delta u, \Delta v)$ of $\mathbf{h_a}$ more simply as $\mathbf{p} = (p,q)$.

a. Verify that $\mathbf{f}(\mathbf{u}) = \mathbf{f}(-\mathbf{u})$ for every $\mathbf{u}$.
b. Verify that $\mathbf{h_a}(\mathbf{0}) = \mathbf{0}$, $\mathbf{h_a}(-\mathbf{a}) = -\frac{1}{2}\mathbf{a}$, and $\mathbf{h_a}(-2\mathbf{a}) = \mathbf{0}$.
c. Show that $\mathbf{h_a}$ fails to be 1–1 on any neighborhood of $-\mathbf{a}$ by showing that $\mathbf{h_a}(-\mathbf{a}(1+\varepsilon)) = \mathbf{h_a}(-\mathbf{a}(1-\varepsilon))$ for any ε.
d. As noted in the text, $\mathbf{h_a}$ is invertible on any sufficiently small square window W centered at $\mathbf{p} = \mathbf{0}$. Give an upper bound on the length of the side of W.
e. Show that $\mathbf{h_a}$ is conformal everywhere inside W (part d) by showing the derivative $d(\mathbf{h_a})_\mathbf{p}$ is a dilation–rotation matrix (or similarity transformation, p. 118) for each point $\mathbf{p}$ in W.
f. Show that the dilation factor of $d(\mathbf{h_a})_\mathbf{p}$ is

$$\frac{(a+p)^2 + (b+q)^2}{a^2+b^2},$$

and conclude that $d(\mathbf{h_a})_{-\mathbf{a}}$ is the zero linear map.
g. Determine the rotation angle θ of $d(\mathbf{h_a})_\mathbf{p}$ in terms of $\mathbf{a}$ and $\mathbf{p}$, and deduce that $\theta > 0$ when $\mathbf{p} = (p,q)$ is above the line $q = (b/a)p$ and $\theta < 0$ below it. Confirm this fact in the figure on page 164 that illustrates the action of of $\mathbf{h}_{(\sqrt{3}/2,1/2)}$.

5.20. Let W be the infinite strip $-\pi/2 \le x \le \pi/2$ in the (x,y)-plane; let $\mathbf{s} : W \to \mathbb{R}^2$ be the map

$$\mathbf{s} : \begin{cases} u = \sin x \cosh y, \\ v = \cos x \sinh y. \end{cases}$$

a. Determine the derivative $d\mathbf{s}_{(x,y)}$. Determine the Jacobian $J(x,y)$ and show that $J > 0$ everywhere except at the two points $(x,y) = (\pm\pi/2, 0)$.
b. Show that the map $\mathbf{s}$ is conformal (cf. Exercise 5.19) everywhere except at the two points $(x,y) = (\pm\pi/2, 0)$.
c. Show that the image of the horizontal line segment $y = b$ is the upper half of the ellipse

$$\left(\frac{u}{\cosh b}\right)^2 + \left(\frac{v}{\sinh b}\right)^2 = 1$$

if $b > 0$, and the lower half if $b < 0$. What happens if $b = 0$?
d. Show that the image of the vertical line $x = a$ is the right branch of the hyperbola

$$\left(\frac{u}{\sin a}\right)^2 - \left(\frac{v}{\cos a}\right)^2 = 1$$

if $0 < a < \pi/2$ and the left branch if $-\pi/2 < a < 0$. What happens if $a = 0$, $-\pi/2$, or $\pi/2$?

e. Conclude that $\mathbf{s}$ is invertible on $W \setminus (\pm\pi/2, 0)$ (i.e., W with the two points $(\pm\pi/2, 0)$ removed), and thus defines a coordinate change there.

f. The coordinate change $\mathbf{s}$ puts curvilinear (x,y) coordinates on the (u,v)-plane. Sketch that coordinate grid in the square $|u| \le 2$, $|v| \le 2$. How does this grid manifest the conformality of the map $\mathbf{s}$?

g. Sketch the same curvilinear (x,y)-grid on the larger square where $|u| \le 20$, $|v| \le 20$. On this square, the grid should look like the polar coordinate grid; does it? Is conformality still evident?

5.21. Let U be the right half-plane $u > 0$, and let $\mathbf{h} : U \to \mathbb{R}^2$ be the map

$$\mathbf{h} : \begin{cases} x = u\mathrm{e}^{-v}, \\ y = u\mathrm{e}^{v}. \end{cases}$$

a. Show that the image $\mathbf{h}(U)$ is the first quadrant $Q : x > 0, y > 0$.

b. Find the inverse $\mathbf{h}^{-1}$ on Q to show that $\mathbf{h}$ is a coordinate change.

c. Determine the Jacobians $\partial(x,y)/\partial(u,v)$ and $\partial(u,v)/(\partial(x,y)$, and show that they are reciprocals.

d. Sketch the curvilinear (u,v)-coordinate grid on Q, and the curvilinear (x,y)-grid on U. Is the map $\mathbf{h}$ conformal?

5.22. Consider the map $\mathbf{m} : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$\mathbf{m} : \begin{cases} p = \mathrm{e}^{s} \cosh t, \\ q = \mathrm{e}^{s} \sinh t. \end{cases}$$

a. Determine the image $M = \mathbf{m}(\mathbb{R}^2)$ in the (p,q)-plane, and sketch the curvilinear (s,t)-coordinate grid there.

b. Determine the inverse $\mathbf{m}^{-1}$ on M to show that $\mathbf{m}$ is a coordinate change. Sketch the curvilinear (p,q)-grid in the (s,t)-plane.

c. Determine the Jacobians $\partial(p,q)/\partial(s,t)$ and $\partial(s,t)/(\partial(p,q)$, and show that they are reciprocals.

