Chapter 9

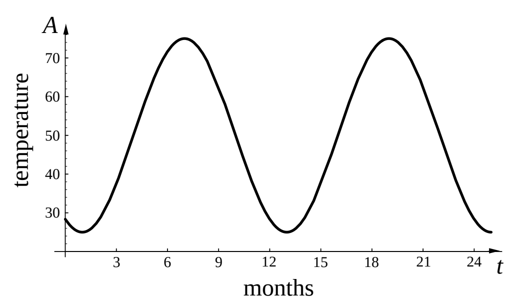
Functions of Several Variables

Functions that depend on several input variables first appeared in the S-I-R model at the beginning of the course. Usually, the number of variables has not been an issue for us. For instance, when we introduced the derivative in chapter 3, we used partial derivatives to treat functions of several variables in a parallel fashion. However, when there are questions of visualization and geometric understanding, the number of variables *does* matter. Every variable adds a dimension to the problem—one way or another. For example, if a function has two input variables instead of one, we will see that its graph is a surface rather than a curve.

The problem of visualizing a function of several variables

This chapter deals with the geometry of functions of two or more variables. We start with graphs and level sets. These are the basic tools for visualization. Then we turn to microscopic views, and see what form the microscope equation takes. Finally, we consider optimization problems using both direct visual methods and dynamical systems.

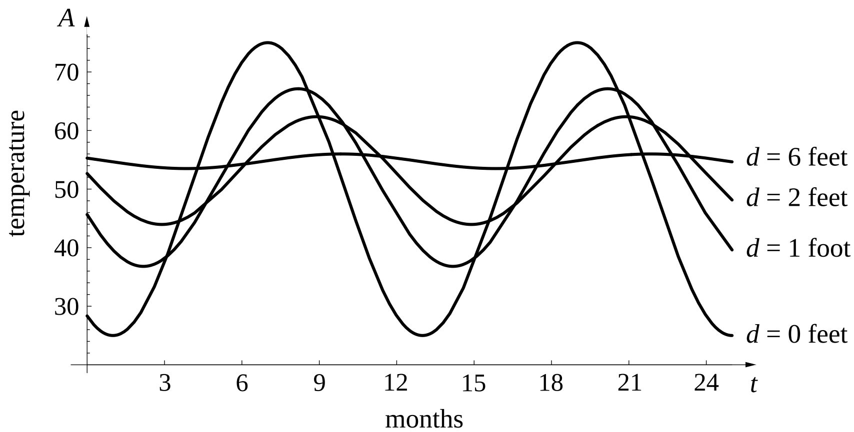
9.1 Graphs and Level Sets



The graph at the right comes from a model that describes how the average daily temperature at one place varies over the course of a year. It shows the temperature A in °F, and the time t in months from January. As we would expect, the temperature is a periodic function (which we can write as $A(t)$), and its period is 12 months. Furthermore, the lowest temperature occurs in February (when $t \approx 1$ or 13) and the highest in July (when $t \approx 7$ or 19). This is about what we would expect.

Underground temperatures fluctuate less

However, all these temperature fluctuations disappear a few feet underground. Below a depth of 6 or 8 feet, the temperature of the soil remains about 55°F year-round! Between ground level and that depth, the temperature still fluctuates, but the range from low to high decreases with the depth. Here is what happens at some specific depths.



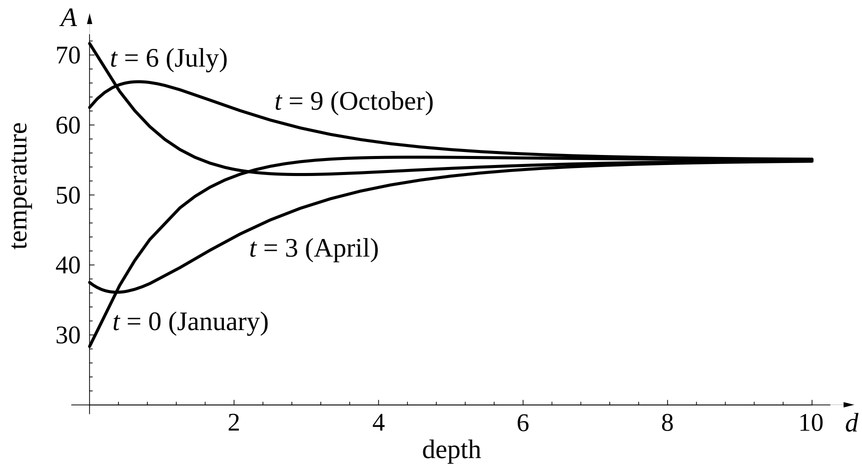
Notice how the time at which the temperature peaks gets later and later as we go farther and farther underground. For example, at $d = 2$ feet the highest temperature occurs in September ($t \approx 9$), not July. It literally takes time for the heat to sink in. In chapter 7 we called this a phase shift. The lowest temperature shifts in just the same way. At a depth of 2 feet, it is colder in March than in January.

The phase shifts with the depth

Thus A is really a function of *two* variables, the depth d as well as the time t. To reflect this addition, let's change our notation for the function to $A(t, d)$. In the figure above, d plays the role of a parameter: it has a fixed value for each graph. We can reverse these roles and make t the parameter. This is done in the figure on the top of the next page. It shows us how the temperature varies with the depth at fixed times of year. Notice that, in April and October, the extreme temperature is not found on the surface. In October, for example, the soil is warmest at a depth of about 9 inches.

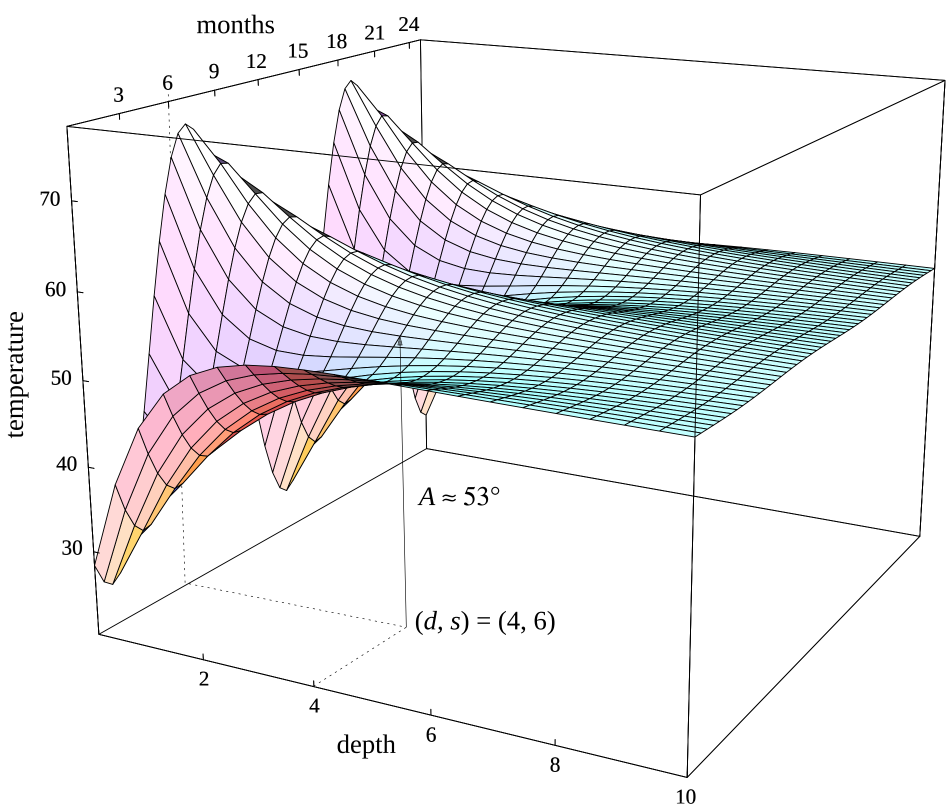
Graphing temperature as a function of depth

The lower figure on the page is a single graph that combines all the information in these two sets of graphs. Each point on the bottom of the box of the box corresponds to a particular depth and a particular time. The height of the surface above that point tells us the temperature at that depth and time. For example, suppose you want to find the temperature 4 feet



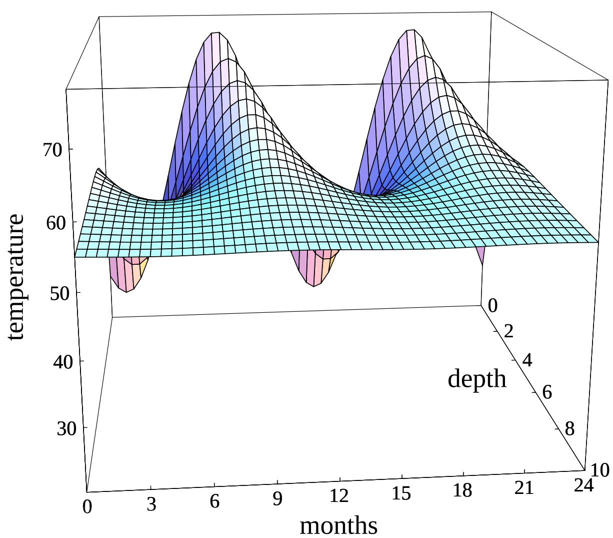
below surface at the beginning of July. Working from the bottom front corner of the box, move 4 feet to the right and then 6 months toward the back. This is the point $(d, s) = (4, 6)$. The height of the graph above this point is the temperature A that we want.

Reading a surface graph



Grid lines are slices of the surface that show how the function depends on each variable separately

There is a definite connection between this surface and the two collections of curves. Imagine that the box containing the surface graph is a loaf of bread. If you slice the loaf parallel to the left or right side, this slice is taken at a fixed depth. The cut face of the slice will look like one of the graphs on page 512. These show how the temperature depends on the time at fixed depths. If you slice the loaf the other way—parallel to the front or back face—then the time is fixed. The cut face will look like one of the graphs on page 513. They show how the temperature depends on the depth at fixed times. The grid lines on the surface are precisely these "slice" marks.

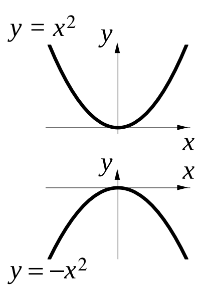


Here is the same graph seen from a different viewpoint. Now time is measured from the left, while depth is measured from the back. The temperature is still the height, though. One advantage of this view is that it shows more clearly how the peak temperature is phase-shifted with the depth.

We now have two ways to visualize how the average daily temperature depends on the time of year and the depth below ground. One is the surface graph itself, and the other is a collection of curves that are slices of the surface. The surface gives us an overall view, but it is not so easy to read the surface graph to determine the temperature at a specific time and depth. Check this yourself: what is the temperature 2 feet underground at the beginning of April? The slices are much more helpful here. You should be able to read from either collection of slices that $A \approx 44°$F.

Comparing the surface to its slices

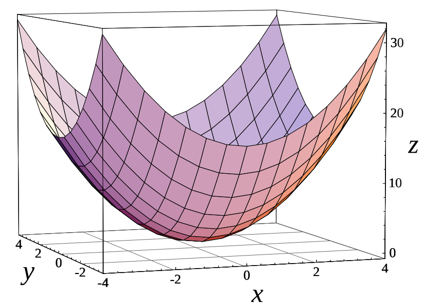
Examples of Graphs



The purpose of this section is to get some experience constructing and interpreting surface graphs. To work in a context, look first at the functions $y = x^2$ and $y = -x^2$. They provide us with standard examples of a minimum and a maximum when there is just one input variable. Let's consider now the corresponding examples for two input variables. Besides an ordinary maximum and an ordinary minimum, we will find a *third* type—called a minimax—that is completely new. It arises because a function can have a minimum with respect to one of its input variables and a maximum with respect to the other.

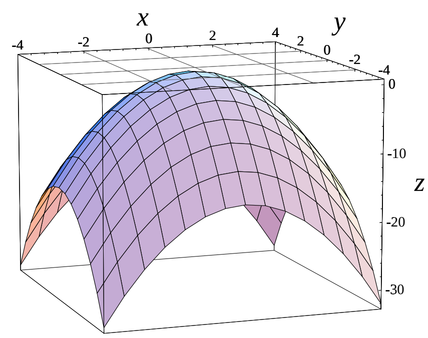
A minimum: $z = x^2 + y^2$

At the origin $(x, y) = (0, 0)$, $z = 0$. At any other point, either x or y is non-zero. Its square is positive, so $z > 0$. Consequently, z has a minimum at the origin. The graph of this function is a parabolic **bowl** whose lowest point sits on the origin. As always, the grid lines are slices, made by fixing the value of x or y. For example, if $y = c$, then the slice is $z = x^2 + c^2$. This is an ordinary parabolic curve.



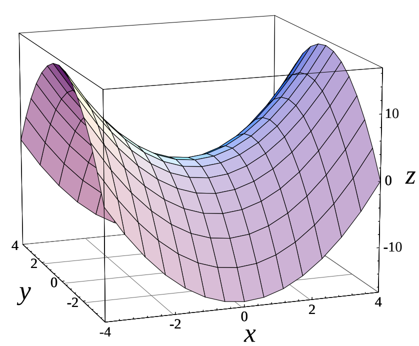
A maximum: $z = -x^2 - y^2$

For any x and y, the value of z in this example is the opposite of its value in the previous one. Thus, z is everywhere negative, except at the origin, where its value is 0. Thus z has a maximum at the origin. Its graph is an upside-down bowl, or **peak**, whose highest point reaches up and touches the origin. Grid lines are the curves $z = -x^2 - c^2$ and $z = -c^2 - y^2$. These are parabolic curves that open downward.

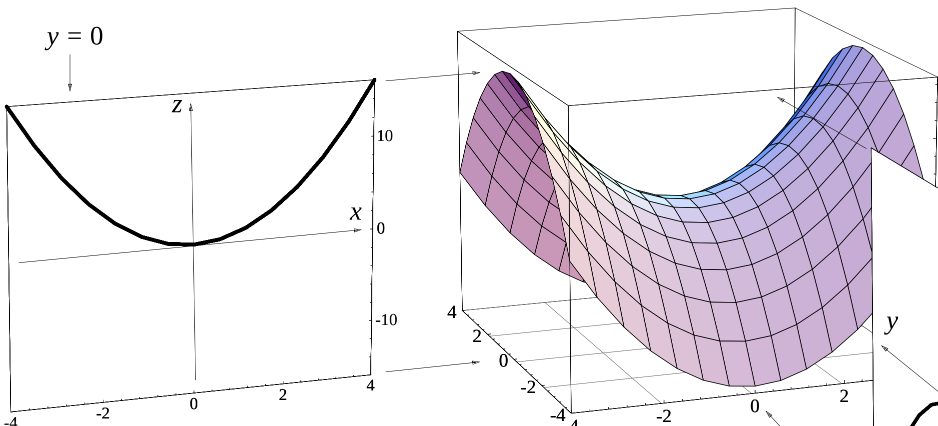


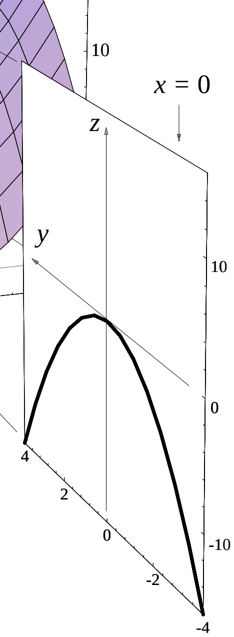
A minimax: $z = x^2 - y^2$

Suppose we fix y at $y = 0$. This slice has the equation $z = x^2$, so it is an ordinary parabola (that opens upward). Thus, as far as the input x is concerned, z has a *minimum* at the origin. Suppose, instead, that we fix x at $x = 0$. Then we get a slice whose equation is $z = -y^2$. It is also an ordinary parabola, but this one opens downward. As far as y is concerned, z has a *maximum* at the origin. It is clear from the graph how upward-opening slices in the x-direction fit together with downward-opening slices in the y-direction. Because of the shape of the surface, a minimax is commonly called a **saddle**, or a **saddle point**.



Here are two slices of $z = x^2 - y^2$ shown in more detail. Points in the box have three coordinates: (x, y, z). If we set $y = 0$ we are selecting the points





of the form $(x, 0, z)$. These make up the x, z-plane. On this plane the equation $z = x^2 - y^2$ becomes simply $z = x^2$. The graph of *this* equation is a curve in the x, z-plane—specifically, the parabola shown. The situation is similar if y is given some other fixed value. For example, $y = -4$ specifies the points $(x, -4, z)$. These describe the plane that forms the front face of the box. The equation $z = x^2 - y^2$ becomes $z = x^2 - 16$. The curve tracing out the intersection of the saddle with the front of the box is precisely the graph of $z = x^2 - 16$.

The points where $y = c$ form a vertical plane parallel to the x, z-plane

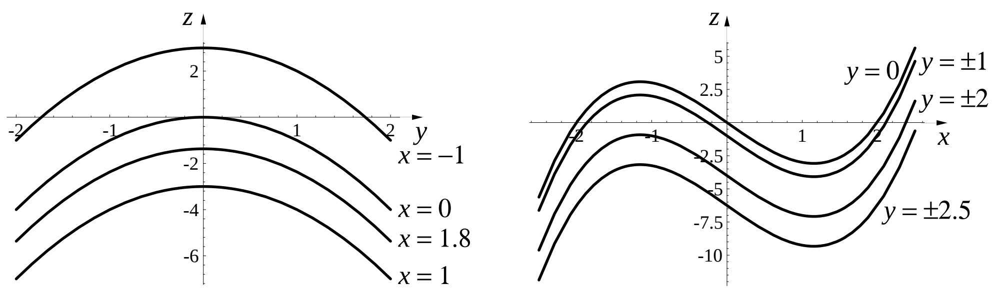
If $x = 0$ we get the points $(0, y, z)$ that make up the y, z-plane. On this plane the equation simplifies to $z = -y^2$, and its graph is the parabolic curve shown. Giving x a different fixed value leads to similar results. A good example is $x = -4$. The points $(-4, y, z)$ lie on the plane that forms the left side of the box. The equation becomes $z = 16 - y^2$ there, and this is the parabolic curve marking the intersection of the saddle with the left side of the box.

The points where $x = c$ form a vertical plane parallel to the y, z-plane

As you can see, it is valuable for you to be able to generate surface graphs yourself. There are now a number of computer utilities which will do the job. Some can even rotate the surface while you watch, or give you a stereo view. However, even without one of these powerful utilities, you should try to generate the slicing curves that make up the grid lines of the surface.

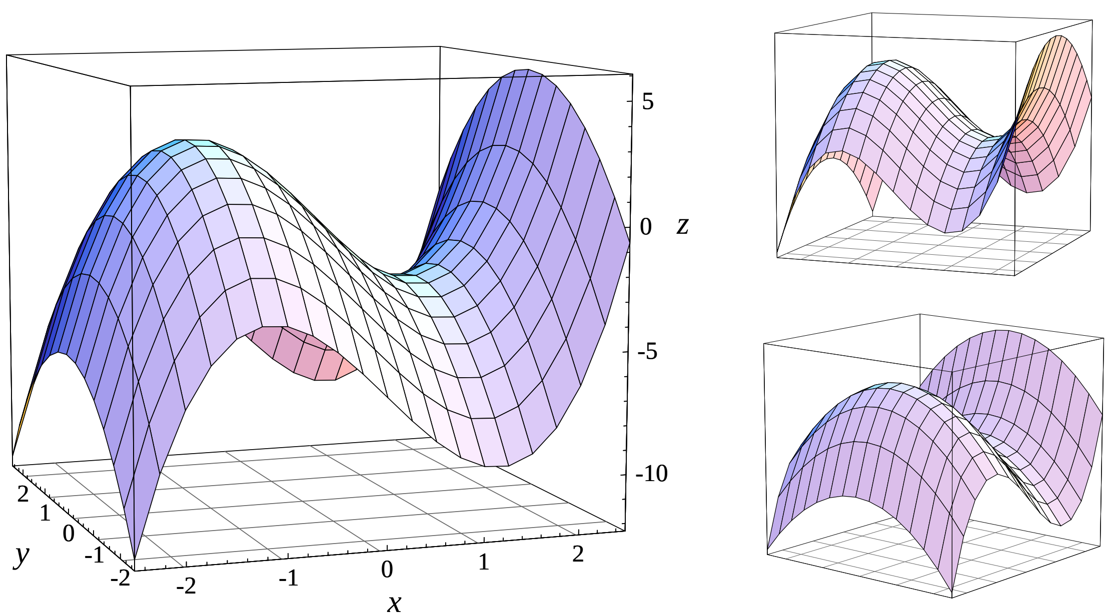
A cubic: $z = x^3 - 4x - y^2$

Slices of this graph are downward-opening parabolas (when $x = c$) and are cubic curves that have the same shape (when $y = c$). Notice that each cubic curve has a maximum and a minimum, and each parabola has



a maximum. The surface graph itself has a *peak* where the cubics have their maximum, but it has a *saddle* where the cubics have a minimum. Do you see why? The saddle point is a minimax for $z = x^3 - 4x - y^2$: z has a minimum there *as a function of* x *alone* but a maximum *as a function of* y *alone*.

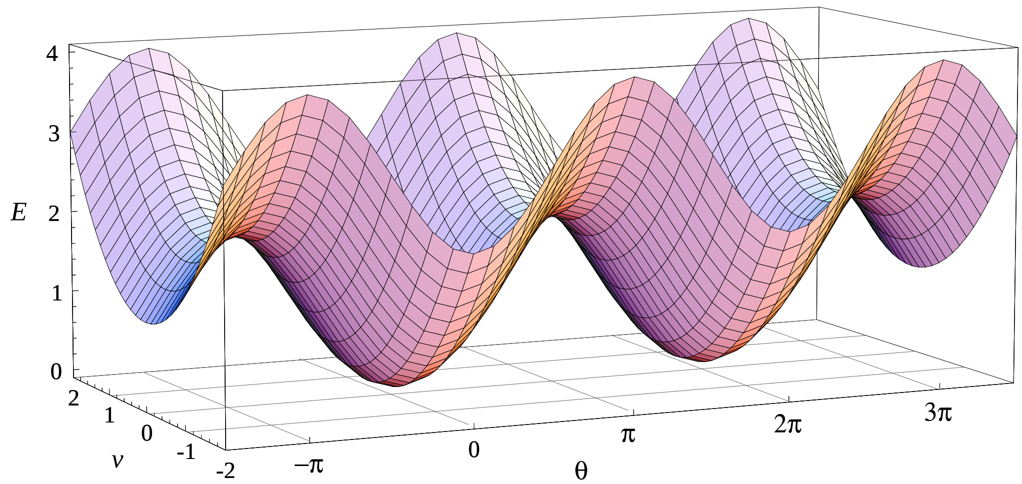
The surface has a saddle

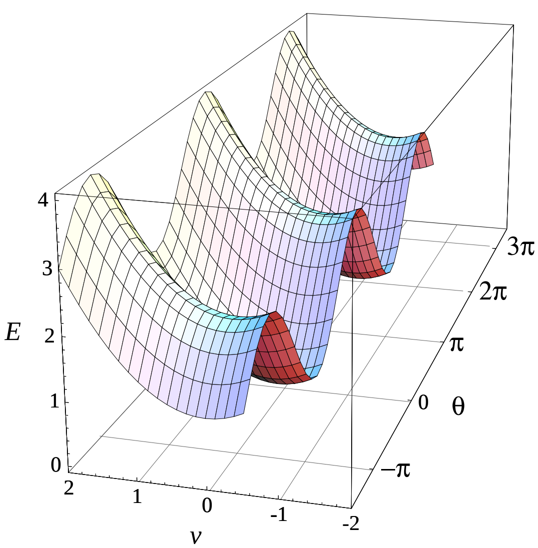


The small figures on the right show the same surface as the large figure; they just show it from different viewpoints. As a practical matter, you should look at these surfaces the way you would look at sculpture: “walk around them” by generating diverse views.

See the graph from different viewpoints

Energy of the pendulum: $E = 1 - \cos\theta + \frac{1}{2}v^2$



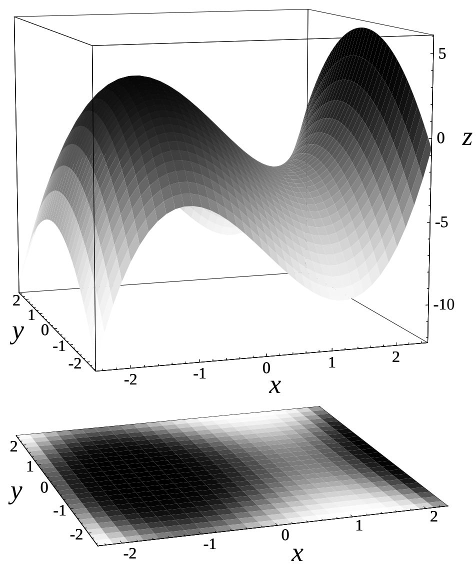


This function first came up in chapter 7, where it was used to demonstrate that a dynamical system describing the motion of a frictionless pendulum had periodic solutions. It was used again in chapter 8 to clarify the phase portrait of that dynamical system. The function E varies periodically with θ, and you can see this in the graph. The minimum at the origin is repeated at $(\theta, v) = (2\pi, 0)$, and so on. The graph also has a saddle at the point $(\theta, v) = (-\pi, 0)$. This too repeats with period 2π in the θ direction.

The figure at the left is the same surface with part cut away by a slice of the form $v = c$. These slices are sine curves: $E = 1 - \cos\theta + \frac{1}{2}c^2$. Slices of the form $\theta = c$ are upward-opening parabolas. From this viewpoint, the saddle points show up clearly.

One way to describe what happens to a real pendulum—that is, one governed by frictional forces as well as gravity—is to say that its energy "runs down" over time. Now, at any moment the pendulum's energy is a point on this graph. As the energy runs down, that point must work its way down the graph. Ultimately, it must reach the bottom of the graph—the minimum energy point at the origin $(\theta, v) = (0, 0)$. This is the stable equilibrium point. The pendulum hangs straight down ($\theta = 0$) and is motionless ($v = 0$). The graph gives us an abstract—but still vivid and concrete—way of thinking of the dissipation of energy.

From Graphs to Levels



There is still another way to picture a function of two variables. To see how it works we can start with an ordinary graph. On the right is the graph of

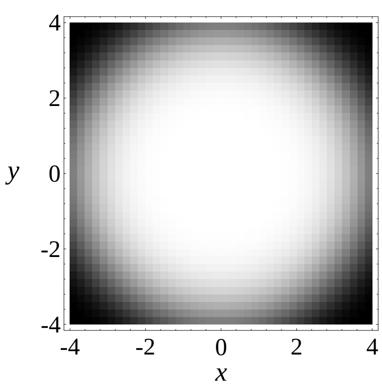
$$z = f(x, y) = x^3 - 4x - y^2,$$

the cubic function we considered on page 517. This graph looks different, though. The difference is that points are shaded according to their height. Points at the bottom are lightest, points at the top are darkest.

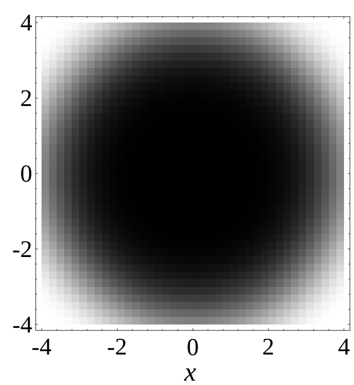
Notice that the flat x, y-plane is shaded exactly like the graph above it. For instance, the dark spot centered at the point $(x, y) = (-1, 0)$ is directly under the peak on the graph. The other dark patch, near the right edge of the plane, is under the highest visible part of the surface. Consequently, the shading on the x, y-plane gives us the same information as the graph. In other words, *the intensity of shading at* (x, y) *is proportional to the value of the function* $f(x, y)$.

Density plots

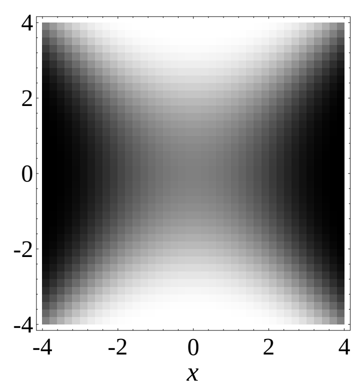
The figure in the x, y-plane is called a **density plot**. Think of the intensity of shading as the *density* of ink on the page. Here are density plots of the standard minimum, maximum, and minimax. Compare these with the



$z = x^2 + y^2$



$z = -x^2 - y^2$



$z = x^2 - y^2$

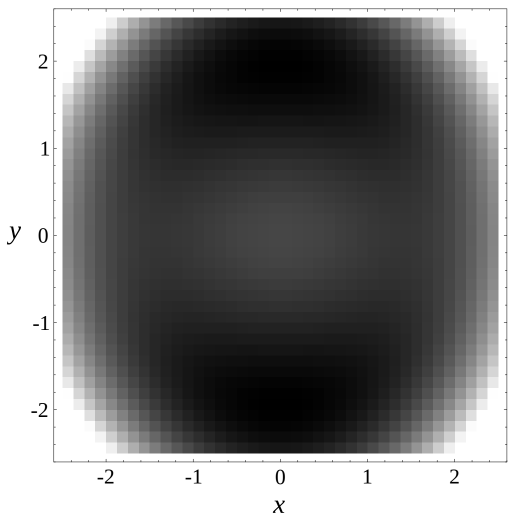
graphs on page 515. The third density plot is the most interesting. From the center of the x, y-plane, the shading increases to the right and left. Therefore,

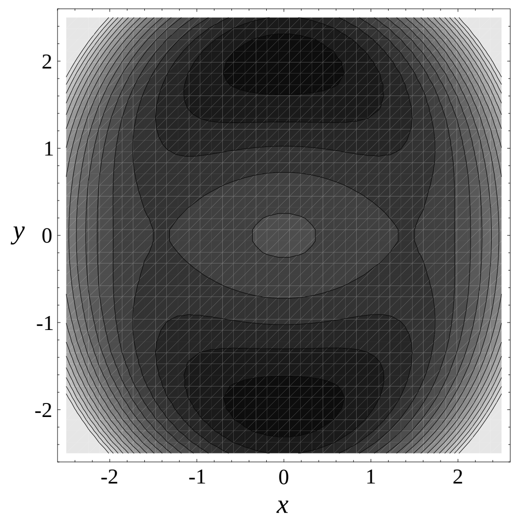
z has a minimum in the horizontal direction. However, the shading *decreases* above and below the center. Therefore, z has a *maximum* in the vertical direction. Thus, you really can see there is a minimax at the origin.

A sample plot

Try your hand at reading the density plot on the left below. You should see two maxima (directly above and below the origin), a minimum (at the origin itself), and two saddles (to the right and the left of the origin). The function defining the plot is

$$f(x, y) = (x^2 + (y-1)^2 - 3)(3 - x^2 - (y+1)^2).$$





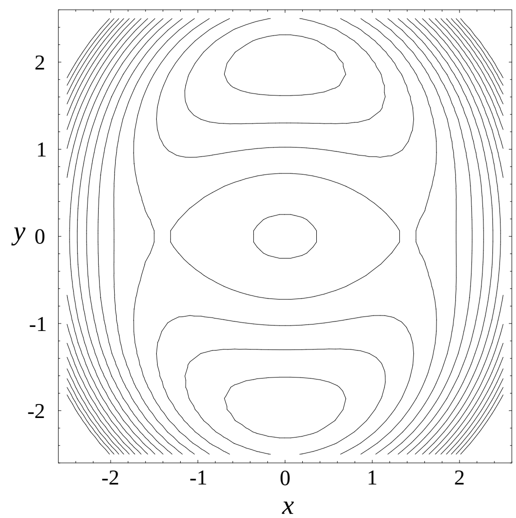
Can you visualize what the graph looks like? This density plot should help you, and you can also construct slices by setting $x = c$ and $y = c$. The slices $x = 0$ and $y = 0$ are especially useful. With them you could determine the exact coordinates of the maxima and the saddles.

These density plots show a "checkerboard" pattern because *Mathematica* (the computer program that produces them) shades each little square according to the value of the function at the center of the square. This pattern is an artefact; it is not inherent to a density plot.

From densities to contours

In a density plot, the shading varies smoothly with the value of the function. This is accurate, but it may be a bit difficult to read. On the right you see a modified density plot. There is still shading, but there are now just a few distinct shades. This makes a sharp boundary between one shade and the next. The boundary is called a **contour**, or a **level**. The figure itself is called a **contour plot**. The two maxima on the vertical line $x = 0$ stand out more clearly on the contour plot. Also, the contour lines around the two saddles help us see that the function has a minimum in the vertical direction and a maximum in the horizontal direction.

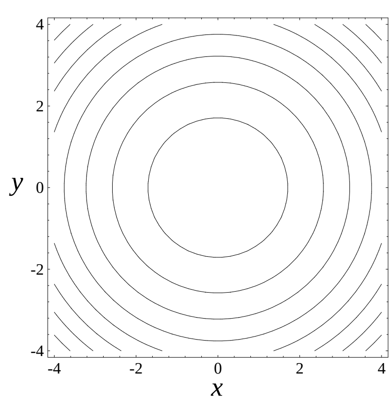
Once we have contour lines to separate one density level from the next, we can even dispense with the shading. The figure on the right is just the contour plot from the opposite page, minus the shading. The contour lines, or level curves, now stand out clearly. On each contour, the value of the function is constant. This is also called a **contour plot**.



There is some loss of information here, however. For example, we can't tell where the value of the function is large and where it is small. Nevertheless, the nested ovals on the vertical line $x = 0$ *do* tell us that there is either a maximum or a minimum at the center of each nest.

Contours of the standard functions

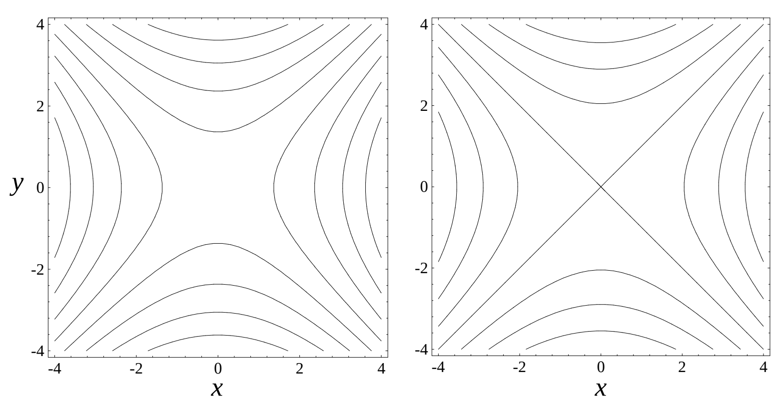
For reference purposes, here are the contour plots for the standard minimum, maximum, and saddle. In the first two cases, the contours are concentric ovals. These look the same, so only one is illustrated. The other two pictures show a saddle. In general, the contours around a saddle are a family of hyperbolas. However, it is possible for one of the contour lines to pass exactly through the minimax point. That contour is a pair of crossed lines, as shown in the version on the right. You should compare these contour plots with the density plots of the same functions on page 519, and with their graphs on page 515.



two functions but one plot

$z = x^2 + y^2$

$z = -x^2 - y^2$

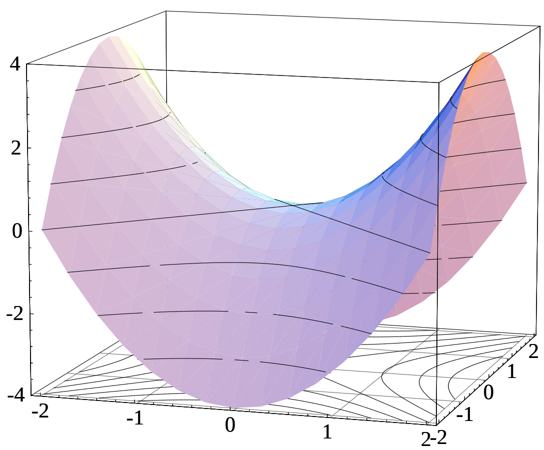


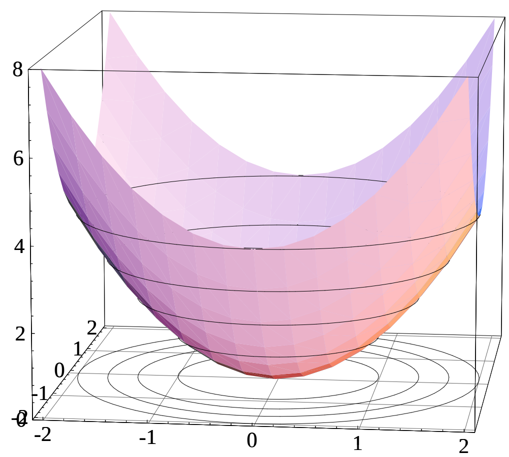
two plots of a single function

$z = x^2 - y^2$

Contours are horizontal slices of a graph

There is a direct connection between the contour plot of a function and its graph. Contours are horizontal slices of the graph, just as grid lines are vertical slices. Below, we use the standard functions $z = x^2 - y^2$ and $z = x^2 + y^2$ to illustrate the connection. Notice that every contour down in the x, y-plane lies exactly below, and has the same shape as, a horizontal slice of the graph. This picture explains why contours are called *level* curves.



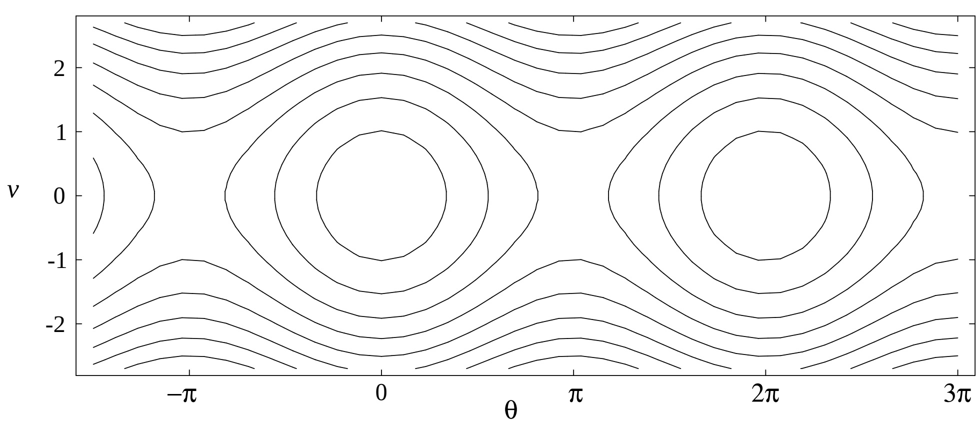


Energy of the pendulum, again

To get some more experience with contour plots, we return to the energy function of the pendulum:

$$E(\theta, v) = 1 - \cos\theta + \tfrac{1}{2}v^2.$$

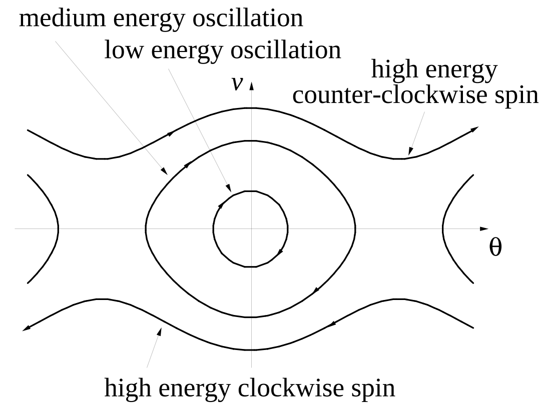
From the contour plot alone you should be able to see that E has either a minimum or a maximum at $(\theta, v) = (0, 0)$, and another at $(2\pi, 0)$. The contours also provide evidence that there is a saddle (minimax) near $(\theta, v) = (-\pi, 0)$ and $(\pi, 0)$. It is also apparent that E is a *periodic* function of θ.



What you should find most striking about this plot, however, is the way it resembles the phase portrait of the pendulum (chapter 8, pages 471–474). Every level curve here looks like a trajectory of the dynamical system. This is no accident. We know from chapter 8 that the energy is a first integral for the dynamics. In other words, energy is constant along each trajectory—this is the law of conservation of energy. But each level curve shows where the energy function has some fixed value. Therefore, each trajectory must lie on a single energy level.

Energy contours are trajectories of the dynamics

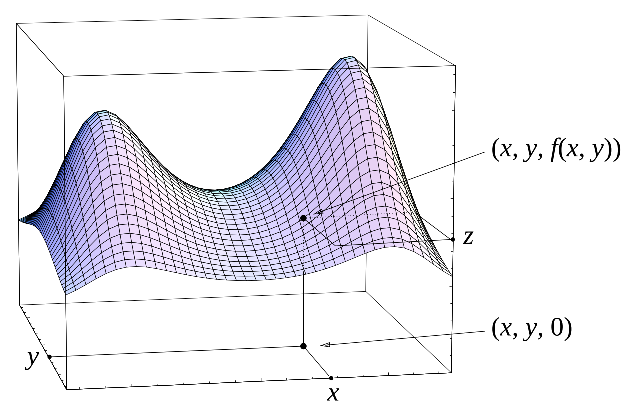
We can carry the connection between contours and trajectories even further. Closed trajectories correspond to *oscillations* of the pendulum. But the closed trajectories are the closed contours, and these are the ones that surround the minimum. In particular, they are *low* energy levels. By contrast, at higher energies ($E > 2$, in fact), the pendulum will just continue to spin in what ever direction it was moving initially. Thus, each high energy level is occupied by *two* trajectories—one for clockwise spinning and one for counter-clockwise.

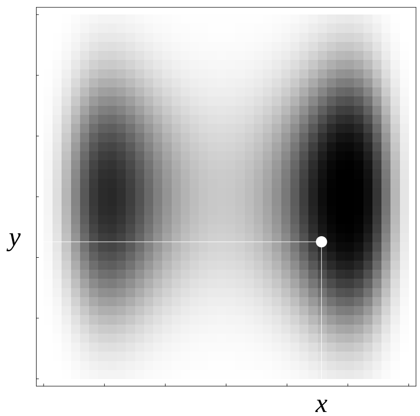


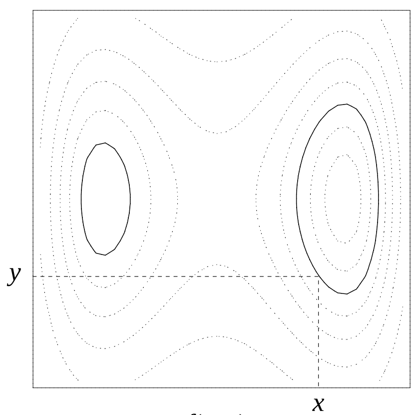
Technical Summary

The examples we have seen so far were meant to introduce some of the common ways of visualizing a function $z = f(x, y)$. To use them most effectively, though, you need to know more precisely how each is defined. We review here the definition of a graph, a density plot, a contour plot, and a terraced density plot.

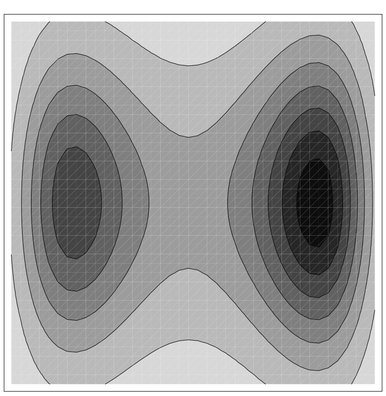
Graph. The graph of $z = f(x, y)$ lies in the 3-dimensional space with coordinates (x, y, z). To construct it, take any input (x, y). Identify this with the point $(x, y, 0)$ in the x, y-plane (which is defined by the condition $z = 0$). The corresponding point on the graph lies at the height $z = f(x, y)$ above the x, y-plane. This point has coordinates $(x, y, f(x, y))$. **The graph is the set of all points of the form $(x, y, f(x, y))$.** This is a 2-dimensional surface.







$f(x, y) = c$



Density plot. The density plot of $z = f(x, y)$ lies in the 2-dimensional x, y-plane. Choose any rectangle where the function is defined, and let m and M be the minimum and maximum values, respectively, of $f(x, y)$ on the rectangle. Define

$$\rho(x, y) = \frac{f(x, y) - m}{M - m}.$$

Then ρ satisfies $0 \le \rho(x, y) \le 1$ on the rectangle; it is called a **density function** (ρ is the Greek letter *rho*). **In the density plot, the density of ink—or darkness—at (x, y) is $\rho(x, y)$.**

Contour plot. A **contour** of $z = f(x, y)$ is the set of points in the x, y-plane where f has some fixed value:

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

That fixed value c is called the **level** of the contour. (The two solid ovals in the figure at the left make up a single contour.) A contour is also called a level curve. **A contour plot of f is a collection of curves $f(x, y) = c_j$ in the x, y-plane.** In the plot it is customary to use constants $c_1, c_2, \ldots$ that are equally spaced; that is, the interval between one c_j and the next always has the same value Δc.

Terraced density plot. This is a contour plot in which the region between two adjacent contours is shaded with ink of a single density. If the contours are at levels c_1 and c_2, then the density that is typically chosen is the one for the level half-way between these two—that is, for their average $(c_1 + c_2)/2$. Each region is called a **terrace**. Often, a terraced density plot is drawn in color, using different colors for each terrace. Television weather programs use terraced density plots to describe the temperature forecast for a large region.

We find density plots everywhere. A photograph is a density plot of the light that fell on the film when it was exposed. A newspaper 'half-tone' illustration is also a density plot of an image.

The pros and cons

Each of these modes of visualization has advantages and disadvantages. All are reasonably good at indicating the extremes (the maxima and minima) of a function. A contour plot needs some additional information—for example, a label on each contour to indicate its level—to distinguish between maxima and minima. However, if you want to know the numerical value of $f(x, y)$ at a particular point (x, y), a contour plot with labels offers more precision than a density plot. It's usually better than a graph, too.

Plots are visually economical in comparison to graphs

Overall, a graph has the biggest visual impact, but there is a cost. It takes three dimensions to represent the graph of a function of two variables, but only two to represent a plot. The cost is that extra dimension. It means that we cannot draw the graph of a function of three variables. That would take four mutually perpendicular axes—an impossibility in our three-dimensional space. However, we can produce a contour plot.

Contours of a Function of Three Variables

Contours and levels

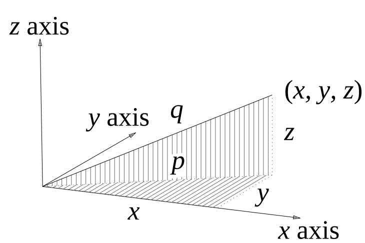
We pause here for a brief glimpse of a large subject. By analogy with the definition for a function of two variables, we say that a **contour** of the function $f(x, y, z)$ is the set of points (x, y, z) that satisfy the equation

$$f(x, y, z) = c,$$

for some fixed number c. We call c the **level** of the contour.

The standard minimum

Let's find the contours of $w = x^2 + y^2 + z^2$. This is completely analogous to the function $x^2 + y^2$ with two inputs. (What do the contours of $x^2 + y^2$ look like?) In particular, w has a minimum when $(x, y, z) = (0, 0, 0)$. As the following diagram shows, $w = x^2 + y^2 + z^2$ is the square of the distance from the origin to (x, y, z). (We use the Pythagorean theorem twice: once for p^2 and once for q^2.) Consequently, all points (x, y, z) where w has a fixed value



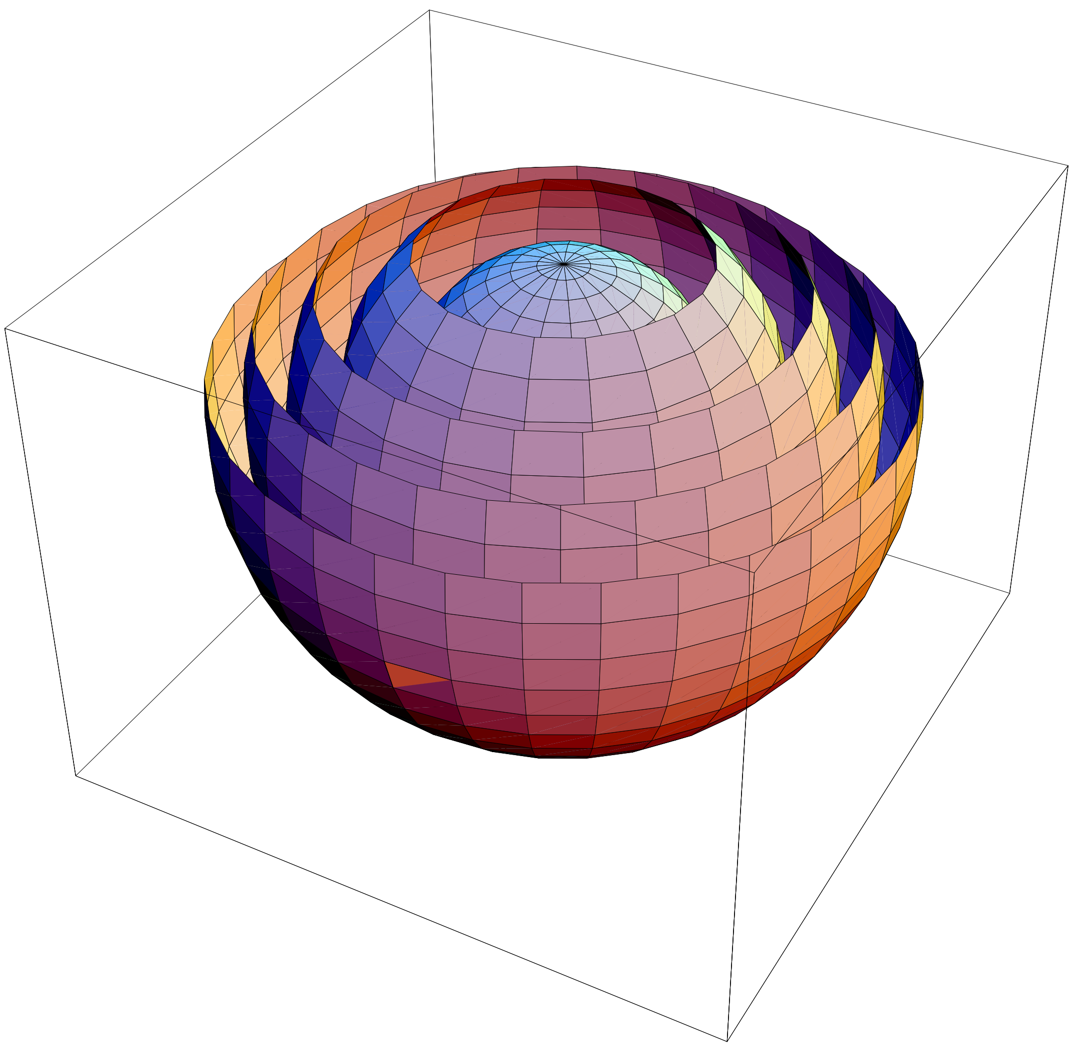
$$\begin{aligned} p^2 &= x^2 + y^2, \\ q^2 &= p^2 + z^2 \\ &= x^2 + y^2 + z^2 \\ &= w. \end{aligned}$$

lie a fixed distance from the origin. Specifically, $w = x^2 + y^2 + z^2 = c$ is the following set:

- $c > 0$: the sphere of radius $\sqrt{c}$ entered at the origin;
- $c = 0$: the origin itself;
- $c < 0$: the empty set.

The contours are spheres

The contour plot of $w = x^2 + y^2 + z^2$ is thus a nest of concentric spheres, as shown in the illustration below. The value of w is constant on each sphere. (The tops of the spheres have been cut away so you can see how the spheres nest; the whole thing resembles an onion.)