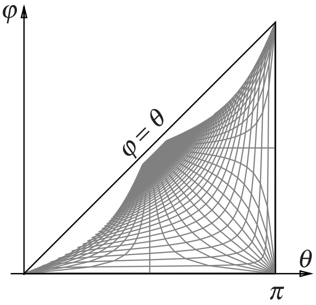
d. You should notice similarities between the maps $\mathbf{h}$ and $\mathbf{m}$ of the previous exercise and this one. Show that the coordinate changes

$$\mathbf{u} : \begin{cases} u = \mathrm{e}^{s}, \\ v = t; \end{cases} \qquad \mathbf{p} : \begin{cases} p = (y + x)/2, \\ q = (y - x)/2; \end{cases}$$

convert $\mathbf{h}$ into $\mathbf{m}$. That is, show $\mathbf{m} = \mathbf{p} \circ \mathbf{h} \circ \mathbf{u}$, and sketch all these maps together.

5.23. Let U be the upper half-plane $y > 0$ and let $\mathbf{a} : U \to \mathbb{R}^2 : (x,y) \to (\theta, \varphi)$ be the map defined by the biangular coordinates (cf. Exercise 5.8):

$$\mathbf{a} : \begin{cases} \theta = \arctan y/(x-1), \\ \varphi = \arctan y/(x+1). \end{cases}$$



a. Show that the image $\mathbf{a}(U)$ is the triangular region

$$T : \begin{array}{l} 0 < \theta < \pi, \\ 0 < \varphi < \theta. \end{array}$$

Conclude that $\mathbf{a} : U \to T$ is a coordinate change.

b. Identify, within T, the images of the lines $x = \pm 1$. Also indicate the limit of $\mathbf{a}(x,y)$ as $y \to 0$ and (i) $x < -1$; (ii) $-1 < x < 1$; and (iii) $1 < x$.

c. The curvilinear (x,y)-coordinate grid in T is shown in the margin. Identify the grid lines $x =$ const. and $y =$ const., and indicate how x and y vary through the grid. Indicate how the grid illustrates the limits you determined in the previous part.

d. Referring to the curvilinear (x,y)-coordinate grid, indicate the geometric action of $\mathbf{a}^{-1}$ on T. That is, indicate how $\mathbf{a}^{-1}$ "opens up" T to become the upper half-plane. Does $\mathbf{a}^{-1}$ reverse orientation? How is the answer to this question indicated in the geometric action?

e. Draw the (θ, φ)-coordinate grid in the (x,y)-plane. (This is, in fact, easy to do; do you see why?)

5.24. Let U be the upper half-plane $y > 0$ and let $\mathbf{b} : U \to \mathbb{R}^2 : (x,y) \to (r_1, r_2)$ be the map defined by the two-center bipolar coordinates (cf. Exercise 5.9):

$$\mathbf{b} : \begin{cases} r_1 = \sqrt{(x-1)^2 + y^2}, \\ r_2 = \sqrt{(x+1)^2 + y^2}. \end{cases}$$

a. Show that r_1 and r_2 satisfy the inequalities

$$2 < r_1 + r_2, \quad -2 < r_2 - r_1 < 2.$$

This defines a "half-infinite" strip S in the (r_1, r_2)-plane; $\mathbf{b}(U) = S$.

b. Explain why the map $\mathbf{b} : U \to S$ is a coordinate change. It follows that the (image of the) Cartesian (x,y)-grid defines curvilinear coordinates in S. Sketch this curvilinear grid. Are the curvilinear grid lines perpendicular?

c. Sketch, in U, the curvilinear coordinate grid defined by r_1 and r_2. (This is easy to do.)

d. The map $\mathbf{b}$ is well defined on the x-axis. Sketch the image of the x-axis in the (r_1, r_2)-plane; note in particular the images of the points $(\pm 1, 0)$. How is the image related to S?

5.25. The following map $\sigma : U^4 \to \mathbb{R}^4 : (r, \mathbf{t}) \to \mathbf{x}$ defines the analogue of spherical coordinates on $\mathbb{R}^4$:

$$\sigma : \begin{cases} x_1 = r\cos t_1\, \cos t_2\, \cos t_3, \\ x_2 = r\sin t_1\, \cos t_2\, \cos t_3, \\ x_3 = \qquad\quad r\,\sin t_2\, \cos t_3, \\ x_4 = \qquad\qquad\qquad r\,\sin t_3; \end{cases} \qquad U^4 : \begin{array}{c} 0 < r, \\ -\pi < t_1 < \pi, \\ -\pi/2 < t_2 < \pi/2, \\ -\pi/2 < t_3 < \pi/2. \end{array}$$

a. Describe the image $V^4 = \sigma(U^4)$.

b. Obtain the derivative $d\mathbf{s}_{(r,\mathbf{t})}$ and show that $\det(d\mathbf{s}_{(r,\mathbf{t})}) = r^3 \cos t_2 \cos^2 t_3$.

c. Deduce that σ is locally invertible everywhere in V^4.

d. Find a formula for the (global) inverse of σ on V^4.