A standard minimax

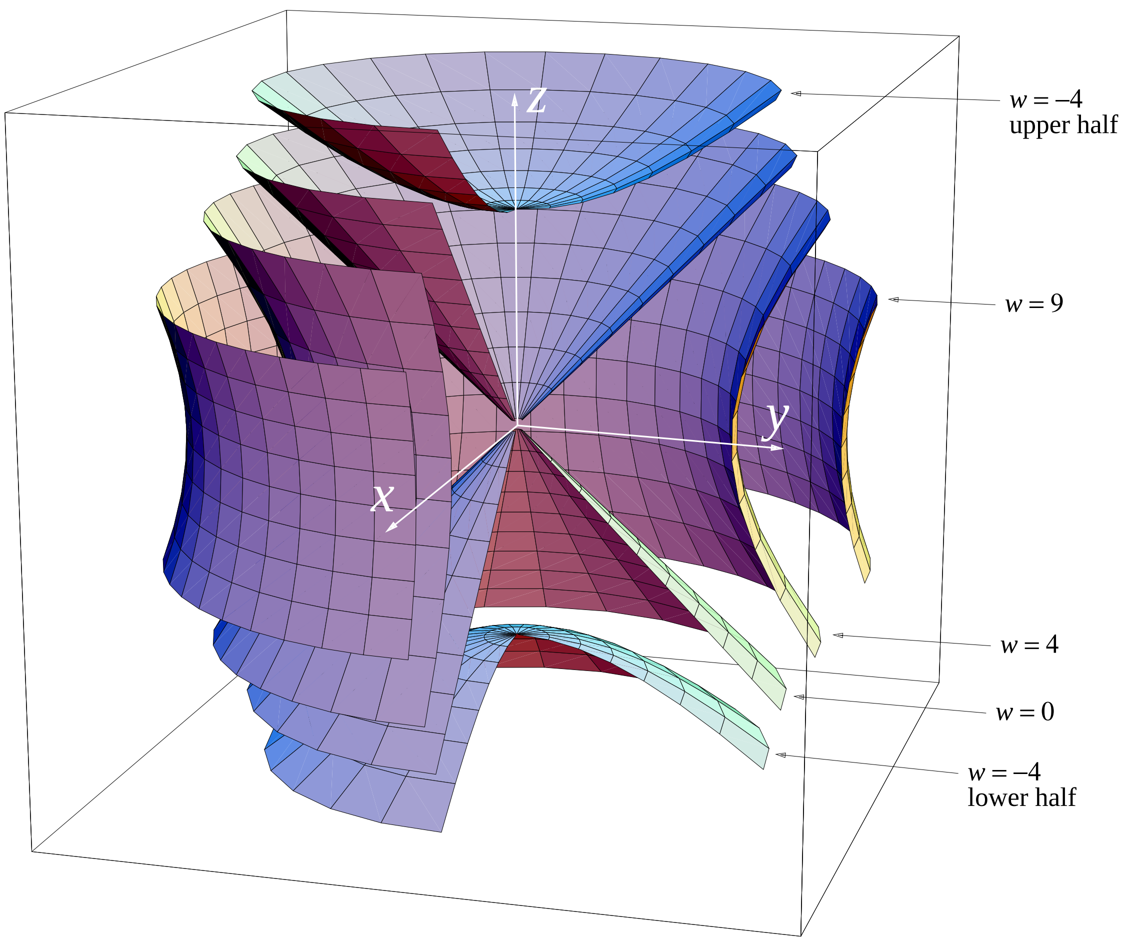
Below is the contour plot of another standard function with three input variables:

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

A quarter of each surface has been cut away so you can see how the surfaces nest together. Note that $w = 0$ is a cone, and every surface with $w < 0$ consists of two disconnected (but congruent) pieces—an upper half and a lower half.

The contours are hyperbolic shapes

You should compare this function to the standard minimax $x^2 - y^2$ in two variables. The three-variable function w has a minimum with respect to *both* of the variables x and y, while it has a maximum with respect to z. (Do you see why? The arguments are exactly the same as they were for two input variables on page 515.) Furthermore, the contours of $x^2 - y^2$ are a family of hyperbolas, and the contours of $x^2 + y^2 - z^2$ are surfaces obtained by rotating these hyperbolas about a common axis.



When there are three input variables, the contours are surfaces

It is a general fact—and our two examples provide good evidence for it—that a single contour of a function of three variables is a *surface*. Thus a contour is a curve or a surface, depending on the number of input variables. We often use the term **level set** (rather than a level *curve* or a level *surface*) as a generic name for a contour.

Exercises

In many of these exercises it will be essential to have a computer program to make graphs, terraced density plots, and contour plots of functions of two variables.

1. a) Use a computer to obtain a graph of the function $z = \sin x \sin y$ on the domain $0 \le x \le 2\pi$, $0 \le y \le 2\pi$. How many maximum points do you see? How many minimum points? How many saddles?

b) Determine, as well as you can from the graph, the location of the maximum, minimum, and saddle points.

2. Continuation. Make the domain $-2\pi \le x \le 2\pi$, $-2\pi \le y \le 2\pi$ and answer the same questions you did in the previous exercise. (Does the graph look like an egg carton?)

3. Obtain a terraced density plot (or a contour plot) of $z = \sin x \sin y$ on the domain $-2\pi \le x \le 2\pi$, $-2\pi \le y \le 2\pi$. Locate the maximum, minimum, and saddle points of the function. Do these results agree with those from the previous exercise?

4. Obtain the graph of $z = \sin x \cos y$ on the domain $0 \le x \le 2\pi$, $0 \le y \le 2\pi$. How does this graph differ from the one in exercise 1? In what ways is it similar?

5. Obtain the graph of $z = 2x + 4x^2 - x^4 - y^2$ when $-2 \le x \le 2$, $-4 \le y \le 4$. Locate all the minimum, maximum, and saddle points in this domain. [Note: the minimum is on the boundary!]

6. Continuation. Obtain a terraced density plot (or contour plot) for the function in the previous exercise, using the same domain. Use the plot to locate all the minimum, maximum, and saddle points. Compare your results with those of the previous exercise.

7. a) Obtain the graph of $z = 2x - y$ on the domain $-2 \leq x \leq 2$, $-2 \leq y \leq 2$. What is the shape of the graph?

b) Graph the same function of the domain $2 \leq x \leq 6$, $0 \leq y \leq 4$. What is the shape of the graph? How does this graph compare to the one in part (a)?

8. a) Continuation. Sketch three different slices of the graph of $z = 2x - y$ in the y-direction. What do the slices have in common? How are they different?

b) Answer the same questions for slices in the x-direction.

9. a) Obtain the graph of $z = .3x + .8y + 2.3$; choose the domain yourself. Where does the graph intercept the z-axis?

b) Describe the vertical slices of this graph in the y-direction and in the x-direction.

10. Describe the vertical slices of the graph of $z = px + qy + r$ in the y-direction and in the x-direction.

11. a) Compare the contours of the function $z = x^2 + 2y^2$ to those of $z = x^2 + y^2$.

b) What is the shape of the graph of $z = x^2 + 2y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

12. a) Compare the contours of the function $z = x^2 - 2y^2$ to those of $z = x^2 - y^2$.

b) What is the shape of the graph of $z = x^2 - 2y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

13. a) Obtain a contour plot of the function $z = x^2 + xy + y^2$.

b) What is the shape of the graph of $z = x^2 + xy + y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

14. a) Obtain a contour plot of the function $z = x^2 + 3xy + y^2$.

b) What is the shape of the graph of $z = x^2+3xy+y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

15. a) Obtain a contour plot of the function $z = x^2 + 2xy + y^2$.

b) What is the shape of the graph of $z = x^2+2xy+y^2$? Decide this first using only the information you have about the contours. Then use a computer to obtain the graph.

16. Complete this statement: The function $f(x, y; p) = x^2 + pxy + 4y^2$, which depends on the parameter p, has a minimum at the origin when _____ and a minimax when _____.

17. a) Obtain the graph and a terraced density plot of the function $z = 3x^2 + 17xy + 12y^2$. What is the shape of the graph?

b) What is the shape of the contours? Indicate how the contours fit on the graph.

18. a) Obtain the graph and a terraced density plot of the function $z = 3x^2 + 7xy + 12y^2$. What is the shape of the graph?

b) What is the shape of the contours? Indicate how the contours fit on the graph.

19. a) Obtain the graph and a terraced density plot of the function $z = 3x^2 + 12xy + 12y^2$. What is the shape of the graph?

b) What is the shape of the contours? Indicate how the contours fit on the graph.

20. Obtain the graph of $z = f(x, y) = xy$ on the domain $-3 \le x \le 3$, $-3 \le y \le 3$. Does this function have a maximum or a minimum or a saddle point? Where?

21. a) Continuation. Sketch slices of the graph of $z = xy$ in the y-direction, for each of the values $x = -2, -1, 0, 1$, and 2. What is the general shape of each of these slices?

b) Repeat part (a), but make the five slices in the x-direction—that is, fix y instead of x.

22. Continuation. Show how the slices you obtained in the previous exercise fit (or appear) on the graph you obtained in the exercise just before that one.

23. a) Continuation. Let $x = u + v$ and $y = u - v$. Express z in terms of u and v, using the fact that $z = xy$. Then obtain the graph of z as a function of the new variables u and v.

b) What is the shape of the graph you just obtained? Compare it to the graph of $z = xy$ you obtained earlier.

24. Can you draw a network of straight lines on the saddle surface $z = x^2 - y^2$?

25. Obtain a terraced density plot of $z = xy$. How do the contours of this plot fit on the graph of $z = xy$ you obtained in a previous exercise?

26. The graphs of $z = x^2 + 5xy + 10y^2$ and $z = 3$ intersect in a curve. What is the shape of that curve?

27. The graphs of $z = x^2 + y^3$ and $z = 0$ intersect in a curve. What is the shape of that curve?

28. The graphs of $z = 2x - y$ and $z = .3x + .8y + 2.3$ intersect in a curve. What is the shape of that curve?

First integrals

29. A hard spring described by the dynamical system

$$\frac{dx}{dt} = v, \qquad \frac{dv}{dt} = -cx - \beta x^3,$$

has a first integral of the form

$$E(x, v) = \tfrac{1}{2}cx^2 + \tfrac{1}{2}\beta x^4 + \tfrac{1}{2}v^2.$$

This is the **energy** of the spring. (See chapter 7.3, especially exercise 13, page 454.)

a) Let $c = 16$ and $\beta = 1$. Obtain the graph of $E(x, v)$ on a domain that has the origin at its center. Locate all the minimum, maximum, and saddle points in this domain.

b) What is the state of the spring (that is, its position x and its velocity v) when it has minimum energy?

30. A soft spring described by the dynamical system

$$\frac{dx}{dt} = v, \qquad \frac{dv}{dt} = \frac{-25x}{1+x^2},$$

has an energy integral of the form

$$E(x,v) = \tfrac{25}{2}\ln(1+x^2) + \tfrac{1}{2}v^2.$$

(See exercise 16, page 455.)

a) Obtain the graph of $E(x,v)$. Experiment with different possibilities for the domain until you get a good representation.

b) Obtain a terraced density plot of $E(x,v)$ over the same domain you chose in part (a). Compare the two representations of E.

c) Does the spring have a state of minimum energy? If so, where is it?

d) Does the spring have a state of *maximum* energy? Explain your answer.

31. a) **The Lotka–Volterra equations**. According to exercise 33 of chapter 7.3 (page 458), the function

$$E(x,y) = .1\ln y + .04\ln x - .005\,y - .004\,x$$

is a first integral of the dynamical system

$$\begin{aligned} x' &= .1x - .005\,xy, \\ y' &= .004\,xy - .04\,y. \end{aligned}$$

Obtain the graph of E on the domain $1 \le x \le 50$, $1 \le y \le 50$. (Why not enlarge the domain to $0 \le x \le 50$, $0 \le y \le 50$?)

b) Find all maximum, minimum, and saddle points on this graph. What is the connection between the maximum of E and the equilibrium point of the dynamical system?

c) Obtain a contour plot of E on the same domain as in part (a). Compare the contours of E and the trajectories of the dynamical system. (This reveals a *conservation of "energy"* for the solutions of the Lotka-Volterra equations.)

32. a) Continuation. Here is another first integral of the same dynamical system as in the previous exercise:

$$H(x, y) = \frac{x^{0.04} y^{0.1}}{\exp(.005\, y + .004\, x)}.$$

Obtain the graph of H and compare it to the graph of E in the previous exercise.

b) Obtain a contour plot of H, and compare the contours to the trajectories of the dynamical system.

9.2 Local Linearity

Local linearity is the central idea of chapter 3: it says that a graph looks straight when viewed under a microscope. Using this observation we were able to give a precise meaning to the *rate of change* of a function and, as a consequence, to see why Euler's method produces solutions to differential equations. At the time we concentrated on functions with a single input variable. In this section we explore local linearity for functions with two or more input variables.

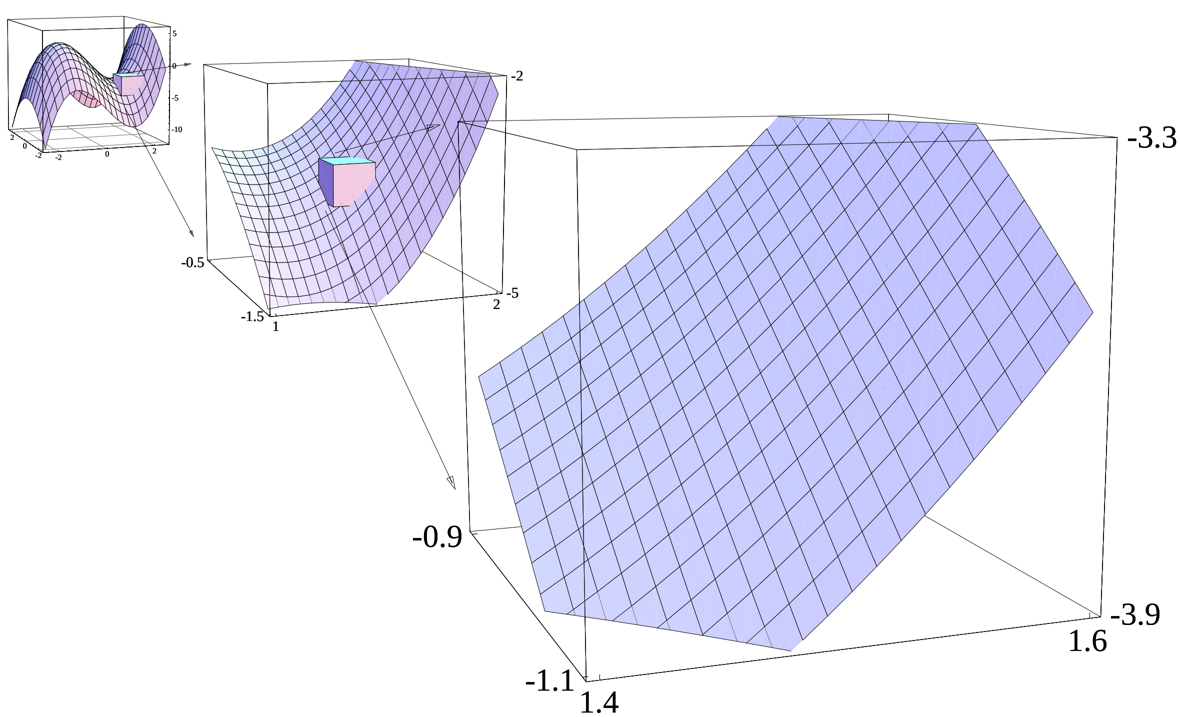
Microscopic Views

Consider the cubic $f(x,y) = x^3 - 4x - y^2$ that we used as an example in the previous section. We'll examine both the graph and the plot of f under a microscope. In the figure below we see successive magnifications of the graph near the point where $(x,y) = (1.5, -1)$. The initial graph, in the left rear, is drawn over the square

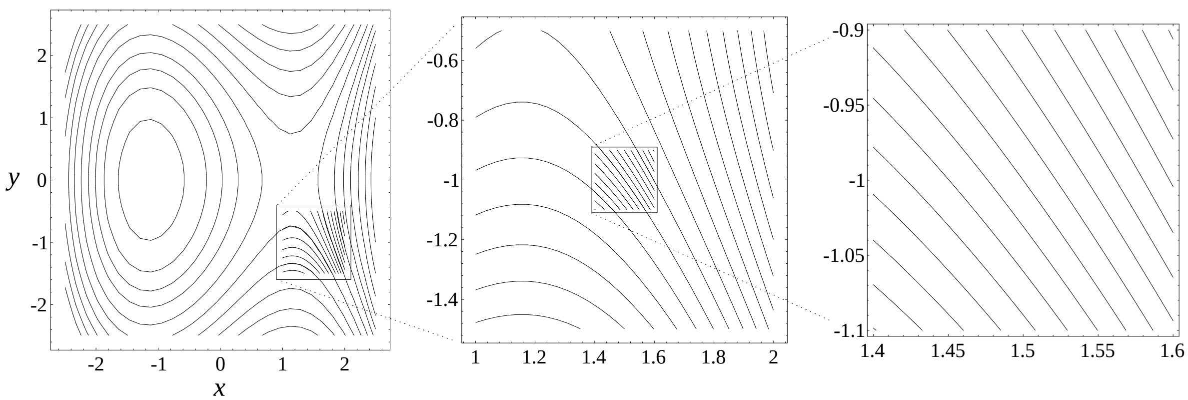
Magnifying a graph

$$-2.5 \leq x \leq 2.5, \qquad -2.5 \leq y \leq 2.5.$$

With each magnification, the portion of the surface we see bends less and less. **The graph approaches the shape of a flat plane**.



Contour plots for $f(x,y) = x^3 - 4x - y^2$ appear below. Again, we magnify near the point $(x,y) = (1.5, -1)$. Each window below is a small part of the window to its left. In the large scale plot, which is the first one on the left, the contours are quite variable in their direction and spacing. With each magnification, that variability decreases. **The contours become straight, parallel, and equally spaced**.



The process of magnification thus leads us to functions whose graphs are flat and whose contours are straight, parallel, and equally-spaced. As we shall now see, these are the linear functions.

Linear Functions

A linear function is defined by the way its output responds to *changes* in the input. Specifically, in chapter 1 we said

Responses to changes in input

$$y = f(x) \quad \text{is linear if} \quad \Delta y = m \cdot \Delta x.$$

This is the simplest possibility: changes in output are strictly proportional to changes in input. The multiplier m is both the *rate* at which y changes with respect to x and the *slope* of the graph of f.

The definition

Exactly same idea defines a linear function of two or more variables: the change in output is strictly proportional to the change in any one of the inputs.

Definition. The function $z = f(x_1, x_2, \ldots, x_n)$ is **linear** if there are multipliers $p_1, p_2, \ldots, p_n$ for which

$$\Delta z = p_1 \cdot \Delta x_1, \quad \Delta z = p_2 \cdot \Delta x_2, \quad \ldots, \quad \Delta z = p_n \cdot \Delta x_n.$$

There is one multiplier for each input variable. The multipliers are constants and they are, in general, all different.

Partial and total changes

The definition describes how z responds to each input separately. We call each $p_j \cdot \Delta x_j$ a **partial change**. The multiplier $p_j = \Delta z/\Delta x_j$ is the corresponding **partial rate of change**. Of course, several input variables may change simultaneously. In that case, the **total change** in z will just be the sum of the individual changes produced by the several variables:

$$\Delta z = p_1 \cdot \Delta x_1 + p_2 \cdot \Delta x_2 + \cdots + p_n \cdot \Delta x_n.$$

Another way to describe a linear function

Of course, if the total change of a function satisfies this condition, then each partial change has the form $p_j \cdot \Delta x_j$. (If only x_j changes, then all the other Δx_k must be 0. So Δz becomes simply $p_j \cdot \Delta x_j$.) Consequently, the function must be linear. In other words, we can use the formula for the total change as another way to define a linear function.

Alternate definition. The function $z = f(x_1, x_2, \ldots, x_n)$ is **linear** if there are multipliers $p_1, p_2, \ldots, p_n$ for which

$$\Delta z = p_1 \cdot \Delta x_1 + p_2 \cdot \Delta x_2 + \cdots + p_n \cdot \Delta x_n.$$

Formulas for linear functions

From the definition to a formula

When $z = f(x_1, x_2, \ldots, x_n)$ is a linear function, we know how Δz depends on the changes Δx_j, but that doesn't tell us explicitly how z itself is related to the input variables x_j. There are several ways to express this relation as a formula, depending on the nature of the information we have about the function. For the sake of clarity, we'll develop these formulas first for a function of two variables: $z = f(x, y)$.

Given the partial rates of change and an initial point

• **The initial-value form**. Suppose we know the value of a linear function at some given point—called the **initial point**—and we also know its partial rates of change. Can we construct a formula for the function? Suppose $z = z_0$ when $(x, y) = (x_0, y_0)$, and suppose the partial rates of change are

$$p = \frac{\Delta z}{\Delta x} \qquad \text{and} \qquad q = \frac{\Delta z}{\Delta y}.$$

If we let

$$\Delta x = x - x_0, \qquad \Delta y = y - y_0, \qquad \Delta z = z - z_0,$$

then we can write

$$\begin{aligned} z - z_0 = \Delta z &= p \cdot \Delta x + q \cdot \Delta y \\ &= p \cdot (x - x_0) + q \cdot (y - y_0). \end{aligned}$$

This is the initial-value form of a linear function. For example, if the initial point is $(x, y) = (4, 3)$, $z = 5$, and the partial rates of change are $\Delta z/\Delta x = -\frac{1}{2}$, $\Delta z/\Delta y = +1$, the equation of the linear function can be written

$$z - 5 = -\tfrac{1}{2}(x - 4) + (y - 3).$$

• **The intercept form**. This is a special case of the initial-value form, in which the initial point is the origin: $(x, y) = (0, 0)$, $z = r$. The formula becomes

Given the partial rates of change and the z-intercept

$$z - r = px + qy, \qquad \text{or} \qquad z = px + qy + r.$$

As we shall see, the graph of this function in x, y, z-space passes through the point $(x, y, z) = (0, 0, r)$ on the z-axis. This point is called the ***z*-intercept** of the graph. Sometimes we simply call the number r itself the z-intercept.

Notice that, with a little algebra, we can convert the previous example to the form $z = -\frac{1}{2}x + y + 4$. This is the intercept form, and the z-intercept is $z = 4$.

If there are n input variables, $x_1, x_2, \ldots, x_n$, instead of two, and an initial point has coordiantes $x_1^0, x_2^0, \ldots, x_n^0$, then a linear equation has the following forms:

initial-value: $z - z_0 = p_1(x_1 - x_1^0) + p_2(x_2 - x_2^0) + \cdots + p_n(x_n - x_n^0),$

***z*-intercept:** $z = p_1x_1 + p_2x_2 + \cdots + p_nx_n + r.$

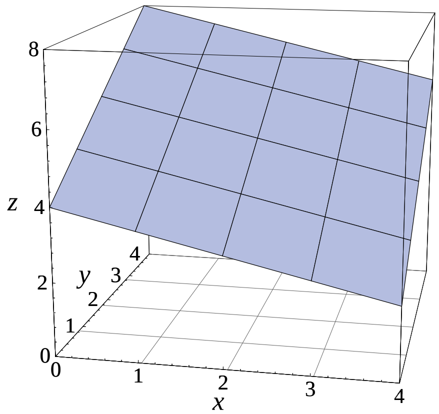
The form of a linear function of n variables

The graph of a linear function

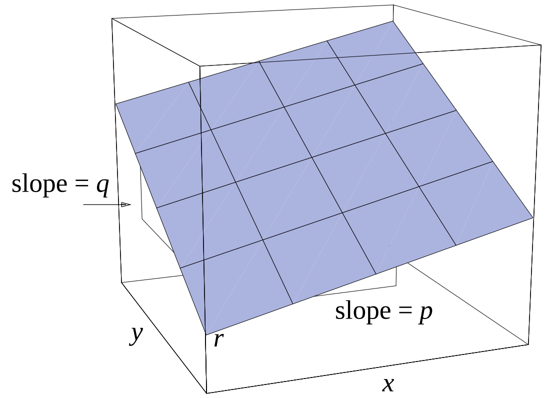
On the left at the top of the next page is the graph of the linear function

$$z = \tfrac{1}{2}x + y + 4.$$

The graph is a flat plane. In particular, grid lines parallel to the x-axis (which represent vertical slices with $y = c$) are all straight lines with the same slope $\Delta z/\Delta x = -\frac{1}{2}$. The other grid lines (with $x = c$) are all straight lines with the same slope $\Delta z/\Delta y = +1$.



$z = -\frac{1}{2}x + y + 4$



$z = px + qy + r$

On the right, above, is the graph of the general linear function written in intercept form: $z = px + qy + r$. The graph is the plane that can be identified by three distinguishing features:

- it has slope p in the x-direction;
- it has slope q in the y-direction.
- it intercepts the z-axis at $z = r$;

The definition of a linear function implies that its graph is a flat plane

Let's see how we can *deduce* that the graph must be this plane. First of all, the partial rate $\Delta z/\Delta x$ tells us how z changes when y is held fixed. But if we fix $y = c$, we get a vertical slice of the graph in the x-direction. The slope if that vertical slice is $\Delta z/\Delta x = p$. Since p is constant, the slice is a straight line. The value of $y = c$ determines which slice we are looking at. Since $\Delta z/\Delta x$ doesn't depend on y, all the slices in the x-direction have the *same* slope. Similarly, all the slices in the y-direction are straight lines with the same slope q. The only surface that can be covered by a grid of straight lines in this way is a flat plane. Finally, since $z = r$ when $(x, y) = (0, 0)$, the graph intercepts the z-axis at $z = r$.

Contours of a linear function

Each contour is a straight line

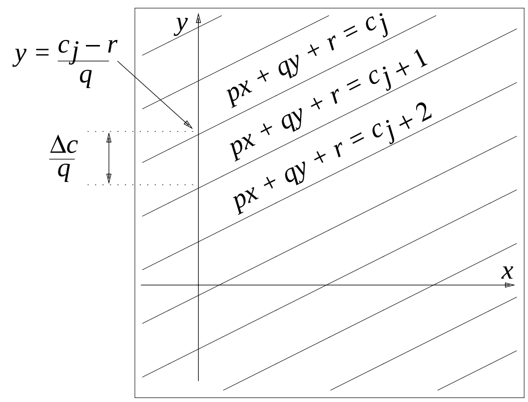
A contour of *any* function $f(x, y)$ is the set of points in the x, y-plane where $f(x, y) = c$, for some given constant c. If $f = px + qy + r$, then a contour has the equation

$$px + qy + r = c \qquad \text{or} \qquad y = -\frac{p}{q}x + \frac{c - r}{q}.$$

This is an ordinary straight line in the x, y-plane. Its slope is $-p/q$ and its y-intercept is $(c-r)/q$. (If $q=0$ we can't do these divisions. However, this causes no problem; you should check that the contour is just the vertical line $x=(c-r)/p$.)

To construct a contour plot, we must give the constant c a sequence of equally-spaced values c_j, with $c_{j+1}=c_j+\Delta c$. This generates a sequence of straight lines

$$px+qy+r=c_j, \quad \text{or} \quad y=-\frac{p}{q}x+\frac{c_j-r}{q}.$$

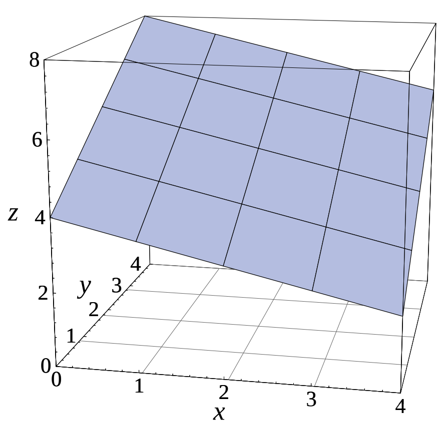


These lines all have the same slope $-p/q$, so they are parallel. (Notice the value of c doesn't affect the slope.) The y-intercept of the j-th contour is $(c_j-r)/q$. Therefore, the distance along the y-axis between one intercept and the next is $\Delta c/q$. The contours are thus straight, parallel, and equally-spaced. (You should check that this is still true if $q=0$.) Note that the figure at the left, above, is drawn with $\Delta c>0$ but $q<0$.

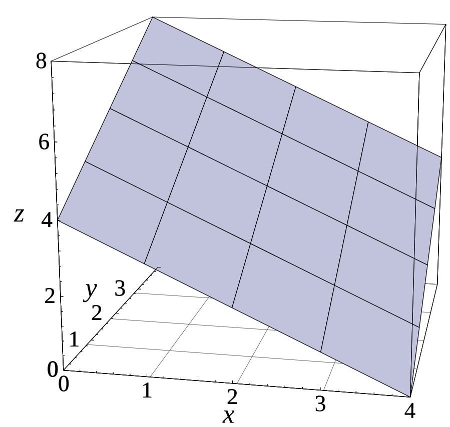
Geometric interpretation of the partial rates

Partial rates and partial slopes

What happens to the graph or the contour plot if you double one of the partial rates of change of a linear function? The graph on the right, below, shows the effect of doubling the partial rate with respect to x of the function $z=-\frac{1}{2}x+y+4$. As you can see, the slope in the x-direction



$z=-\frac{1}{2}x+y+4$



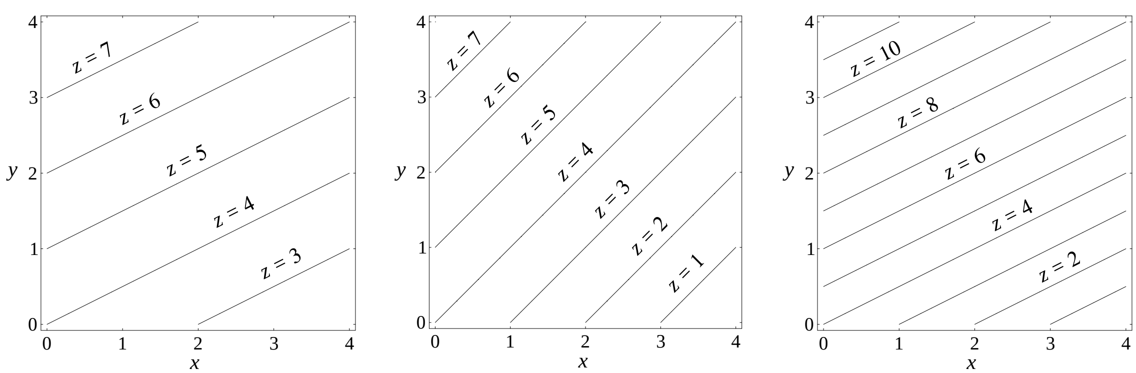
$z=-x+y+4$

The overall tilt of a graph is altered

is doubled (from $-\frac{1}{2}$ to -1). Had we increased the partial rate by a factor of 10, the slope would have increased by a factor of 10 as well. Notice that the slope in the y-direction is not affected. Nevertheless, the overall 'tilt' of the graph *has* been altered. We shall have more to say about this feature in a moment, when we introduce the **gradient** of a linear function to describe the overall tilt.

Partial rate and the spacing of contours

A change in the partial rates has a more complex effect on the contour plot. Perhaps it is more surprising, too. To make valid comparisons, we have constructed all three plots below with the same spacing between levels (namely $\Delta z = 1$). Notice how the levels meet the x- and y-axes in the plot on the left ($z = -tfrac12x + y + 4$). For each unit step we take along the y-axis, the z-value increases by 1. This is the meaning of $\Delta z/\Delta y = +1$. By contrast, we have to take *two* unit steps along the x-axis to produce the same size change in z. Moreover, z *decreases* by 1 when x increases by 2. This is the meaning of $\Delta z/\Delta x = -\frac{1}{2}$. In particular, the relatively wide spacing between z-levels along the x-axis reflects the relative smallness of $\Delta z/\Delta x$.



$z = -\frac{1}{2}x + y + 4$ $\qquad$ $z = -x + y + 4$ $\qquad$ $z = -x + 2y + 4$

The larger the partial rate, the closer the contours

Therefore, when we double the size of $\Delta z/\Delta x$—as we do in the middle plot—we should cut in half the spacing between z-levels along the x-axis. As you can see, this is exactly what happens. Notice that the spacing along the y-axis is not altered. Consequently, the contours change direction and they get packed more closely together.

Suppose we double *both* partial rates—as we do in the plot on the right. Then the spacing between contours is cut in half along both axes. Because the change is uniform, the contours keep their original direction.

The Gradient of a Linear Function

The vector of partial rates

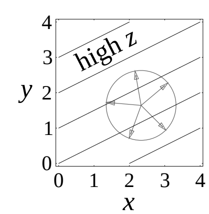
By making use of the concept of a vector, introduced in the last chapter, we can construct still another geometric interpretation of the partial rates of a linear function. This vector is called the gradient, and it is defined in the following way.

> **Definition**. The **gradient** of a linear function $z = f(x, y)$ is the vector whose components are its partial rates of change:
> $$\text{grad } z = \nabla z = \left(\frac{\Delta z}{\Delta x}, \frac{\Delta z}{\Delta y}\right).$$

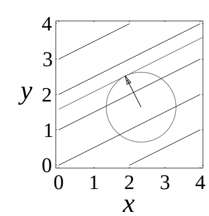
The direction of most rapid growth

The gradient is perhaps the most concise and useful tool for describing the growth of a function of several variables. To get an idea of the role that it plays, consider this question: *In what direction should we move from a given point in the x, y-plane so that the value of a linear function increases most rapidly*?

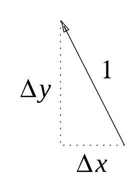
$z = -\frac{1}{2}x + y + 4$



Of course, the answer will depend on the linear function. Let's use $z = -\frac{1}{2}x + y + 4$ and start from the point $(x, y) = (2.4, 1.6)$. We can make z undergo a very large change simply by moving very far from this point. Therefore, to make valid comparisons, we will restrict ourselves to motions that carry us exactly one unit of distance in various directions. The vectors in the figure at the right show some of the possibilities. Their tips lie on a circle of radius 1.



Thus, to choose the direction in which z increases most rapidly, we must simply find the point on this circle where the value of z is largest. The contour line at this level must be tangent to the circle. The vector perpendicular to this contour line (see the second figure) therefore points in the direction of most rapid growth. Since perpendiculars have negative reciprocal slopes, and since all the contour lines have slope $+1/2$, it follows that the vector must have slope $-2/1$.



At the left is a magnified view of this vector. We know

$$\Delta x < 0, \qquad \frac{\Delta y}{\Delta x} = -2, \qquad \text{and} \qquad (\Delta x)^2 + (\Delta y)^2 = 1.$$

Thus $\Delta y = -2 \cdot \Delta x$, so $(\Delta x)^2 + 4\,(\Delta x)^2 = 1$. This implies

$$5\,(\Delta x)^2 = 1, \qquad \text{so} \qquad \Delta x = \frac{-1}{\sqrt{5}}, \quad \Delta y = \frac{2}{\sqrt{5}}.$$

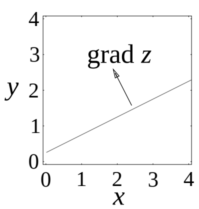
Thus, among all the motions $(\Delta x, \Delta y)$ we have considered, we obtain the greatest change in z by choosing

$$(\Delta x, \Delta y) = \left(\frac{-1}{\sqrt{5}}, \frac{2}{\sqrt{5}}\right).$$

The magnitude of most rapid growth

To determine how large this change is, we can use the alternate definition of a linear function (see page 536)

$$\Delta z = \frac{\Delta z}{\Delta x} \cdot \Delta x + \frac{\Delta z}{\Delta y} \cdot \Delta y = -\frac{1}{2} \cdot \frac{-1}{\sqrt{5}} + 1 \cdot \frac{2}{\sqrt{5}} = \frac{5}{2\sqrt{5}} = \frac{\sqrt{5}}{2}.$$



The gradient vector quickly gives us all this information. First of all, the gradient vector has the value

$$\text{grad } z = \left(\frac{\Delta z}{\Delta x}, \frac{\Delta z}{\Delta y}\right) = \left(-\tfrac{1}{2}, 1\right).$$

Information from the gradient

Since its slope is $1/-\frac{1}{2} = -2$, we see that it does indeed point in the direction of most rapid growth. Consequently, it is also perpendicular to the contour line. Furthermore, its *length* gives the maximum growth rate. We can see this by calculating the length using the Pythagorean theorem:

$$\text{length} = \sqrt{\left(\frac{\Delta z}{\Delta x}\right)^2 + \left(\frac{\Delta z}{\Delta y}\right)^2} = \sqrt{\frac{1}{4} + 1} = \sqrt{\frac{5}{4}} = \frac{\sqrt{5}}{2}.$$

Our findings with this example point to the following conclusion.

> **Theorem**. The gradient of the linear function $z = px + qy + r$ is perpendicular to its contour lines. It points in the direction in which z increases most rapidly, and its length is equal to the maximum rate of increase.

A proof

Let's see why this is true. According to the observation on the previous page, the direction of most rapid increase will be perpendicular to the contour lines. The gradient of $z = px + qy + r$ is the vector $\nabla z = (p, q)$. Its slope is q/p. On page 539 we saw that the slope of the contour lines is $-p/q$. Since these slopes are negative reciprocals, the gradient is indeed perpendicular to the contour lines.

To determine the maximum rate of increase, we must see how much z increases when we move exactly 1 unit of distance in the gradient direction. The gradient vector is (p, q), and its length is $\sqrt{p^2 + q^2}$. Therefore, the vector

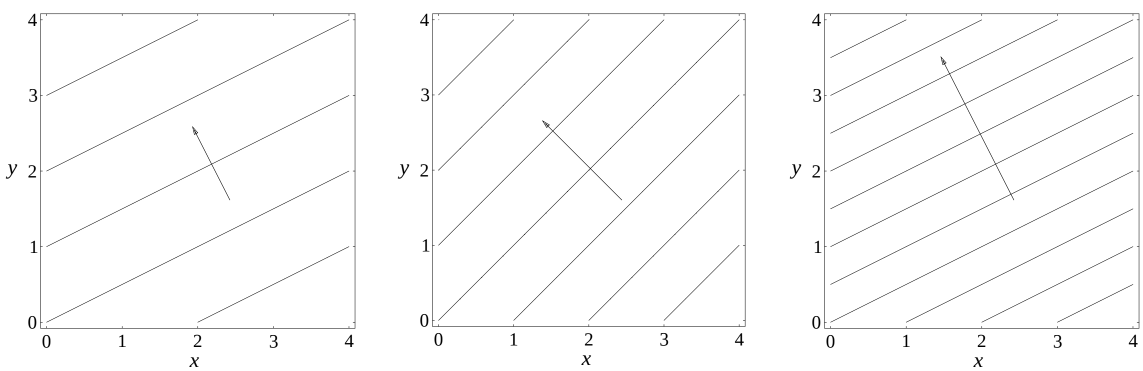
$$(\Delta x, \Delta y) = \left(\frac{p}{\sqrt{p^2 + q^2}}, \frac{q}{\sqrt{p^2 + q^2}} \right)$$

is 1 unit long and in the same direction as the gradient. The increase in z along this vector is

$$\Delta z = p \cdot \Delta x + q \cdot \Delta y = p \cdot \frac{p}{\sqrt{p^2 + q^2}} + q \cdot \frac{q}{\sqrt{p^2 + q^2}} = \frac{p^2 + q^2}{\sqrt{p^2 + q^2}} = \sqrt{p^2 + q^2}.$$

End of the proof

This *is* the length of the gradient vector, so we have confirmed that the length of the gradient is equal to the maximum rate of increase.



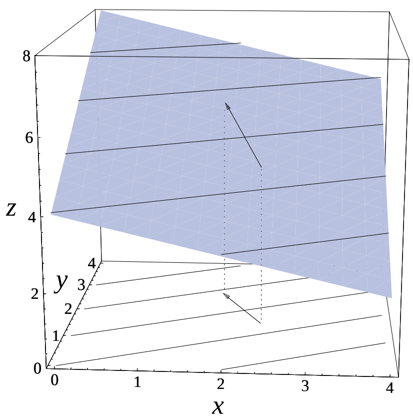
$z = -\frac{1}{2}x + y + 4$ $z = -x + y + 4$ $z = -x + 2y + 4$

Contour spacing and the length of the gradient

Shown above are the three linear functions we've already examined. In each case the gradient vector is perpendicular to the contours, and it gets longer as the space between the contours *decreases*. This is to be expected because the space between contours is also an indicator of the maximum rate of growth of the function. Widely-spaced contours tell us that z changes relatively little as x and y change; closely-spaced contours tell us that z changes a lot as x and y change.

The gradient points directly uphill

The connection between the gradient and the graph is particularly simple. Since the gradient (which is a vector in the x, y-plane) points in the direction of greatest increase, it points in the direction in which the graph is tilted up. If we project the gradient vector onto the graph, as in the figure at the top of the next page, it points directly "uphill". Putting it another way, we can



say that the gradient shows us the "overall tilt" of the graph. There are two parts to this information. First, the direction of the gradient tells us which way the graph is tilted. Second, the length tells us how steep the graph is.

The figure at the left combines all the visual elements we have introduced to analyze a linear function: contours, graph, and gradient. Study it to see how they are related.

The Microscope Equation

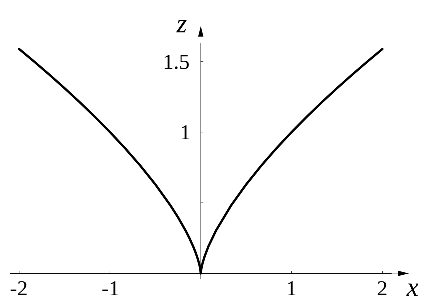
Local linearity

Local linearity

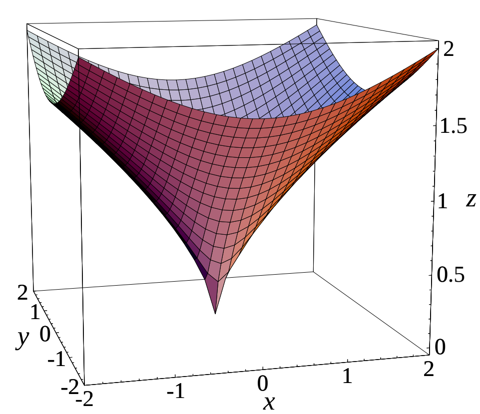
Let's return to arbitrary functions of two variables—that is, ones that are not necessarily linear. First we looked at magnifications of their graphs and contour plots under a microscope. We found that the graph becomes a plane, and the plot becomes a series of parallel, equally-spaced lines. Next, we saw that it is precisely the linear functions which have planar graphs and uniformly parallel contour plots. Hence this function is **locally linear**.

Exceptions

Of course, not *every* function is locally linear, and even a function that *is* locally linear at most points may fail to be so at particular points. We have already seen this with functions of a single variable in chapter 3. For example, $g(x) = x^{2/3}$ is locally linear everywhere *except* the origin. It has a sharp spike there. The two-variable function $f(x, y) = (x^2 + y^2)^{1/3}$ has the same sort of spike at the origin. The two graphs help make it clear that g is just a slice of f



$z = g(x) = x^{2/3}$



$z = f(x, y) = (x^2 + y^2)^{1/3}$

(constructed by taking $y = 0$). (Compare chapter 3.2, pages 113–114.) The spike is just one example; there are many other ways that a function can fail to be locally linear.

The relation between calculus and fractals

Functions that are *nowhere* locally linear are now being used in science to construct what are called **fractal** models. However, calculus does not deal with fractals. On the contrary, we remind you of the stipulation first made in chapter 3:

Calculus studies functions that are locally linear almost everywhere.

The microscope equation with two input variables

The equation of a microscopic view

If the function $z = f(x, y)$ is locally linear, then its graph looks like a plane when we view it under a microscope. The linear equation that describes that plane is the **microscope equation**. Since the plane is part of the graph of f, f itself must determine the form of the microscope equation. Let's see how that happens.

The idea is to reduce f to a function of one variable and then use the microscope equation for one-variable functions (described in chapter 3.3 and 3.7). Suppose the microscope is focused at the point $(x, y) = (a, b)$. If we fix y (at $y = b$), then z depends on x alone: $z = f(x, b)$. The microscope equation for this function at $x = a$ is just

The microscope equation in the x-direction. . .

$$\Delta z \approx \frac{\partial f}{\partial x}(a, b) \cdot \Delta x.$$

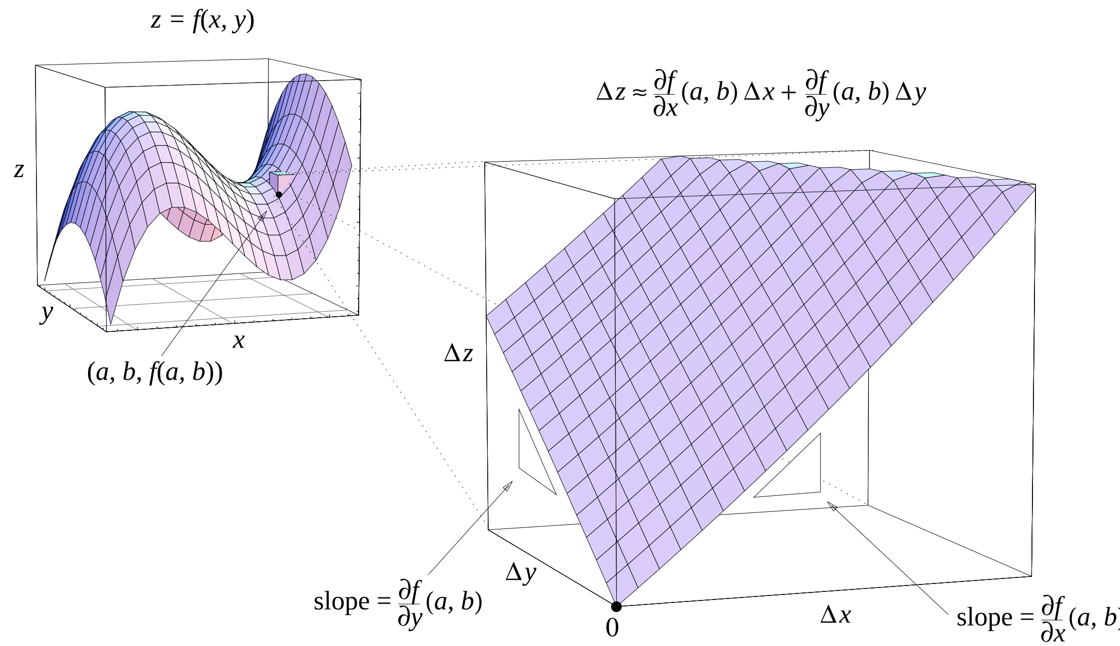
The multiplier $\partial f/\partial x$ is the rate of change of f with respect to x. We need to write it as a partial derivative because f is a function of two variables. Geometrically, the multiplier tells us the slope of a vertical slice of the graph in the x-direction.

Now reverse the roles of x and y, fixing $x = a$. The microscope equation for the function $z = f(a, y)$ at $y = b$ is

. . . and in the y-direction

$$\Delta z \approx \frac{\partial f}{\partial y}(a, b) \cdot \Delta y.$$

The multiplier $\partial f/\partial y$ in this equation is the slope of a vertical slice of the graph in the y-direction. The slopes of the two vertical slices are indicated in the microscope window that appears in the foreground of the following figure.



From partial changes to total change

The separate microscope equations for the x- and y-directions give us the partial changes in z. However, as we saw when we were defining linear functions (page 536), when we know all the *partial* changes, we can immediately write down the *total* change:

The microscope equation:

$$\Delta z \approx \frac{\partial f}{\partial x}(a,b) \cdot \Delta x + \frac{\partial f}{\partial y}(a,b) \cdot \Delta y$$

As always, the origin of the microscope window corresponds to the point $(a, b, f(a, b))$ on which the microscope is focused. The microscope coordinates Δx, Δy, and Δz measure distances from this origin. For the sake of clarity in the figure above, we put the origin at one corner of the (three-dimensional) window.

An example

Incidentally, the function shown above is $f(x, y) = x^3 - 4x - y^2$, and the microscope is focused at $(a, b) = (1.5, -1)$. Since

$$\frac{\partial f}{\partial x} = 3x^2 - 4 = 2.75, \qquad \frac{\partial f}{\partial y} = -2y = 2$$

when $(x, y) = (1.5, -1)$, the microscope equation is

$$\Delta z \approx 2.75\,\Delta x + 2\,\Delta y.$$

Linear Approximation

The microscope equation gives a linear approximation

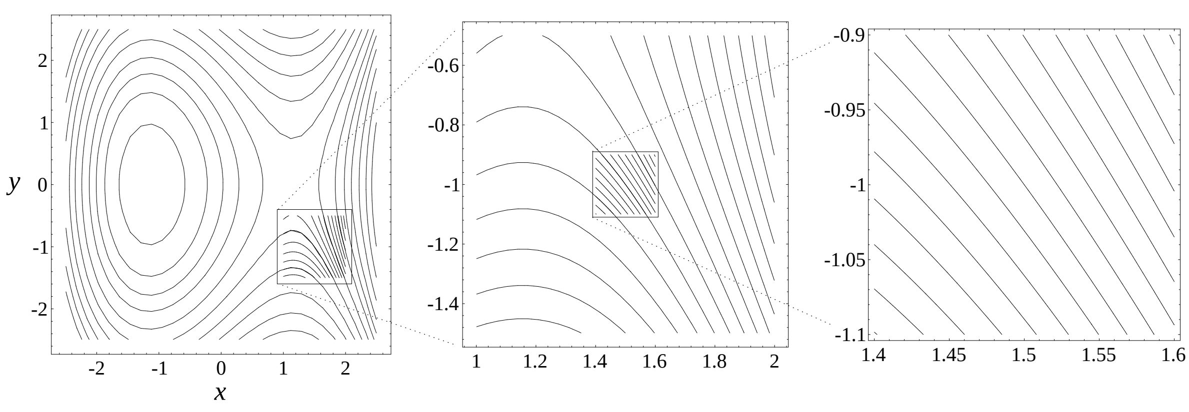
The microscope equation describes a linear function that approximates the original function near the point on which the microscope is focused. It is easy to see exactly how good the approximation is by comparing contour plots of the two functions. This is done below. In the window on the right, which shows the highest magnification, the solid contours belong to the original function $f = x^3 - 4x - y^2$. They are curved, but only slightly so. The dotted contours belong to the linear function

$$(z + 3.625) = 2.75(x - 1.5) + 2(y + 1) \quad \text{or} \quad z = 2.75\,x + 2\,y - 5.75.$$

(This is the microscope equation expressed in terms of the original variables x, y and z instead of the microscope coordinates $\Delta x = x - 1.5$, $\Delta y = y - (-1)$ and $\Delta z = z - (-3.625)$.)

Comparing the contours . . .

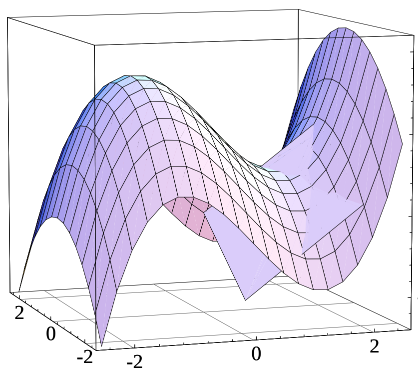
The difference between the two sets of contours shows us just how good the approximation is. As you can see, the two functions are almost indistinguishable near the center of the window, which is the point $(x, y) = (1.5, -1)$. As we look farther from the center, we find the contours of f depart more and more from strict linearity.

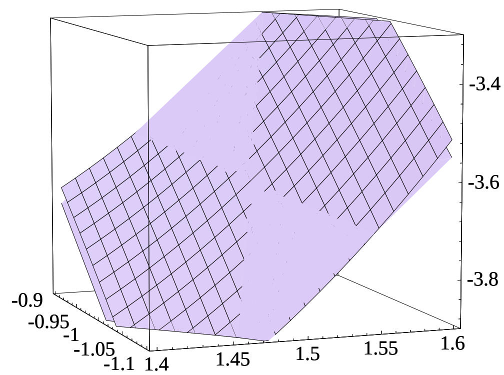


. . . and the graph of the function and its linear approximation

We can also compare the *graphs* of a function and its linear approximation. The graph of the linear approximation is a plane, of course. It is, in fact, the plane that is tangent to the graph of the function.

In the left figure immediately below we see the tangent plane to the graph of $z = f(x, y) = x^3 - 4x - y^2$ at the point where $(x, y) = (1.5, -1)$. On the right is a magnified view at the point of tangency. The graph of f is almost flat.

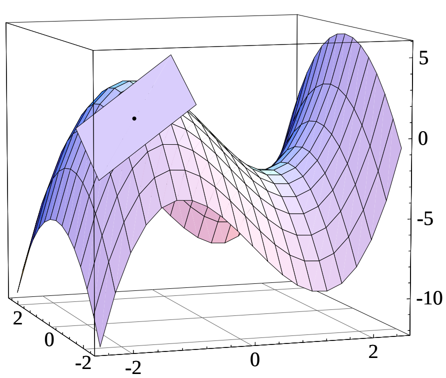


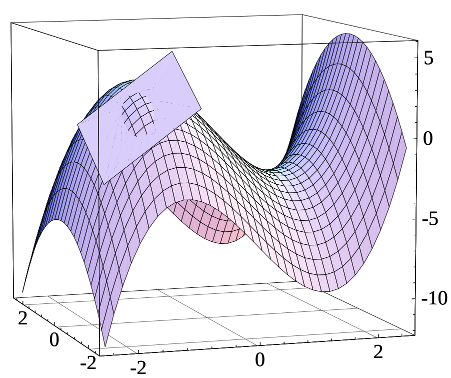


The grid helps us distinguish it from the tangent plane, which is solid gray. Such close agreement between the graph and the plane demonstrates how good the linear approximation is near the point. Notice, however, that the plane diverges from the graph as we move away from the point of tangency.

Tangent planes

At first glance you may not think that the plane in the figures above is tangent to the surface. The word "tangent" comes from the Latin *tangere*, to touch. We sometimes take this to mean "touch at one point", like the plane in the figure at the left, below. More properly, though, two objects are **tangent** if they have the same direction at a point where they meet. The plane in the figures above *does* meet this condition—as the microscopic view helps make clear.

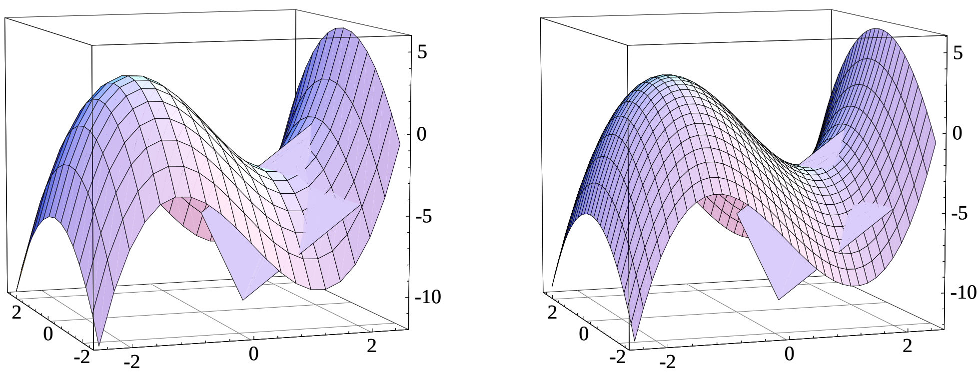




Elliptic points...

There are many different ways that a tangent plane can intersect a surface. What happens depends on the shape of the surface at the point of tangency.

The surface could bend the same way in all directions at that point, or it could bend up in some directions and down in others. In the first case, it will bend away from its tangent plane, so the two will meet at only one point. This is called an **elliptic point**, because the intersection turns into an ellipse if we push the plane in a bit. The figure on the right, above, shows what happens.



Suppose, on the other hand, that the surface bends up in some directions but down in others. Then, in some intermediate directions, it will not be bending at all. In those directions it will meet its tangent plane—which doesn't bend, either. Typically, there are two pairs of such directions. The surface and the tangent plane then intersect in an X. This always happens at a minimax (or saddle point) on a graph. and it also happens at the first point we considered on the graph of $z = x^3 - 4x - y^2$. This is shown again in the figure on the left above. It is called a **hyperbolic point**, because the intersection turns into a hyperbola if we push the plane a bit—as on the right. The lines where the tangent plane itself intersects the surface are called **asymptotic lines**, because they are the asymptotes of the hyperbolas. (To make it easier to see the elliptical and hyperbolic intersections we made the surface grid finer.)

. . . and hyperbolic points

The curve of intersection between a surface and a shifted tangent plane is called the *Dupin indicatrix*. The Dupin indicatrix can take many forms besides the ones we have described here. However, at almost all points on almost all surfaces it turns out to be an ellipse or a hyperbola. More precisely, the indicatrix is *approximately* an ellipse or a hyperbola—in the same way that the surface itself is only approximately flat.

Most points on a surface fall into one of two regions; one region consists of elliptic points, the other of hyperbolic. Points on the boundary between these two regions are said to be **parabolic**. On the graph of $z = x^3 - 4x - y^2$, if $x < 0$ the point is elliptic; if $x > 0$ it is hyperbolic; and if $x = 0$, it is parabolic. Try to confirm this yourself, just by looking at the surface.

Parabolic points

The classification of the points into elliptical, parabolic, and hyperbolic types is one of the first steps in studying the curvature of a surface. This is part of *differential geometry*; calculus provides an essential language and tool. Differential geometry is used to model the physical world at both the cosmic scale (general relativity) and the subatomic (string theory).

The Gradient

Consequences of local linearity

Local linearity is a powerful principle. It says that an arbitrary function looks linear when we view it on a sufficiently small scale. In particular, all these statements are approximately true in a microscope window:

- the contours are straight, parallel, and equally spaced;
- the graph is a flat plane (the tangent plane);
- the function has a linear formula (the microscope equation);
- the partial derivatives are the slopes of the graph in the directions of the axes.

Extending the gradient to non-linear functions

There is one more aspect of a linear function for us to interpret—the gradient. Since partial rates become partial derivatives, we can make the following definition for an arbitrary locally linear function.

Definition. The **gradient** of a function $z = f(x, y)$ is the vector whose components are the partial derivatives:

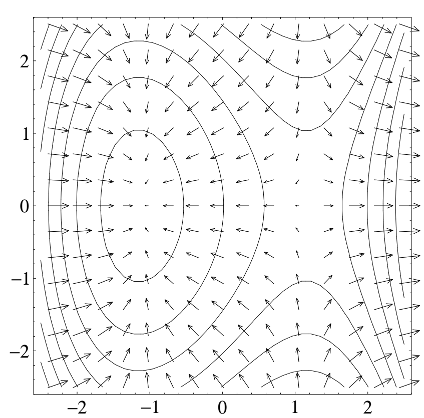
$$\text{grad } f = \nabla f = (f_x, f_y) = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right).$$

The gradient vector field

The partial derivatives are functions, so the gradient varies from point to point. Thus, gradients form a **vector field** in the same sense that a dynamical system does (see chapter 8). We draw the gradient vector $(f_x(x, y), f_y(x, y))$ as an arrow whose tail is at the point (x, y).

Example. $f(x, y) = x^3 - 4x - y^2$, grad $f = (3x^2 - 4, -2y)$

Contours and gradient vectors together



There is one thing you should notice about the previous figure: we drew the gradient vectors much shorter than they actually are. For instance, at the origin grad $f = (-4, 0)$, but the arrow *as drawn* is closer to $(-.25, 0)$. The purpose of rescaling is to keep the vectors out of each other's way, so the overall pattern easier to see.

The vectors are rescaled for clarity

The example shows contours as well as gradients so we can see how the two are related. The result is very striking. Even though the vectors vary in length and direction, and the contours vary in direction and spacing, the two are related the same way they were for a *linear* function (page 542ff). First of all, each vector is perpendicular to the level curve that passes through its tail. Second, the vectors get longer where the spacing between level curves gets smaller. The similarity is no accident, of course; it is a consequence of local linearity. We can summarize and extend our observations in the following theorem. It is just a modification of the earlier theorem on the gradient of a linear function (page 542).

The relation between gradients and contours

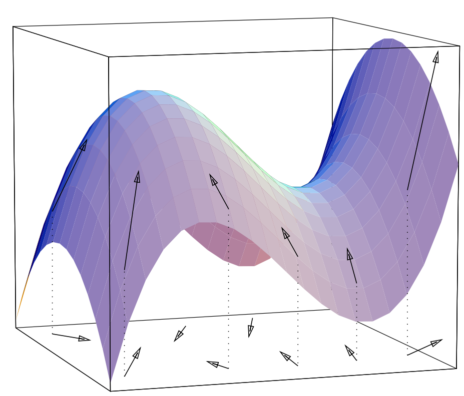
> **Theorem**. The gradient vector field of the function $z = f(x, y)$ is perpendicular to its contour lines. At each point, the *direction* of the gradient is the direction in which z increases most rapidly; the *length* is equal to the maximum rate of increase.

To see why this theorem is true, just look in a microscope. The gradient and the contours become the gradient and the contours of the linear approximation at the point where the microscope is focused. But we already know the theorem is true for linear functions, so there is nothing more to prove.

A proof

There is a direct connection between the gradient field of a function $z = f(x, y)$ and its graph. Since the gradient (which is a vector in the x, y-plane) points in the direction in which z increases most rapidly, it points in the direction in which the graph is tilted up. Thus, if we project the gradient onto the graph, as we do in the figure at the left, it points directly "uphill". Since f is not a linear function, both the steepness and the uphill direction vary from point to point. The gradient vector field also varies; in this way, it keeps track of the steepness of the graph and the direction of its tilt.

The gradient points uphill



The Gradient of a Function of Three Variables

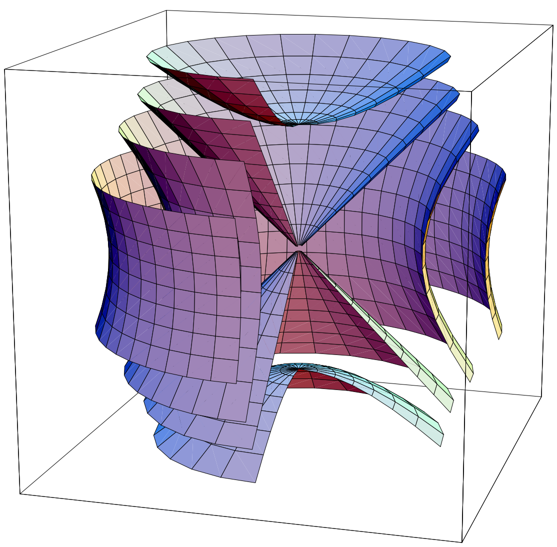
Let's take a brief glance at the gradient of a function $f(x, y, z)$. It has three components:

$$\text{grad} f = \nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right) = (f_x, f_y, f_z),$$

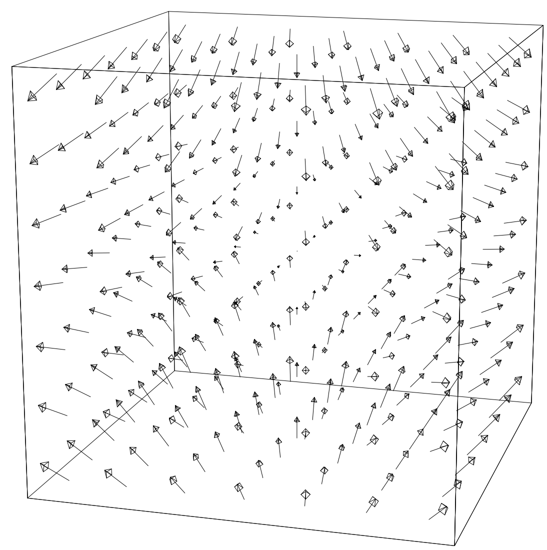
and it defines a vector field in x, y, z-space. At each point, the gradient of f is perpendicular to the level set through that point, and it points in the direction in which f increases most rapidly.

Visualizing a three-dimensional vector field

Example. In the two boxes below you can compare the gradient field of $f(x, y, z) = x^2 + y^2 - z^2$ with its level sets. At first glance, you may not find a clear pattern to the gradient vectors. After all, the picture is three-dimensional, and it is difficult to tell whether an arrow is near the front of the box or the back. However, there is a pattern: at the top and the bottom of the box, arrows point inward; closer to the middle of the box, they flare outward. The lowest values of the function occur along the z-axis, inside the shallow bowls that sit at the top and the bottom of the box. The highest values occur outside the "equatorial belt" formed by the outermost level set. This is where the x, y-plane meets the middle of the box. Notice also that the level sets are symmetric around the z-axis. The gradient field has the same symmetry, though this is harder to see.



$x^2 + y^2 - z^2 = c$



$\nabla f = (2x, 2y, -2z)$

Exercises

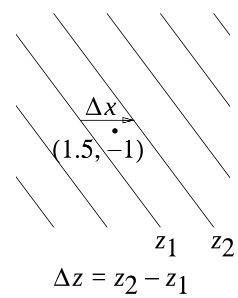
In many of these exercises it will be essential to have a computer program to make graphs and contour plots of functions of two variables.

1. a) Obtain the graph of $z = x^3 - 4x - y^2$ on a domain centered at the point $(x, y) = (1.5, -1)$, and magnify the graph until it looks like a plane.

b) Estimate, by eye, the slopes in the x-direction and the y-direction of the plane you found in part (a). [You should find numbers between +2 and +3 in both cases.]

2. Obtain a contour plot of $z = x^3 - 4x - y^2$ on a domain centered at the point $(x, y) = (1.5, -1)$, and magnify the plot until the contours look straight, parallel, and equally spaced. Compare your results with the plots on page 535.

3. Continuation. The purpose of this exercise is to estimate the rate of change $\Delta z/\Delta x$ at $(x, y) = (1.5, -1)$, using the most highly magnified contour plot you constructed in the last exercise.



a) What is the horizontal spacing Δx between the two contours closest to the point $(1.5, -1)$? See the illustration at the right.

b) Find the z-levels z_1 and z_2 of those contours, and then compute $\Delta z = z_2 - z_1$.

c) Compare the value of $\Delta z/\Delta x$ you now obtain with the slope in the x-direction that you estimated in exercise 1.

4. Continuation. Repeat all the work of the last exercise, this time for $\Delta z/\Delta y$.

Linear functions

5. a) Find the z-intercept of the graph of the linear function given by the formula

$$z - 3 = 2(x - 4) - 3(y + 1).$$

b) Write the formula for this linear function in intercept form.

6. a) Write, in initial-value form, the formula for the linear function $z = L(x, y)$ for which $\Delta z/\Delta x = 3$, $\Delta z/\Delta y = -2$, and $L(1, 4) = 0$.

b) Write the intercept form of the same function. What is the z-intercept of the graph of L?

7. a) Suppose z is a linear function of x and y, and $\Delta z/\Delta x = -7$, $\Delta z/\Delta y = 12$. If (x, y) changes from $(35, 24)$ to $(33, 33)$, what is the total change in z?

b) Suppose $z = 29$ when $(x, y) = (35, 24)$. What value does z have when $(x, y) = (33, 33)$? What value does z have when $(x, y) = (0, 0)$?

c) Write the intercept form of the formula for z in terms of x and y.

8. Suppose z is a linear function of x and y for which we have the following information:

x	5	7	0		4	0	4	
y	9	1	4	6	7	0		-1
z	2		12	-7	2		20	20

a) Fill in the blanks in this table.

b) Write the formula for z in intercept form.

9. a) Sketch the graph of $z = x - 2y + 7$ on the domain $0 \le x \le 3$, $0 \le y \le 3$.

b) Determine the slope of this graph in the x-direction, and indicate on your sketch where this slope can be found.

c) Determine the slope of this graph in the y-direction, and indicate on your sketch where this slope can be found.

10. Continuation. Draw the gradient vector of $z = x - 2y + 7$ in the x, y-plane, and then lift it up so it sits on the graph you drew in the previous exercise. Does the gradient point directly "uphill"?

11. Sketch the graph of the linear function $z = L(x, y)$ for which

$$\frac{\Delta z}{\Delta x} = -1, \qquad \frac{\Delta z}{\Delta y} = .6, \qquad L(1, 1) = 8.$$

Be sure your graph shows clearly the slopes in the x-direction and the y-direction, and the z-intercept.

12. Continuation. Draw the gradient of the function L from the previous exercise, and lift it up so it sits on the graph of L you drew there. Does the gradient point directly uphill?

13. What is the equation of the linear function $z = L(x, y)$ whose graph (a) has a slope of -4 in the x-direction, (b) a slope of $+5$ in the y-direction, and (c) passes through the point $(x, y, z) = (2, -9, 0)$?

14. Suppose $z = L(x, y)$ is a linear function whose graph contains the three points

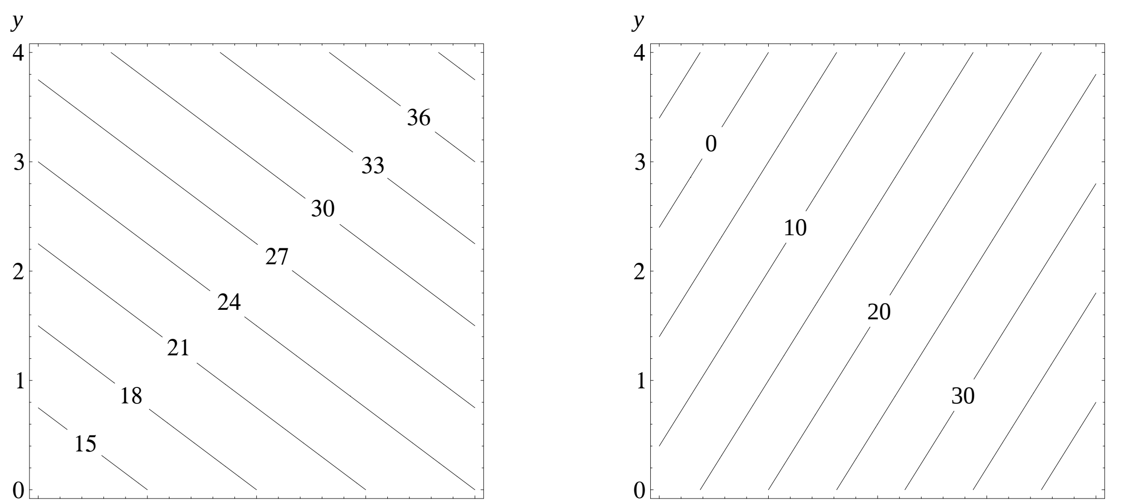
$$(1, 1, 2), \qquad (0, 5, 4), \qquad (-3, 0, 12).$$

a) Determine the partial rates of change $\Delta z/\Delta x$ and $\Delta z/\Delta y$.

b) Where is the z-intercept of the graph?

c) If $(4, 1, c)$ is a point on the graph, what is the value of c?

d) Is the point $(2, 2, 4)$ on the graph? Explain your position.



15. The figure on the left above is the contour plot of a function $z = L_1(x, y)$.

a) What are the values of $L_1(1, 0)$, $L_1(2, 0)$, $L_1(3, 0)$, $L_1(0, 3)$, $L_1(2, 3)$, and $L_1(0, 0)$?

b) What are the partial rates $\Delta L_1/\Delta x$ and $\Delta L_1/\Delta y$?

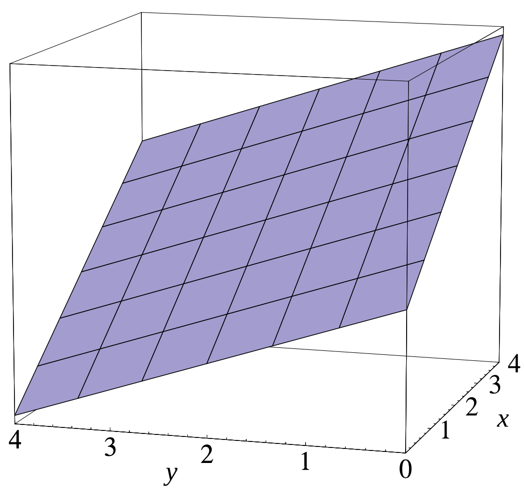
16. Continuation. Find the values of $L_1(3,2)$, $L_1(7,0)$, $L_1(7,7)$, $L_1(1.4,2.9)$, $L_1(-2,9)$, $L(-10,-100)$.

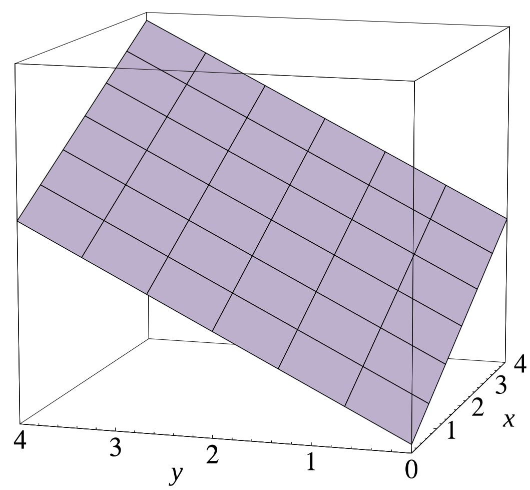
a) Find x so that $L_1(x,0)=0$. Find y so that $L_1(0,y)=0$.

17. Continuation. Write the intercept form of the formula for $L_1(x,y)$.

18. a) Find the partial rates $\Delta L_2/\Delta x$ and $\Delta L_2/\Delta y$ of the linear function L_2 whose contour plot is shown at the right on the previous page.

b) Obtain the intercept form of the formula for $L_2(x,y)$.





19. The figures above are the graphs of L_1 and L_2. Which is which? Explain your choice. (Note that both graphs are shown from the same viewpoint. The x-axis is on the right, and the y-axis is in the foreground. The z-axis has no scale on it.)

20. Determine the gradient vectors of L_1 and L_2.

a) Sketch the gradient vector of each function on the x,y-plane and on its own graph. Does the gradient point directly uphill in each case?

21. Suppose the gradient vector of the linear function $p=L(q,r)$ is grad $p=\nabla p=(5,-12)$. If $L(9,15)=17$, what is the value of $L(11,11)$?

22. a) What is the gradient vector of the function $w=2u+5v$?

b) At what point on the circle $u^2+v^2=1$ does w have its largest value? What is that value?

23. a) Write the formula of a linear function $z = L(x,y)$ whose gradient vector is grad $z = \nabla z = (-3, 4)$

b) Using your formula for L, calculate the total change in L when $\Delta x = 2$, $\Delta y = 1$.

24. a) Continuation; in particular, continue to use your formula for $z = L(x,y)$. What is the value of $L(0,0)$?

b) What is the maximum value of z on the circle $x^2 + y^2 = 1$? At what point (x,y) does z achieve that value?

c) Determine the difference between the maximum and minimum values of L on the circle $x^2 + y^2 = 1$.

[Answer: The difference is 10, independent of the formula you use.]

25. a) What value does $z = 7x + 3y + 31$ have when $x = 5$ and $y = 2$?

b) If x increases by 2, how must y change so that the value of z doesn't change. (The change in y needed to keep z fixed when x changes is called the **trade-off**. See also chapter 3, page 174.)

The concept of a *trade-off*

c) What is the trade-off in y when x increases by α?

d) What is the trade-off in x when y increases by β?

26. a) Suppose $z = L(x,y)$ is a linear function for which $\Delta z/\Delta x = 5$ and $\Delta z/\Delta y = -2$. What is the trade-off in y when x increases by 50?

b) What is the trade-off in x when y increases by 1?

27. Suppose $z = L(x,y)$ is a linear function and suppose the trade-off in y when x increases by 1 is -4.

a) What is the trade-off in y when x is *decreased* by 3?

b) What is the trade-off in x when y is increased by 10? [Note that x and y are reversed here, in comparison to the earlier parts of this question.]

28. Suppose $z = L(x,y)$ is a linear function for which we know

$$L(3,7) = -2, \qquad \frac{\Delta z}{\Delta x} = 2.$$

Suppose also that the trade-off in y when x increases by 10 is -4.

a) What is the value of $L(7,7)$?

b) If $L(7, \beta) = -2$, what is the value of β?

c) What is the value of $\Delta z/\Delta y$?

d) Write the formula for $L(x, y)$ in intercept form.

29. a) Suppose the graph of the linear function $z = L(x, y)$ has a slope of -1.5 in the x-direction and -2.4 in the y-direction. What is the trade-off between x and y? That is, how much should y change when x is increased by the amount α?

b) Suppose the partial slopes become $+1.5$ and $+2.4$ in the x- and y-directions, respectively. How does that affect the trade-off? Explain.

30. a) Sketch in the x, y-plane the set of points where $z = 2x + 3y + 7$ has the value 34.

b) If $x = 10$ then what value must y have so that the point (x, y) is on the set in part (a)?

c) If x increases from 10 to 14, how must y change so that the point (x, y) stays on the set in part (a)? In other words, what is the trade-off?

The set in the last question is called a *trade-off line*. Do you see that it is just a contour line by another name?

31. a) Write, in intercept form, the formula for the linear function

$$w - 4 = 3(x - 2) - 7(y + 1) - 2(z - 5).$$

b) What is the gradient vector of the linear function in part (a)?

32. Suppose the gradient of the linear function $w = L(x, y, z)$ is $\nabla w = (1, -1, 4)$. If $L(3, 0, 5) = 10$, what is the value of $L(1, 2, 3,)$?

33. Describe the level sets of the function $w = f(x, y, z) = x + y + z$.

The microscope equation

34. Find the microscope equation for the function $f(x, y) = 3x^2 + 4y^2$ at the point $(x, y) = (2, -1)$.

35. a) Continuation. Use the microscope equation to estimate the values of $f(1.93, -1.05)$ and $f(2.07, -.99)$

b) Calculate the *exact* values of the quantities in part (a), and compare those values with the estimates. In particular, indicate how many digits of accuracy the estimates have.

36. Find the microscope equation for the function $f(x, y, z) = x^2 y \sin z$ at the point $(x, y, z) = (1, 1, \pi)$.

37. Suppose $f(87, 453) = 1254$ and

$$\frac{\partial f}{\partial x}(87, 453) = -3.4, \qquad \frac{\partial f}{\partial y}(87, 453) = 4.2.$$

Estimate the following values: $f(90, 453)$, $f(87, 450)$, $f(90, 450)$, and $f(100, 500)$. Explain how you got your estimates.

38. a) Continuation. Find an estimate for y to solve the equation $f(87, y) = 1250$.

b) Find an estimate for x to solve $f(x, 450) = 1275$.

39. Continuation: a trade-off. Go back to the starting values $x = 87$, $y = 453$, and $f(87, 453) = 1254$. If x increases from 87 to 88, how should y change to keep the value of the function fixed at 1254?

40. a) Suppose $Q(27.3, 31.9) = 15.7$ and $Q(27.9, 31.9) = 15.2$. Estimate the value of $\partial Q/\partial x(27.6, 31.9)$.

b) Estimate the value of $Q(27, 31.9)$.

41. Suppose $S(105, 93) = 10$, $S(110, 93) = 10.7$, $S(105, 95) = 9.3$. Estimate the value of $S(100, 100)$. Explain how you made your estimate.

42. Let P be the point $(x, y, z) = (173, -29, 553)$. Suppose $f(P) = 48$ and

$$\frac{\partial f}{\partial x}(P) = 7, \qquad \frac{\partial f}{\partial y}(P) = -2, \qquad \frac{\partial f}{\partial z}(P) = 5.$$

Estimate the value of $f(175, -30, 550)$, and explain what you did.

43. a) What is the equation of the tangent plane to the graph of $z = xy$ at the point $(x, y) = (2, -3)$?

b) Which has a higher z-intercept: the graph or the tangent plane?

44. a) Suppose the function $H(x, y)$ has the microscope equation

$$\Delta H \approx 2.53\,\Delta x - 1.19\,\Delta y$$

at the point $(x, y) = (35, 26)$. Sketch the gradient vector ∇H at that point.

b) Pick the point exactly one unit away from $(35, 26)$ at which you estimate H has the largest possible value.

[Answer: At $(35.324, 25.848)$, H is about 2.796 units larger than it is at $(35, 26)$.]

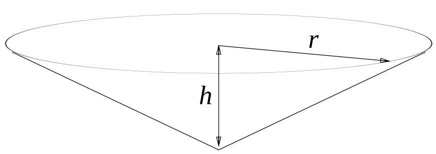
45. Write the microscope equation for the function $V(x, y) = x^2 y$ at the point $(x, y) = (25, 10)$.

Refer to the discussion of error propagation in chapter 3.4

46. a) Continuation. Suppose a cardboard carton has a square base that is 25 inches on a side and a height of 10 inches. If there is an error of Δx inches in measuring the base and an error of Δy inches in measuring the height, how much error will there be in the calculated volume?

b) Why is this a continuation of the previous question?

47. Continuation. Which causes a larger percentage error in the calculated volume: a 1% error in the measurement of the length of the base, or a 1% error in the measurement of the height?



48. A large basin in the shape of a cone is to be used as a water reservoir. If the radius r is 186 meters and the depth h is 31 meters, how much water can the basin hold, in cubic meters?

49. a) Continuation. If there were a 3% error in the measurement of the radius, how much error would that lead to in the calculation of the capacity of the basin?

b) If there were a 5% error in the measurement of the depth, how much error would there in the calculated capacity of the basin?

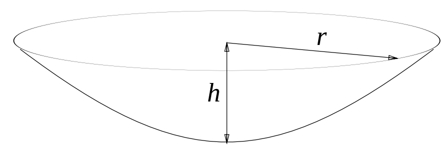
c) If *both* errors are present in the measurements, what is the total error in the calculated capacity of the basin?

50. Continuation. Suppose the measured radius of the basin ($r = 186$ meters) is assumed to be accurate to within 2%. The depth has been measured at 31 meters. Is it possible to make that measurement so accurate that the

calculated capacity is known to within 5%? How accurate does the depth measurement have to be?

51. Continuation. Suppose the accuracy of the radius measurement can only be guaranteed to be 3%. Is it still possible to measure the depth accurately enough to guarantee that the calculated capacity is accurate to within 5%? Explain.

52. Let's give the basin the more realistic shape of a parabolic bowl. If the radius r is still 186 meters and the depth h is still 31 meters, what is the capacity of the basin now? Is it larger or smaller than the conical basin of the same dimensions?



53. Continuation. Determine the error in the calculated capacity of the bowl if there were a 3% error in the measurement of the radius and, at the same time, a 5% error in the measurement of the depth.

54. Continuation. Is it possible for the calculated capacity to be accurate to within 5% when the measured radius is accurate to within 2%? How accurate does the measurement of the depth have to be to achieve this?

55. The energy of a certain pendulum whose position is x and velocity is v can be given by the formula

$$E(x, v) = 1 - \cos x + \tfrac{1}{2}v^2.$$

Suppose the position of the pendulum is known to be $x = \pi/2$ with a possible error of 5% and its velocity is $v = 2$ with possible error of 10%. What is the calculated value of the energy, and how accurately is that value known?

56. a) A frictionless pendulum conserves energy: as the pendulum moves, the value of E does not change. Suppose $x = \pi/2$ and $v = 2$, as in the previous exercise. When x decreases by $\pi/180$ (this is 1 degree), does v increase or decrease to conserve energy?

The conservation of energy leads to a trade-off

b) Approximately how much does v change when x decreases by $\pi/180$?

Linear approximations

57. a) Write the linear approximation to $f(x, y) = \sin x \cos y$ at the point $(x, y) = (0, \pi/2)$.

b) Write the equation of the tangent plane to the graph of $f(x, y)$ at the point $(x, y) = (0, \pi/2)$.

c) Where does the tangent plane in part (b) meet the x, y-plane?

58. Suppose $w(3, 4) = 2$ while

$$\frac{\partial w}{\partial x}(3, 4) = -1, \qquad \frac{\partial w}{\partial y}(3, 4) = 3.$$

Write the equation of the tangent plane of w at the point $(3, 4)$.

59. a) Write the equation of the tangent plane to the graph of the function $\varphi(x, y) = 3x^2 + 7xy - 2y^2 - 5x + 3y$ at an arbitrary point $(x, y) = (a, b)$.

b) At what point (a, b) is the tangent plane horizontal?

c) Magnify the graph of φ at the point you found in part (b) until it looks flat. Is it also horizontal?

The next few exercises concern the Lotka–Volterra differential equations and their linear approximations at an equilibrium point. Specifically, consider the *bounded growth* system from chapter 4.1 (pages 187–189):

$$\begin{aligned} R' &= .1\,R - .00001\,R^2 - .005\,RF, \\ F' &= .00004\,RF - .04\,F. \end{aligned}$$

60. Confirm that the system has an equilibrium point at $(R, F) = (1000, 18)$. Then obtain the phase portrait of the system near that point. What kind of equilibrium is there at $(1000, 18)$?

61. Obtain the linear approximations of the functions

$$\begin{aligned} g_1(R, F) &= .1\,R - .00001\,R^2 - .005\,RF, \\ g_2(R, F) &= .00004\,RF - .04\,F, \end{aligned}$$

at the point $(r, F) = (1000, 18)$. Call them $\ell_1(R, F)$ and $\ell_2(R, F)$, respectively.

62. Obtain the phase portrait of the **linear** dynamical system

$$\begin{aligned} R' &= \ell_1(R, F), \\ F' &= \ell_2(R, F). \end{aligned}$$

a) Does this system have an equilibrium at $(1000, 18)$? Compare this phase portrait to the phase portrait of the original *non*-linear system.

The gradient field

63. a) Make a sketch of the gradient vector field of $f(x, y) = x^2 - y^2$ on the domain $-2 \leq x \leq 2$, $-2 \leq y \leq 2$.

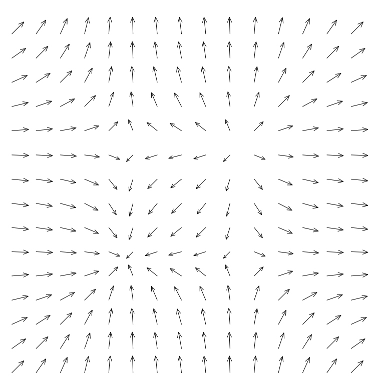
b) Mark on your sketch where the gradient field indicates the maximum and the minimum values of f are to be found in the domain.

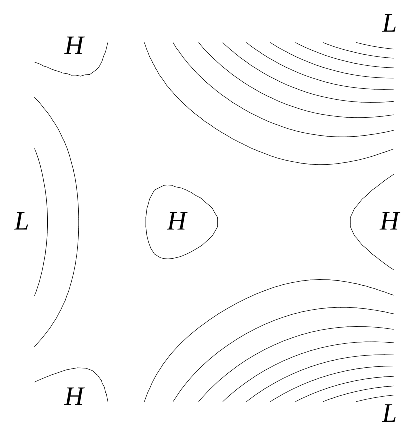
64. Repeat the previous exercise, using $f(x, y) = 2x + 4x^2 - x^4 - y^2$ and the domain $-2 \leq x \leq 2$, $-4 \leq y \leq 4$. (See exercises 5 and 6 in the previous section, page 528.)

65. Continuation. Add to your sketch in the previous exercise a contour plot of the function $f(x, y) = 2x + 4x^2 - x^4 - y^2$ and confirm that each gradient vector is perpend1 of 82icular to the contour that passes through its base. (Note: most vectors have no contour passing through their bases, so you have to infer the shape and position of such a contour from the contours that *are* drawn.)

66. Draw a plausible set of contour lines for the function whose gradient vector field is plotted on the left below.

67. Draw a plausible gradient vector field for the function whose contour plot is shown on the right below. H marks a local maximum and L a local minimum.





9.3 Optimization

The contexts for optimization

Optimization is the process of making the best choice from a range of possibilities. ("Optimum" is the Latin word for *best.*) We are all familiar with optimization in the economic arena: managers of an enterprise typically seek to to maximize profit or minimize cost by making conscious choices. It is perhaps more surprising to learn that we sometimes use the same language to describe physical processes. For instance, the atoms in a molecule are arranged so that their total energy is minimized. A light ray travels from one point to another along the path that takes the least time. Of course atoms and photons don't make conscious choices. Nevertheless, the imagery of optimization is so vivid and useful that we try to invoke it whenever we can.

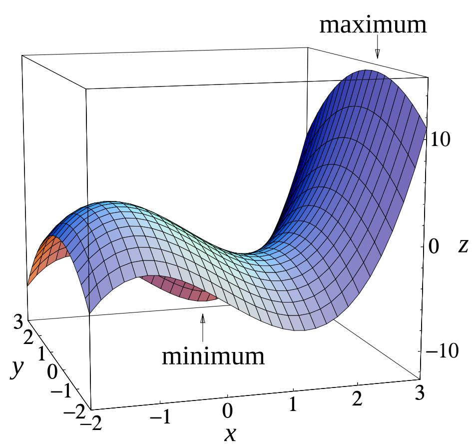
Constrained optimization

Usually, there is a restriction—called a **constraint**—on the choices that can be made to achieve to best possible outcome. For instance, consider a factory that makes tennis rackets. We can expect that the factory managers are instructed to minimize cost *while producing a given number of rackets.* This is their constraint. Production cost is a function of many quantities that the managers can control—the number of workers, the wage scale, and the cost of the raw materials are just a few. When the managers choose values for these quantities that minimize the cost function, they must be sure those values will also satisfy the constraint.

Mathematical optimization

In mathematical terms, optimization is the process of finding the minimum or maximum value of a function. The presence of constraints complicates this task, as you shall see.

Visual Inspection



The maximum value of a function is the highest point on its graph; the minimum value is the lowest. Shown at the right is the graph of $z = x^3 - 4x - y^2$ on the domain

$$-2 \le x \le 3 \qquad -2 \le y \le 3.$$

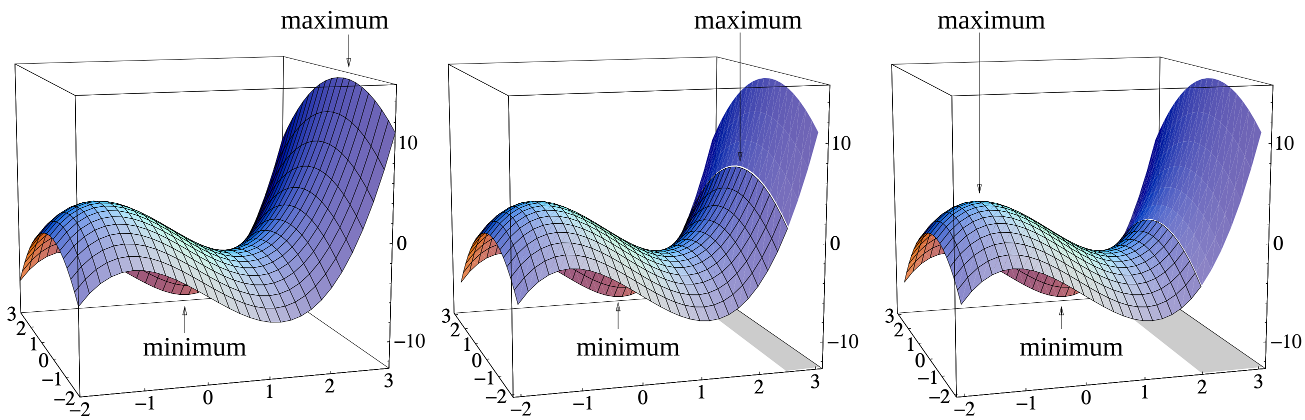
The maximum occurs where $(x, y) = (3, 0)$, and it has the value $z = 15$. The minimum is $z \approx -12.1$, when $(x, y) \approx (1.2, 3)$. Confirm this yourself by inspecting the graph and then calculating z.

Extremes, and local extremes

We'll use the term **extreme** to refer to either a maximum or a minimum. In the example we see several *local* extremes. These are points where the value of the function is larger or smaller than it is at any *nearby* point. There are local minima at both left-hand corners of the graph, at $(-2, -2)$ and $(-2, 3)$. There is another along the front edge, near $(1.2, -2)$. There is a local maximum in the interior, near $(-1.2, 0)$. To decide which local minimum is the *true* minimum, we must simply look. It is the same for local maxima.

The domain is a constraint

In our example, the domain of definition of the function acts as a *constraint*. If we change the domain, the positions of the extreme points can change. For example, suppose we change the position of the right-hand border in stages: first, $x \le 3$; second, $x \le 2.5$; third, $x \le 2$. (Since the graph itself doesn't change, we use a grid to show the part of the graph that satisfies



$x \le 3$ $\quad$ $x \le 2.5$ $\quad$ $x \le 2$

the constraint in each case below.) At the start, the maximum is on the boundary. When we first move the boundary to the left (from $x = 3$ to $x = 2.5$), the maximum just moves with it. However, when we move the boundary farther (from $x = 2.5$ to $x = 2$), the maximum jumps from the boundary to an interior point (near $(-1.2, 0)$). During these changes the minimum is not affected. It stays at the same place.

The maximum moves —it even jumps

Catastrophes

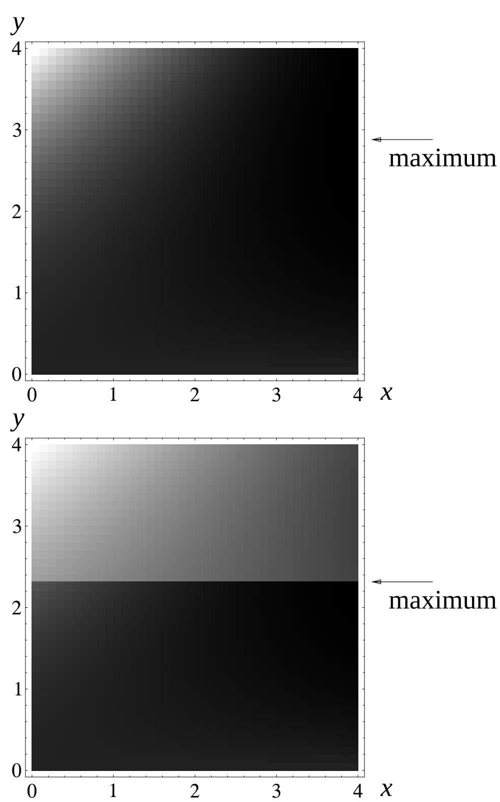
The sudden jump in the maximum is called a **catastrophe**. The figures explain what happens. We impose a constraint $x \le a$, and then we reduce the value of the parameter a. At first, the maximum is at the boundary point $(a, 0)$. The position of this point changes smoothly with a, causing a gradual drop in the value of the maximum. Eventually, the maximum reaches the same value as the interior local maximum. (This happens when $a = 4/\sqrt{3}$;

see the exercises.) If a continues to decrease, the local maximum at $(a, 0)$ then has a lower value than the local maximum at the interior point, so the true maximum jumps to the interior.

There are many variations on this pattern. Whenever a function depends on a parameter the positions of its extremes do, too. There are many ways for the position of an extreme to jump suddenly *while the parameter is changing gradually*. Any such jump is called a catastrophe.

Catastrophes make the task of optimization more interesting. If the maximum of a certain function gives the optimal solution to a problem, and that maximum jumps to a new position, then the optimal solution changes radically. For example, suppose the problem is to determine the minimum-energy configuration of atoms in a molecule. When the minimum jumps catastrophically, there is a new configuration of the same atoms, producing an *isomeric* form of the molecule.

The quest for an optimum does not have to involve mathematical tools. A good example is John Stuart Mill's philosophy of "the greatest good for the greatest number". In politics a catastrophe is called a revolution. Even scientific research pursues an optimum in raising the question: "What is the best way to explain certain phenomena?" The consensus in the scientific community can change catastrophically, in what is called a *paradigm shift* or an intellectual revolution. The geological theory of plate tectonics is a familiar example. Though proposed in the 1920s, it was dismissed until the 1960s, when it was suddenly and overwhelmingly accepted.



Density plots

We can also solve optimization problems by inspecting density plots. Suppose z is the yield from a process that is controlled by two inputs x and y, and

$$z = 3xy - 2y^2.$$

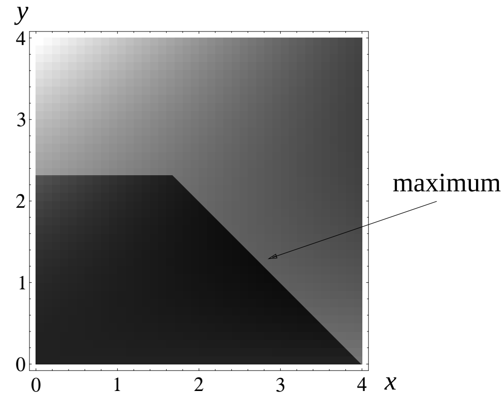
Initially we take $0 \leq x \leq 4$, $0 \leq y \leq 4$. The maximum yield is at the darkest spot in the upper density plot. It occurs on the right boundary, near $y = 3$.

The position of the maximum is subject to change if we have to impose further constraints. For instance, suppose the resource y is more limited than we first assumed, requiring us to set $y \leq 2.3$. With this added constraint we see that the maximum shifts to the corner $(x, y) = (4, 2.3)$. (The points shown in a lighter gray are the ones that have been removed from consideration by the new constraint.)

Besides limits on individual resources, we are often faced with a limit on *total* resources. In our case, let's suppose the limit has the form

$$x + y \leq 4.$$

That means all points above the line $x + y = 4$ must be removed from consideration. (They are shown in lighter gray.) This new constraint causes the maximum to shift yet again. It now appears near the point $(x, y) = (3, 1)$.

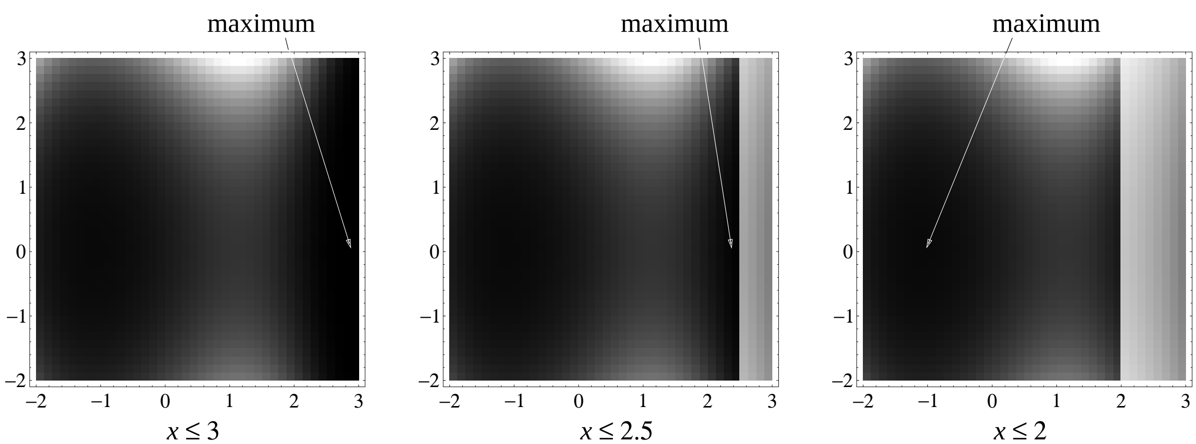


Constraints reduce optimality

As we add constraints that force the maximum to move, the density at each new maximum is less than it was at the previous one. Thus, the value of the maximum itself decreases. In other words, each added constraint makes the optimal solution slightly "less optimal" than it had been. This is only to be expected.

Extremes on the interior of a density plot

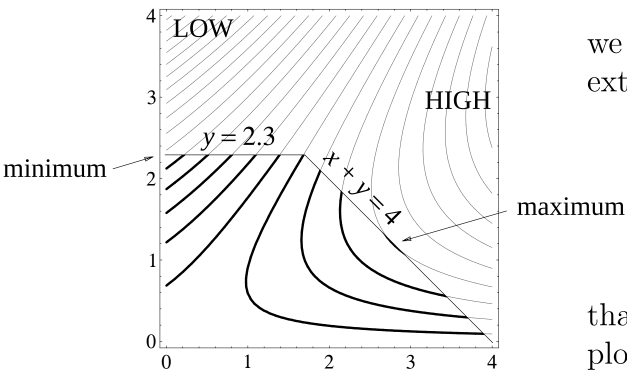
Of course the extremes may appear in the interior of the domain as well as on the boundary—and density plots can show this. Here are the density plots that correspond to the graphs of $z = x^3 - 4x - y^2$ that we saw on page 565. The maximum jumps to the interior when the value of a in the constraint $x \leq a$



drops below $a = 4/\sqrt{3}$. In all three plots the minimum appears as the bright spot at the top of the rectangle near where $x = 1$. (The exact location is $(x, y) = (2/\sqrt{3}, 3)$.) We can also see a local minimum at the bottom of the rectangle near $x = 1$. From the graph we know there are two more at the left corners of the rectangle, but they are harder to notice in the density plots.

Contour plots

A density plot is useful for showing the general location of the highs and lows, but we can get a more precise picture by switching to a contour plot. Let's look at the contour plots of the same problems we just analyzed using density plots.



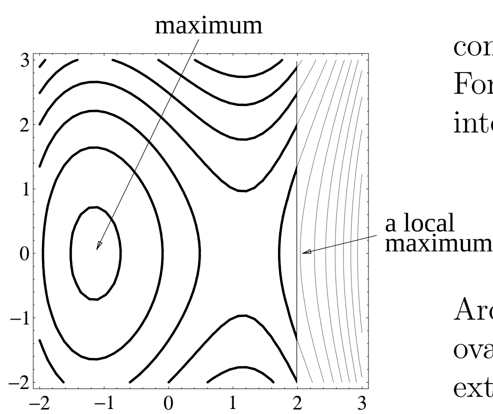
We'll start with the function $z = 3xy - 2y^2$ we first considered on page 566 and search for its extremes subject to the constraints

$$0 \le x,$$
$$0 \le y \le 2.3,$$
$$x + y \le 4,$$

that we introduced earlier. From the density plot on page 566 it was obvious that the values of z increase steadily from the upper left corner of the large square to the upper part of the right side. In the contour plot, though, we need some sort of labels to show us where the low and high values of z are to be found.

An extreme can occur where a contour is tangent to the boundary

To locate the extremes within the constrained region, we need to find contours that carry the lowest and the highest values of z. We can do this quite precisely. The contour that just passes through the upper left corner—and meets the region at that point alone—carries the lowest value of z. If you study the plot you can see that the contour carrying the *highest* value of z also touches the boundary at just a single point. It is the contour that is tangent to the line $x + y = 4$, and elsewhere lies outside the constrained region.



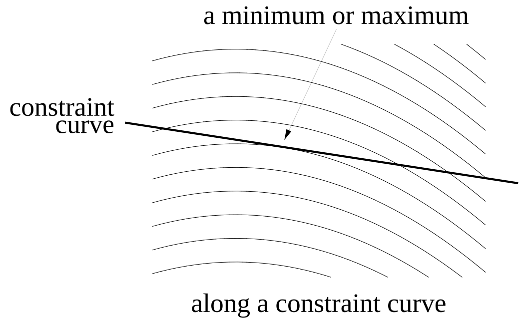
Suppose the extreme is in the *interior* of the constrained region, rather than on the boundary. For example, the function $z = x^3 - 4x - y^2$ has an interior maximum when

$$-2 \le x \le 2$$
$$-2 \le y \le 3.$$

Around the maximum there is a nest of concentric ovals. This pattern is characteristic for an interior extreme. Notice the local maximum where a contour is tangent to the boundary line $x = 2$.

Contour patterns at an extreme

These examples demonstrate that there are characteristic patterns that contour lines make near an extreme point. One pattern appears along a constraint curve; another pattern appears at interior points of a domain. We first



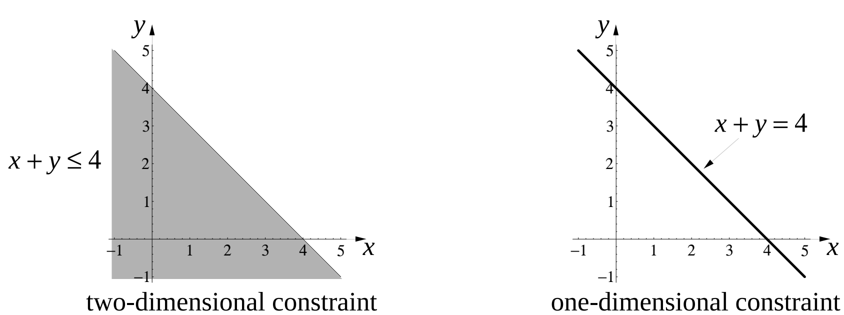
Patterns of contour lines at an extreme point

met the pattern that appears at an interior extreme point when we discussed the 'standard' minimum ($x^2 + y^2$) and maximum ($-x^2 - y^2$) in section 1 (pages 519–522). **Caution**: the patterns you see here are "typical", but they do not *guarantee* the presence of an extreme point. The exercises give you a chance to explore some of the subtleties.

Dimension-reducing Constraints

How a constraint can reduce the dimension of a problem

Constraints appear frequently in optimization problems. Thus far, however, the constraints we considered have been described by *inequalities*, like $x+y \leq 4$. Initially, x and y give us the coordinates of a point in a two-dimensional plane. The effect of the constraint $x + y \leq 4$ is to restrict the points (x, y) to just a part of that plane—but it is still a *two-dimensional* part. Sometimes, though, the constraint is given by an *equality*, like $x + y = 4$. In that case, the points (x, y) are restricted to lie on a line—which is a one-dimensional set in the plane. The second constraint therefore reduces the dimension of the problem.



A standard form

There is a standard form for any constraint that reduces a two-dimensional optimization problem to a one-dimensional problem. The form is this:

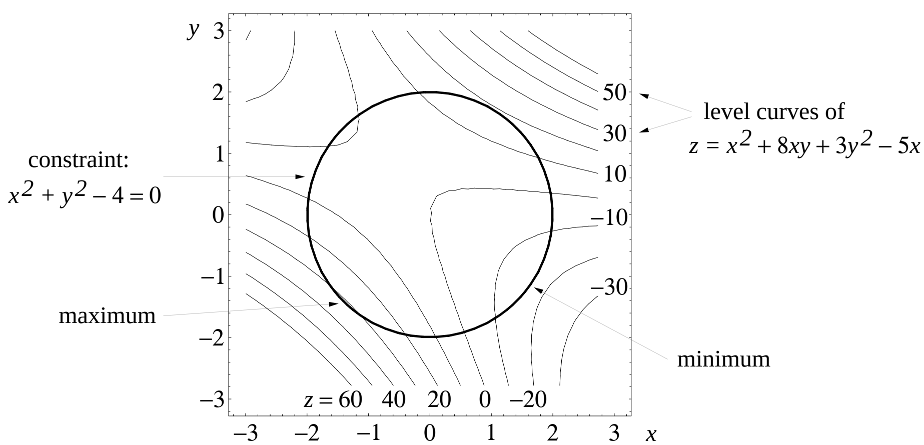
$$\text{constraint:} \quad g(x, y) = 0.$$

For instance, the constraint $x + y = 4$ can be written this way by setting $g(x, y) = x + y - 4$. We'll look at another example in a moment.

A comment on dimension

First, though, notice that $g(x, y) = 0$ is one of the contour lines of the function $g(x, y)$, namely, the contour at level zero. Since the contours of g are curves, we often call $g(x, y) = 0$ a **constraint curve**. This curve is *one*-dimensional, even though it might twist and turn in a two-dimensional plane. We say the curve is one-dimensional because it looks like a straight line under a microscope. Likewise, a curved surface is two-dimensional because it looks like a flat plane under a microscope.

Example. Find the extreme values of $f(x, y) = x^2 + 8xy + 3y^2 - 5x$ subject to the constraint $g(x, y) = x^2 + y^2 - 4 = 0$. The constraint curve is a circle of radius 2, and the level curves of $z = f(x, y)$ form a set of hyperbolas. We need to find the highest and the lowest z-levels on the constraint curve.



Finding the highest and lowest levels on the constraint curve

The z-levels are labelled around the right and the bottom edges of the figure. It is in the third quadrant, near the point $(-1.4, -1.5)$, that the constraint curve meets the highest z-level. At that point z is slightly more than 30. The constraint curve is evidently tangent to the contour there. The lowest z-level that the constraint curve meets is about $z = -16$, near the point $(1.6, -1.2)$. Picture in your mind how the contours between $z = -10$ and $z = -20$ fit together. Can you see that the constraint curve is tangent to contour at the minimum?

Example, continued. There is still more to say about the way the constraint $x^2+y^2-4=0$ reduces the dimension of our problem. First of all, we can describe the coordinates of any point on the constraint circle by using the circular functions:

Locate points on the constraint circle with a single variable t

$$x = 2\cos t, \qquad y = 2\sin t.$$

We need the factor 2 because the circle has radius 2. These equations mean that x and y are now functions of a *single* variable, t. Next, consider the function

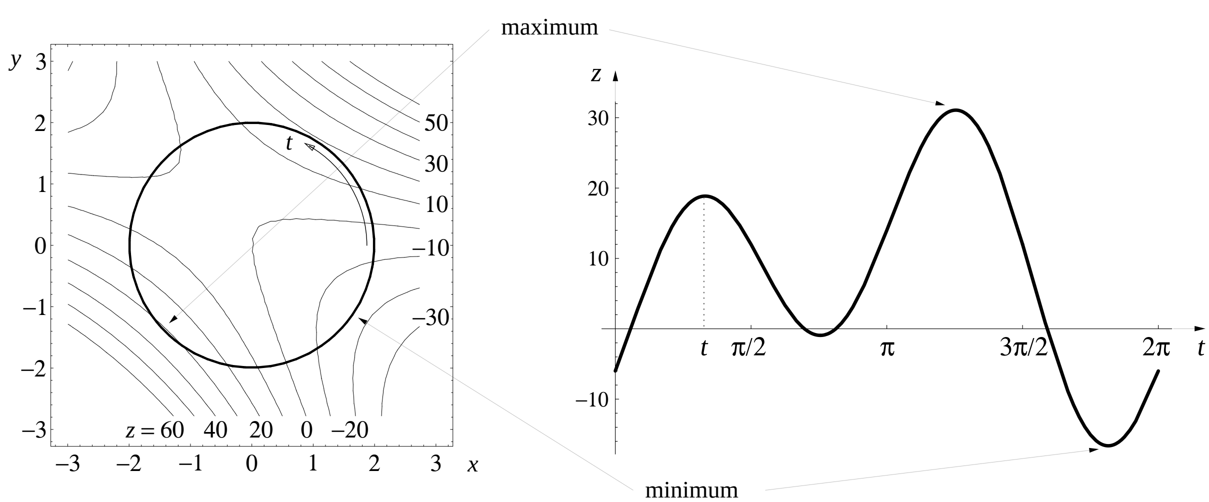
$$z = f(x,y) = x^2 + 8xy + 3y^2 - 5x$$

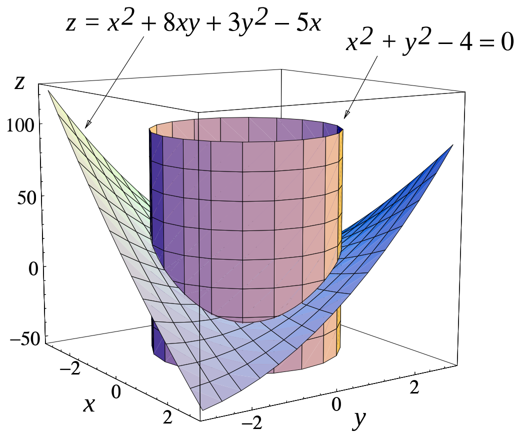
that we seek to optimize *subject to the constraint*. But the constraint makes x and y functions of t. Thus, *when we take the constraint into account*, z itself becomes a function of t:

z becomes a function of t

$$\begin{aligned} z &= f(2\cos t, 2\sin t) \\ &= (2\cos t)^2 + 8\,(2\cos t)(2\sin t) + 3(2\sin t)^2 - 5(2\cos t). \end{aligned}$$

The graph of *this* function is thus just an ordinary curve in the t,z-plane. It is shown on the right, below. A value of t determines a point on the circle, as shown in the contour plot on the left. (We have taken $t \approx \pi/3$.) The value of z at that point then determines the height of the graph on the right. As you can see, our chosen value of t puts z near a local maximum.

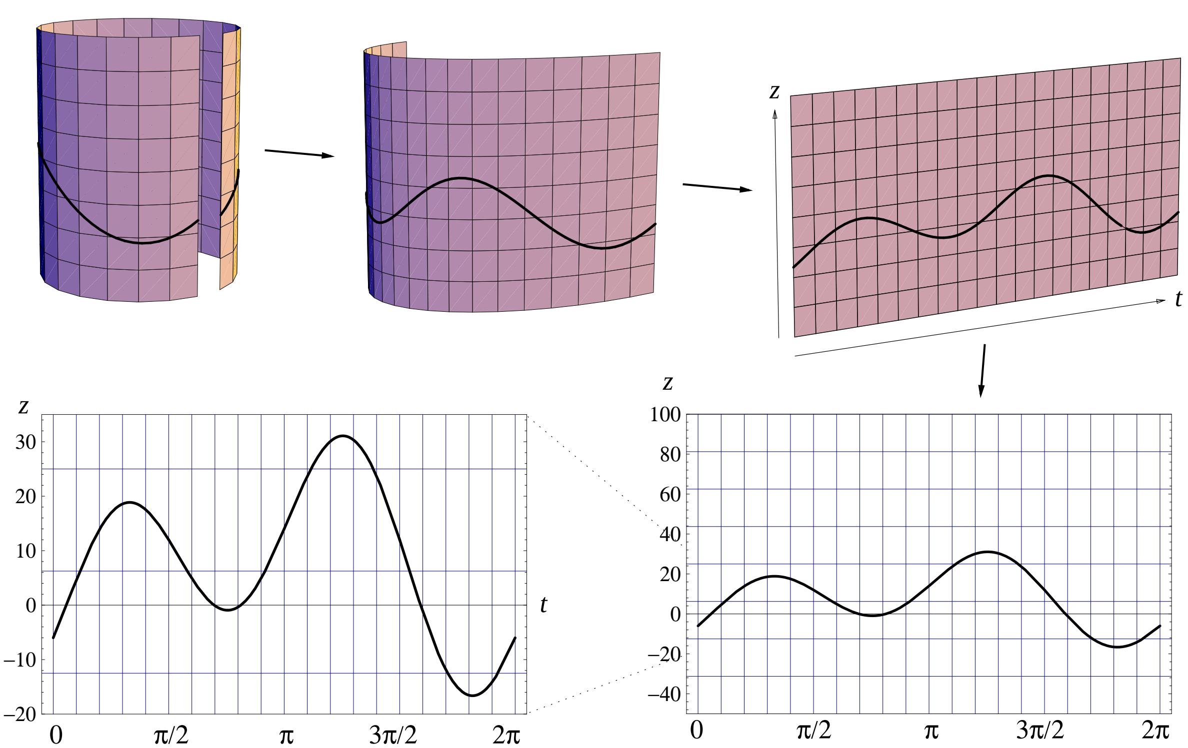




Example, continued. Let's look at the graph of $z = x^2 + 8xy + 3y^2 - 5x$. The constraint tells us that we should look *only* at the points on the graph that lie above the circle $x^2 + y^2 - 4 = 0$. These are the points where the graph intersects the cylinder that you see at the right. The intersection is a curve that goes up and down around the cylinder. At some point on the curve z has a maximum, and at some other point it has a minimum. (In fact, both the maximum and the minimum are visible in this view.)

Unwrap the cylinder

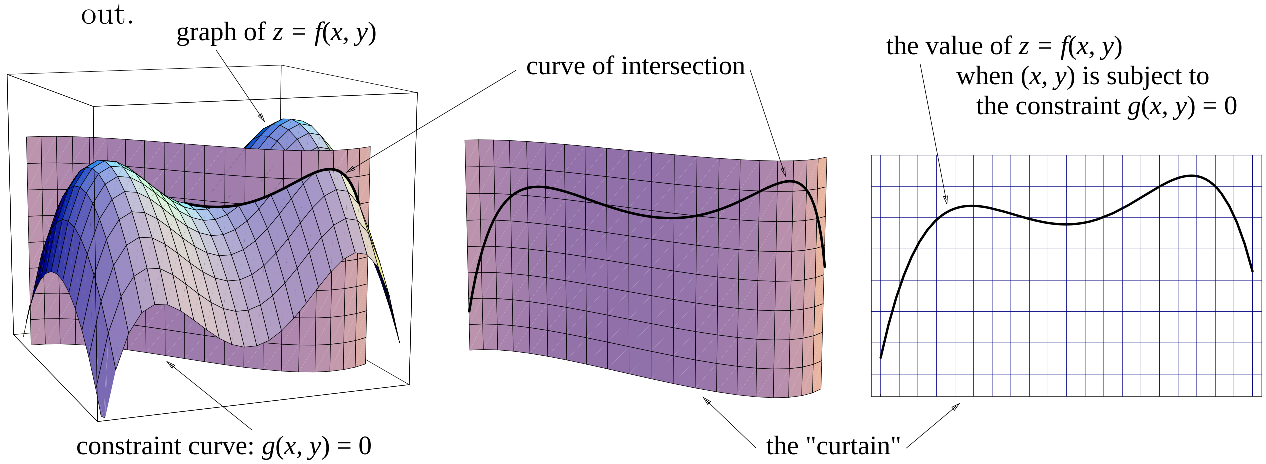
Since the intersection curve lies on the cylinder, we can get a better view of the curve if we slit open the cylinder and unwrap it. Follow the sequence clockwise from the upper left. We can use coordinates to describe the curve on the flattened cylinder. The t variable takes us around the cylinder, so it becomes the horizontal coordinate. The z variable measures vertical height. The z-range in the figure above is larger than we need: it goes from -50 to $+100$. In the bottom row on the left we have rescaled the z-axis so it runs from -20 to $+35$. Compare this graph with the one on the opposite page.



The example shows us how a constraint works to reduce the dimension of a problem in general. Suppose we want to maximize the value of the function $z = f(x, y)$, subject to the constraint $g(x, y) = 0$. Then:

A general view of the dimension-reducing effect of a constraint

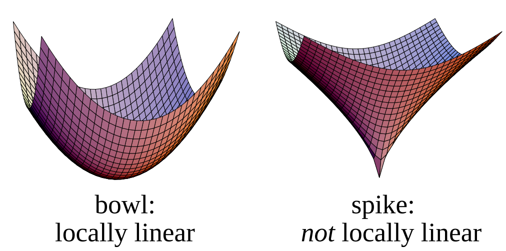
- *Without* the constraint, we would look for the highest point on a *two*-dimensional surface in three-dimensional space.
- *With* the constraint, we would restrict our search to the vertical "curtain" that lies above the constraint curve. The graph intersects this curtain in a curve, so we end up looking for the highest point on a *one*-dimensional curve in a two-dimensional plane. We can think of this plane as the curtain after it has been unwrapped and straightened out.



This is just a picture of the relation between the function and the constraint. We may still have to determine *analytically* the form that the function $f(x, y)$ takes when we impose the constraint $g(x, y) = 0$. You can find a number of possibilities in the exercises.

Extremes and Critical Points

Suppose that a function $f(x, y)$ has a maximum or a minimum at an *interior* point (a, b). Suppose also that the function is *locally linear* at (a, b), so we have a "bowl" rather than a "spike" or some other irregularity in the graph. Then we must have



$$\operatorname{grad} f(a, b) = \left(\frac{\partial f}{\partial x}(a, b), \frac{\partial f}{\partial y}(a, b) \right) = (0, 0).$$

A proof

Here is why. The gradient vector grad$f(a, b)$ tells us how the graph of $z = f(x, y)$ is tilted at the point (a, b). (We discuss the geometric meaning of the gradient on page 551.) At a maximum or a minimum, though, the graph is *not* tilted; it must be flat. Therefore, the gradient must be the zero vector.

Moving to an arbitrary number of input variables

The ideas here carry over to functions that have any number of input variables. First, we give a name to a point where the gradient is zero.

> **Definition**. A **critical point** of a locally linear function is one where the gradient vector is zero. Equivalently, all the first partial derivatives of the function are zero.

The observation we just made can now be restated as a theorem that connects extreme points and critical points.

> **Theorem**. If a locally linear function has a maximum or a minimum at an interior point of its domain, then that point must be a critical point.

A statement and its converse

The direction of the implication in this theorem is important. Here is the theorem, written in a very abbreviated form:

$$\textit{statement}: \qquad \text{extreme} \Longrightarrow \text{critical}.$$

When we reverse the direction of the implication, we get a new statement, abbreviated the same way:

$$\textit{converse}: \qquad \text{critical} \Longrightarrow \text{extreme}.$$

The converse of *this* theorem is not true

The converse says that a critical point must be an extreme point. But that is just not true. For example, an ordinary saddle point (a minimax) is a critical point, but it is not a minimum or a maximum.

Searching critical points for extremes

The theorem and the observation about its converse are both important in the *optimization process*—that is, the search for extremes. Together they offer us the following guidance:

- Search for the extremes of a function among its critical points.
- A critical point may be neither a maximum nor a minimum.

To see how we can find extremes by searching among the critical points of a function, we'll do a few examples. All the examples use the same basic idea. However, as the details get more complicated we bring in more powerful techniques.

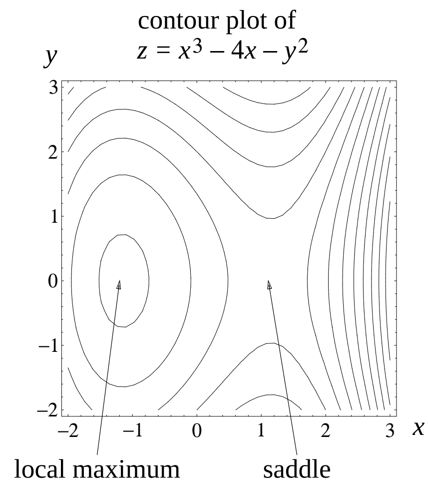
Example 1. We'll start with the function

$$z = f(x, y) = x^3 - 4x - y^2$$

we have frequently used as a test case in this chapter.

The critical points of f are the points that *simultaneously* satisfy the two equations

$$\frac{\partial f}{\partial x} = 3x^2 - 4 = 0,$$
$$\frac{\partial f}{\partial y} = -2y = 0.$$



It is clear that $y = 0$ and $x = \pm\sqrt{4/3} \approx \pm 1.1547$. The two critical points are therefore

$$\left(+\sqrt{4/3}, 0\right) \qquad \text{and} \qquad \left(-\sqrt{4/3}, 0\right).$$

If you check the contour plot you can see that $(-\sqrt{4/3}, 0)$ is a local maximum and $(+\sqrt{4/3}, 0)$ is a saddle point.

Example 2. Here is a somewhat more complicated function:

$$g(x, y) = 2x^2y - y^2 - 4x^2 + 3y.$$

The critical points are the solutions of the equations

$$\frac{\partial g}{\partial x} = 4xy - 8x = 0, \qquad \frac{\partial g}{\partial y} = 2x^2 - 2y + 3 = 0.$$

Algebraic methods will still work, even though both variables appear in both equations. For example, we can rewrite $\partial g/\partial y = 0$ as

$$2y = 2x^2 + 3 \qquad \text{or} \qquad y = x^2 + \tfrac{3}{2}.$$

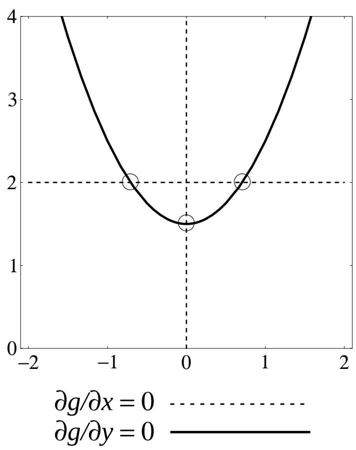
We can then substitute this expression for y into $\partial g/\partial x = 0$ and get

$$4x\left(x^2 + \tfrac{3}{2}\right) - 8x = 4x\left(x^2 + \tfrac{3}{2} - 2\right) = 4x\left(x^2 - \tfrac{1}{2}\right) = 0.$$

This implies $x = 0$ or $x = \pm\sqrt{1/2}$. For each x we can then find the corresponding y by from the equation $y = x^2 + \frac{3}{2}$.

Let's work through this a second time using a geometric approach. The equations $\partial g/\partial x = 0$ and $\partial g/\partial y = 0$ both define curves in the x, y-plane. The curve $\partial g/\partial y = 0$ is a parabola: $y = x^2 + \frac{3}{2}$. The equation $\partial g/\partial x = 0$ factors as

$$4x(y-2) = 0 \qquad \text{or} \qquad x(y-2) = 0.$$

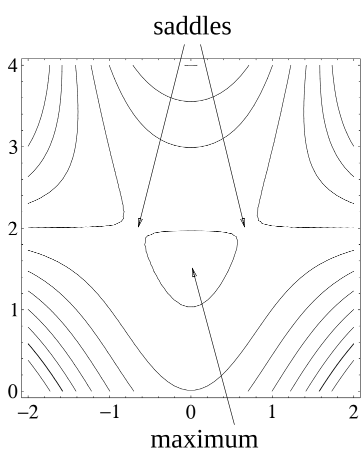


Now a product equals 0 precisely when one of its factors equals 0, so $\partial g/\partial x = 0$ implies that *either* $x = 0$ *or* $y - 2 = 0$. In other words, the "curve" $\partial g/\partial x = 0$ consists of two lines:

$$\begin{aligned} x &= 0, \quad \text{a vertical line,} \\ y &= 2, \quad \text{a horizontal line.} \end{aligned}$$

The two curves are shown in the figure at the right. They intersect in three points. (The place where the horizontal and vertical lines cross is *not* one of the intersection points.)

One of the immediate benefits of the geometric approach is to make it clear that the y-coordinate of a critical point is either 2 or $\frac{3}{2}$. The critical points are therefore



$$\left(-\sqrt{1/2}, 2\right), \quad \left(0, \tfrac{3}{2}\right), \quad \left(+\sqrt{1/2}, 2\right).$$

A glance at the contour plot of g makes it clear that the first and third of these are saddle points. The middle point is a local maximum. There are several ways to determine this. One is to look at the graph of g. Another is to look at a vertical slice of the graph through the line $x = 0$. Then $z = g(0, y) = -y^2 + 3y$. Here z has a maximum when $y = \frac{3}{2}$.

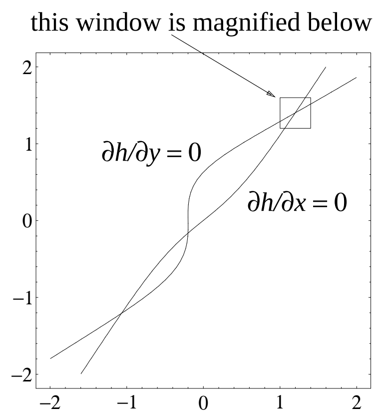
We first identified the local maximum of the function $z = x^3 - 4x - y^2$ on page 565. At the time, though, we could only estimate its position by eye. Now, however, we can specify its location exactly, because we have analytical tools for finding critical points. As the next example shows, these tools are useful even when we can't carry out the algebraic manipulations.

Example 3. Here is a function whose critical points we *can't* find using just algebraic computations:

$$z = h(x, y) = x^2y^2 - x^4 - y^4 - 2x^2 + 5xy + y$$

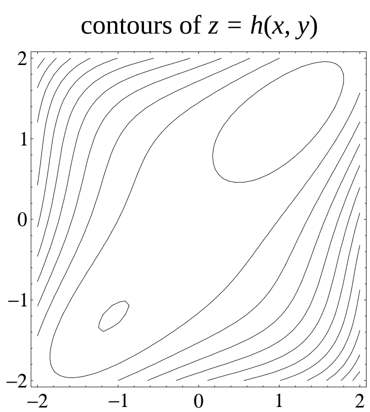
The equations for the critical points are

$$\frac{\partial h}{\partial x} = 2xy^2 - 4x^3 - 4x + 5y = 0,$$
$$\frac{\partial h}{\partial y} = 2x^2y - 4y^3 + 5x + 1 = 0.$$

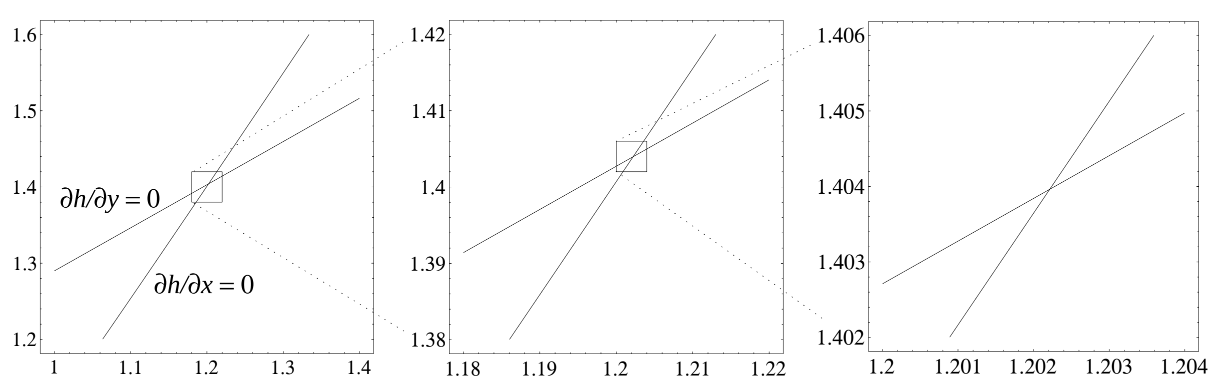


Even though we can't solve the equations algebraically, we can plot the curves they define by using the contour-plotting program of a computer. This is done at the right. The curves intersect in three points. (If you draw the plots on a large scale, you will find that these are still the only intersections.)

A contour plot of $z = h(x, y)$ itself reveals that the middle point is a saddle and the outer two are local maxima. Let's focus on the maximum in the upper right and determine its position more precisely.



We can always use a microscope. But if we magnify the contour plot, we just get a set of nested ovals. The maximum would lie somewhere inside the smallest—but we wouldn't know quite where. By contrast, the curves $\partial h/\partial x = 0$ and $\partial h/\partial y = 0$ give us a pair of "crosshairs" to focus on. Even with relatively little magnification we can see that the maximum is at $(1.202\ldots, 1.404\ldots)$.



The Method of Steepest Ascent

Walk uphill to get to a maximum

Here is yet another way to find the maximum of a function $z = f(x,y)$. Imagine that the graph of f is a landscape that you're standing on. You want to get to the highest point. To do that you just walk uphill. But the uphill direction on the landscape is given by the gradient vector field of f (see page 551). So you move to higher ground by following the gradient field.

Trajectories of the gradient dynamical system...

In fact, the gradient field defines a **dynamical system** of exactly the sort we studied in chapter 8. The differential equations are

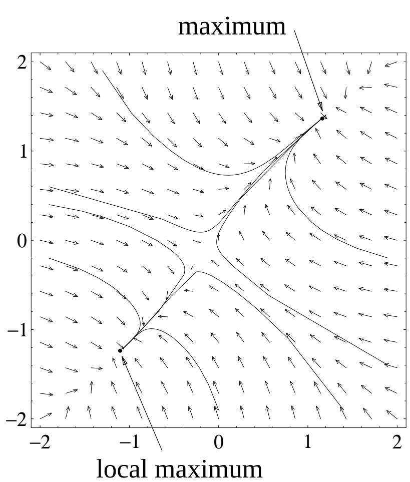
$$\frac{dx}{dt} = \frac{\partial f}{\partial x}(x,y),$$
$$\frac{dy}{dt} = \frac{\partial f}{\partial y}(x,y).$$

...lead to the local maxima

Because the gradient points uphill, the trajectories of this dynamical system also go uphill. Trajectories flow to the attractors of the system; these are the local maxima of the function. Furthermore, since the gradient points in the direction f increases *most rapidly*, the trajectories follow paths of **steepest ascent** to the maxima. This explains the name of the method.

Example 1. Let's see how the method of steepest ascent will find the local maxima of the function

$$z = h(x,y) = x^2y^2 - x^4 - y^4 - 2x^2 + 5xy + y$$



we considered in the previous example. The gradient field is

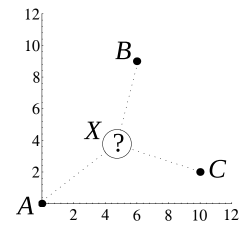
$$\frac{dx}{dt} = 2xy^2 - 4x^3 - 4x + 5y.$$
$$\frac{dy}{dt} = 2x^2y - 4y^3 + 5x + 1.$$

As you can see at the left, some of the trajectories flow to a local maximum near $(-1,-1)$, while other flow to the maximum whose position we determine on the opposite page. Each attractor has its own basin of attraction (as described in chapter 8). Therefore, the maximum found by the method of steepest ascent depends on the initial point of the trajectory.

If we replace the gradient vectors by their negatives, the new field will point directly *downhill*—to the local minima. Using the trajectories of the negative gradient field to find the minima is thus called the method of steepest descent. In the following example we use this method to investigate an economic question.

The method of steepest *descent*

Example 2. Manufacturing companies ship their products to regional warehouses from large distribution centers. Suppose a company has regional warehouses at A, B, and C, as shown on the map at the right. Where should it put its distribution center X so as to minimize the total cost of supplying the three regional warehouses?



This is a complicated problem that depends on many factors. For example, X probably should be put near major roads. The managers may also want to choose a location where labor costs are lower. Certainly, the total distance between the center and the three warehouses is important. Let's simply get a *first approximation* to a solution by concentrating on the last factor. We will find the position for X that minimizes the total straight-line distance (the distance "as the crow flies") from X to the three points A, B, and C. The map shows these distances as three dotted lines.

Minimize the total distance

To describe the various positions we have introduced a coordinate system in which

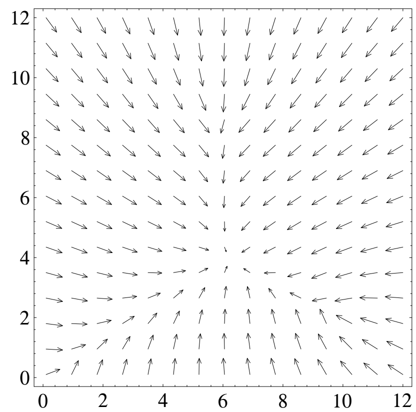
$$A: \ (0,0), \qquad B: \ (6,9), \qquad C: \ (10,2).$$

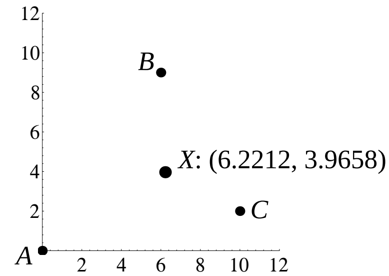
The coordinates here are arbitrary. That is, they don't represent miles, or kilometers, or any of the usual units of distance—but they are *proportional* to the usual units, so we can measure with them. If we let the unknown position of X be (x, y), then we seek to minimize the function

$$S(x,y) = \sqrt{x^2+y^2} + \sqrt{(x-6)^2+(y-9)^2} + \sqrt{(x-10)^2+(y-2)^2}.$$

According to the method of steepest descent, we want to find the attractor of this dynamical system:

$$\frac{dx}{dt} = -\frac{x}{\sqrt{x^2+y^2}} - \frac{x-6}{\sqrt{(x-6)^2+(y-9)^2}} - \frac{x-10}{\sqrt{(x-10)^2+(y-2)^2}},$$
$$\frac{dy}{dt} = -\frac{y}{\sqrt{x^2+y^2}} - \frac{y-9}{\sqrt{(x-6)^2+(y-9)^2}} - \frac{y-2}{\sqrt{(x-10)^2+(y-2)^2}}.$$





the optimum location for
the distributiuon center

The attractor is a global minimum

As you can see from the vector field of the dynamical system (above left), there is a single attractor, near the point $(6, 4)$. This implies the total distance function $S(x, y)$ has a single global minimum—which is what our intuition about the problem would lead us to expect.

To find the position of X more exactly, you can do the following. First, obtain a solution $(x(t), y(t))$ to the system of differential equations with an arbitrary initial condition—for example,

$$x(0) = 1, \qquad y(0) = 1.$$

The attractor is the limit point of a solution

Then, obtain the coordinates of the attractor by evaluating $(x(t), y(t))$ for larger and larger values of t, stopping when the values of $x(t)$ and $y(t)$ stabilize. You will find that

$$X = \lim_{t\to\infty} (x(t), y(t)) = (6.22120\ldots, 3.96577\ldots).$$

The method needs only a differential equation solver

The important point to note here is that it is not necessary to plot the vector field—or any other graphic aid, like a contour plot or graph. You simply need to solve a system of differential equations. For example, the values above were found by modifying the computer program SIRVALUE we introduced in chapter 2. In summary: *the method of steepest descent (or ascent) requires no graphical tools, but only a basic differential equation solver.*

Lagrange Multipliers

In searching for the interior extremes of a function $f(x, y)$, we have seen that it is helpful to solve the critical point equations:

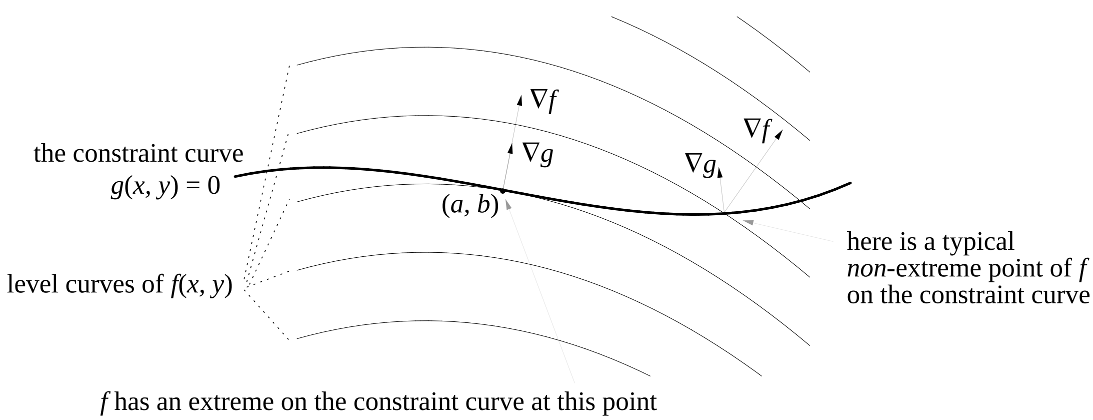
$$\frac{\partial f}{\partial x} = 0, \qquad \frac{\partial f}{\partial y} = 0.$$

There is a similar set of equations we can use in the search for the *constrained* extremes of f. These equations involve a new variable called a Lagrange multiplier.

Equations for a constrained extreme

So, suppose the function $f(x, y)$ has an extreme on the constraint curve $g(x, y) = 0$ at the point (a, b). According to the following diagram, the gradient vectors ∇f and ∇g must be parallel at (a, b). (This means they are in the same direction or in opposite directions.) Here is why. We already know (see page 569) that the level curve of f that passes through the constrained maximum or minimum must be tangent to the constraint curve. Now, the gradient vector ∇f at any point is perpendicular to the level curve of f through that point—and the same is true for g. At a point where the level curves are tangent, the gradients ∇f and ∇g are perpendicular to the *same* curve, and must therefore be parallel.

∇f and ∇g must be parallel at a point where f has an extreme along the constraint curve $g = 0$



Parallel vectors are multiples of each other. Specifically, at a point (a, b) where ∇f and ∇g are parallel, there must be a number λ for which

The multiplier equation

$$\nabla g(a, b) = \lambda \cdot \nabla f(a, b).$$

The multiplier λ is called a **Lagrange multiplier**. In the figure above, $\lambda \approx 1/2$. If ∇g and ∇f were in *opposite* directions, then λ would be negative.

Joseph Louis Lagrange (1736–1813) was a French mathematician and a younger contemporary of Leonhard Euler. Like Euler he played an important role in making calculus the primary analytical tool for study the physical world. He is particularly noted for his contributions to celestial mechanics, the field where Isaac Newton first applied the calculus.

If we write out the multiplier equation using the components of ∇f and ∇g, we get

$$\left(\frac{\partial g}{\partial x}(a,b), \frac{\partial g}{\partial y}(a,b)\right) = \lambda\left(\frac{\partial f}{\partial x}(a,b), \frac{\partial f}{\partial y}(a,b)\right) = \left(\lambda\,\frac{\partial f}{\partial x}(a,b), \lambda\,\frac{\partial f}{\partial y}(a,b)\right).$$

In a vector equation, the vectors are equal component by component. Thus,

$$\frac{\partial g}{\partial x}(a,b) = \lambda\,\frac{\partial f}{\partial x}(a,b)$$
$$\frac{\partial g}{\partial y}(a,b) = \lambda\,\frac{\partial f}{\partial y}(a,b)$$

How can we find a constrained maximum or minimum?

Let's return to the main question, which can be stated this way: How do we determine where the function $f(x,y)$ has a maximum or a minimum, subject to the constraint $g(x,y) = 0$? If we let (a,b) denote the point we seek, then we see that a and b satisfy three equations:

$$\text{the constraint equation}: \qquad g(a,b) = 0,$$
$$\text{the multiplier equations}: \quad \begin{cases} \dfrac{\partial g}{\partial x}(a,b) = \lambda\,\dfrac{\partial f}{\partial x}(a,b), \\[2mm] \dfrac{\partial g}{\partial y}(a,b) = \lambda\,\dfrac{\partial f}{\partial y}(a,b). \end{cases}$$

In fact, there are *three* unknowns in these equations: a, b, and λ. When we solve the three equations for the three unknowns, we will determine the location of the constrained extreme. (We'll also have a piece of information we can throw away: the value of λ.)

Example. Find the maximum of $f(x,y) = x^p y^{1-p}$ subject to the constraint $x + y = c$. There are two parameters in this problem: p and c. We assume that $0 < p < 1$ and $0 < c$. We introduce parameters to remind you that analytic methods (such as Lagrange multipliers) are especially valuable in solving problems that depend on parameters.

We let $g(x,y) = x + y - c$. Then

$$\nabla g = (1,1), \qquad \nabla f = \left(px^{p-1}y^{1-p}, (1-p)x^p y^{-p}\right),$$

so the three equations we must solve are

$$x + y - c = 0,$$
$$1 = \lambda p x^{p-1} y^{1-p},$$
$$1 = \lambda (1-p) x^p y^{-p}.$$

Since the second and third equations both equal 1, we can set them equal to each other:

$$\lambda p x^{p-1} y^{1-p} = \lambda (1-p) x^p y^{-p}.$$

We can cancel the two λs and combine the powers of x and y to get

$$p x^{-1} y = 1 - p \qquad \text{or} \qquad \frac{y}{x} = \frac{1-p}{p}.$$

According to the first of the three equations, $y = c - x$. If we substitution this expression in for y into the last equation, we get

$$\frac{c-x}{x} = \frac{1-p}{p} \qquad \text{or} \qquad p(c-x) = (1-p)x.$$

This equation reduces to $pc = x$, which gives us the x-coordinate of the maximum. To get the y-coordinate, we use $y = c - x = c - pc = c(1-p)$. To sum up, the maximum is at

$$(x, y) = (cp, c(1-p)) = c(p, 1-p).$$

Exercises

When searching for an extreme, be sure to zoom in on the graph or plot you are using as you narrow down the location of the point you seek.

1. Inspect the graph of $z = xy$ to find the maximum value of z subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad 3x + 8y \leq 120.$$

2. Inspect the graph of $z = 5x + 2y$ to find the minimum value of z subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad xy \geq 10.$$

3. Inspect a contour plot of $z = 3xy - y^2$ to find the maximum value of z subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad x + y \leq 5.$$

4. Continuation. Add the constraint $x \leq a$ to the preceding three, where a is a parameter that takes values between 0 and 5. Find the maximum value of z subject to all four constraints. Describe how the position of the constraint depends on the value of the parameter a.

[Answer: The maximum is found at $(x, y) = (25/8, 15/8)$, as long as $a \geq 25/8$. When $a < 25/8$, the maximum is at $(x, y) = (a, 5 - a)$.]

5. Find the maximum and minimum values of $z = 12x - 5y$ when (x, y) is exactly 1 unit from the origin.

How is the gradient involved here?

6. Find the maximum and minimum value of $z = px + qy$ when (x, y) is exactly 1 unit from the origin.

a) At what point is the maximum achieved; at what point is the minimum achieved?

7. Use a graph to locate the maximum value of the function

$$z = 2xy - 5x^2 - 7y^2 + 2x + 3y.$$

There are no constraints.

8. Use a graph to find the maximum value of $z = 6x + 12y - x^3 - y^3$, subject to the constraints

$$x \geq 0, \qquad y \geq 0, \qquad x^2 + y^2 \leq 100.$$

9. a) Locate the position of the minimum of $x^4 - 2x^2 - \alpha x + y^2$ as a function of the parameter α.

b) The position of the minimum jumps catastrophically when α passes through a certain value. At what value of α does this happen, and what jump occurs in the minimum?

10. a) Locate the maximum of $x^3 + y^3 - 3x - 3y$ subject to

$$x \leq 3, \qquad y \leq 0, \qquad x + y \leq \beta.$$

The position of the maximum depends on the value of the parameter β, which you can assume lies between 0 and 5.

b) The position of the maximum jumps catastrophically when β crosses a certain threshold value β_0. What is β_0?

11. Find the maximum value of x^2y in the first quadrant, subject to the constraint $x + 5y = 10$.

12. Find the maximum and minimum values of $z = 3x + 4y$ subject to the single constraint $x^2 + 4xy + 5y^2 = 10$.

13. Find all the critical points of the following functions.

a) $3x^2 + 7xy + 2y^2 + 5x - 6y + 3$.

b) $\sin x \sin y$ on the domain $-4 \le x \le 4$, $-4 \le y \le 4$.

c) $\sin xy$ on the domain $-4 \le x \le 4$, $-4 \le y \le 4$.

d) $\exp(x^2 + y^2)$.

e) $x^3 + y^3 - 3x - 3y$.

f) $x^3 - 3xy^2 - x^2 - y^2$. [There are four critical points; three are saddles.]

14. a) Find the nine critical points of the function

$$C(x, y) = (x^2 + xy + y^2 - 1)(x^2 - xy + y^2 - 1).$$

Four are minima, four are saddles, and one is a maximum.

b) Mark the locations of the critical points on a suitable contour plot of $C(x, y)$.

15. Locate and classify the critical points of the energy integral of a pendulum:

$$E(x, v) = 1 - \cos x + \tfrac{1}{2}v^2.$$

a) Compare the *critical points* of E with the *equilibrium points* of the dynamical system associated with this energy integral.

16. Use the method of steepest descent to find the minimum of the function

$$z = p(x, y) = e^{2+y-x^2-y^2} \sin x.$$

The distribution problem

The next two exercises are modifications of the distribution problem on pages 579–580. In the example we assumed that deliveries from the center X to each of the regional warehouse A, B, and C happened equally often. These exercises assume that deliveries to some warehouses are more frequent than others.

17. Suppose that one truck makes deliveries to A 5 times each week, while a second truck is used to make 3 deliveries to B and 2 to C each week. It makes sense to locate X so that the *total weekly travel* is minimized, rather than just the total distance to the three warehouses. To get the total weekly travel, we should:

- multiply the distance from X to A by 5;
- multiply the distance from X to B by 3;
- multiply the distance from X to C by 2.

(Actually, the total *round-trip* distances are twice these values, but the proportions would remain the same, so we can use these numbers.) Thus, the function to minimize is

$$T(x, y) = 5\sqrt{x^2 + y^2} + 3\sqrt{(x - 6)^2 + (y - 9)^2} + 2\sqrt{(x - 10)^2 + (y - 2)^2}.$$

a) Use the method of steepest descent to find the minimum of T.

b) Compare the location of the distribution center X as determined by T to its location determined by the function S of example 2 in the text. Would you *expect* the location to change? In what direction? Does the calculated change in position agree with your intuition?

18. Suppose that deliveries to A are twice as frequent as deliveries to either B or C. (For example, two trucks make the round-trip to A each day, but only one truck to B and one to C.) Where should the distribution center X be located in these circumstances? Explain how you got your answer.

[Answer: X should be at A. Does this surprise you?]

19. A company which has four offices around the country holds an annual meeting for its top executives. The location of each office, and the number of executives at that office, are given in the following table. (The coordinates x and y of the position are given in arbitrary units.)

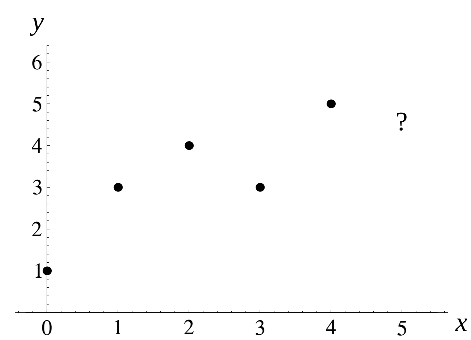
office	executives	x	y
A	32	200	300
B	17	1920	1100
C	20	2240	450
D	41	2875	1150

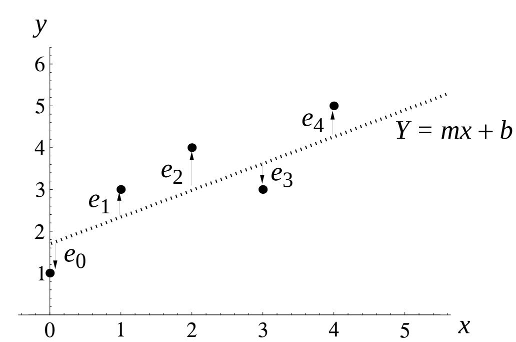
Where should the meeting be held if the location depends *solely* on the total travel cost for all the participants? Assume that the travel cost, per mile, is the same for every participant.

The best-fitting line

Suppose we've taken measurements of two quantities x and y, and obtained the results shown in the table and graph below. We assume that y depends on x according to some rule that we don't happen to know. In particular, we'd like to know what y is when $x = 5$. We have no data. Can we *predict* what y should be?

x	y
0	1
1	3
2	4
3	3
4	5
5	?





Here is a common approach to the question. We assume there is a simple underlying relation between y and x. However, the measurements that give us the data contain errors or "noise" of some sort that obscure the relationship. The simplest relation is a linear function, so we assume that there is a formula $Y = mx + b$ that describes the connection between x and y.

Which line should we choose? In other words, how should we choose m and b? Since the data points don't lie on a line, there is no perfect solution. For any choices, we must expect a difference e_j between the the

j-th data value y_j and the value $Y_j = m\,j + b$ predicted by the formula. These differences are the *errors* we assume are present.

A reasonable way to proceed is to **minimize** the total error. Even this involves choices. In the figure above, e_0 and e_3 are negative, so the ordinary total could be zero, or nearly so, even if the individual errors were large. We need a total that *ignores* the signs of the errors. Here is one:

$$\text{absolute error:} \qquad |e_0| + |e_1| + |e_2| + |e_3| + |e_4| = AE.$$

Here is another:

$$\text{squared error:} \qquad e_0^2 + e_1^2 + e_2^2 + e_3^2 + e_4^2 = SE.$$

In the following table we compare the data y_j with the calculated values Y_j and the resulting errors e_j.

x	y	Y	$e = y - Y$
0	1	b	$1 - b$
1	3	$m + b$	$3 - m - b$
2	4	$2m + b$	$4 - 2m - b$
3	3	$3m + b$	$3 - 3m - b$
4	5	$4m + b$	$5 - 4m - b$

The total errors are functions of m and b

To get the values of AE and SE, we take either the absolute values or the squares of the elements of the rightmost column, and then add. In particular, the table makes it clear that both total errors are functions of m and b. The absolute error is

$$AE(m, b) = |1 - b| + |3 - m - b| + |4 - 2m - b| + |3 - 3m - b| + |5 - 4m - b|.$$

20. Inspect a graph and a contour plot to determine the values of m and b which minimize $AE(m, b)$.

[Answer: Remarkable as is may seem, there is an entire line segment of solutions to this problem in the m, b-plane. One end of the line is near $(m, b) = (.67, 2.3)$, the other is near $(m, b) = (1, 1)$.]

21. a) Using a best-fitting line from the previous question, find the predicted value of y when $x = 5$.

b) Since there is a range of best-fitting lines, there should be a range of predicted values for y when $x = 5$. What is that range?

22. Write down the function $SE(m, b)$ that describes the squared error in the fit of a straight line to the data given above.

23. a) Use a graph and a contour plot to locate the minimum of the function $SE(m, b)$ from the previous exercise. Indicate how many digits of accuracy your answer has.

b) Use the method of steepest descent to locate the minimum. How many digits of accuracy does *this* method yield?

9.4 Chapter Summary

The Main Ideas

- The **graph** of a function of two variables is a two-dimensional surface in a three-dimensional space.
- A function of two variables can also be viewed using a **density plot**, a **terraced density plot**, or a **contour plot**. The latter is a set of **level curves** drawn on a flat plane.
- A **contour plot** of a function of three variables is a collection of **level surfaces** in three-dimensional space.
- The graph of a **linear function** is a flat plane, and its contour plot consists of straight, parallel, and equally-spaced lines.
- The **gradient** of a linear function is a vector whose components are the partial rates of change of the function.
- Under a **microscope**, the graph of a function of two variables becomes a flat plane. A contour plot turns into a set of straight, parallel, and equally-spaced lines.
- The multipliers in the **microscope equation** for a function are its **partial derivatives**:
$$\Delta z = \frac{\partial f}{\partial x_1}(a, b)\Delta x_1 + \cdots + \frac{\partial f}{\partial x_n}(a, b)\Delta x_n.$$
- The **gradient** of a function is a vector whose components are the partial derivatives of the function. Its magnitude and direction give the greatest **rate of increase** of the function at each point.
- **Optimization** is a process that involves finding the **maximum** or **minimum** value of a function. There may be **constraints** present that limit the scope of the search for an **extreme**.
- Extremes can be found at **critical points**, where all partial derivatives of a locally linear function are zero.
- The **method of steepest ascent** introduces the power of dynamical systems into the optimization process.

Expectations

- Using appropriate computer software, you should be able to make a **graph**, a **terraced density plot**, and a **contour plot** of a function of two variables.
- Using appropriate graphical representations of a function of two variables, you should be able to recognize the **maxima**, **minima**, and **saddle points**.
- You should be able to estimate the **partial rates of change** of a function of two variables at a point by zooming in on a contour plot.
- You should be able to recognize the various forms of a **linear function** of several variables and transform the representation of the function from one form to another.
- You should be able to describe the geometric meaning of the partial rates of change of a linear function of two variables.
- You should be able to find the **gradient** of a function of several variables at a point.
- You should know how the gradient of a function of two variables is related to its level curves.
- You should be able to write the **microscope equation** for a function of two variables at a point.
- You should be able to use the microscope equation for a function of two variables at a point to estimate values of the function at nearly points, to find the **trade-off** in one variable when the other changes by a fixed amount, and to estimate errors.
- You should be able to find the **linear aprroximation** to a function of two variables at a point.
- You should be able to find the equation of the **tangent plane** to the graph of a function of two variables at a point.
- You should be able to sketch the **gradient vector field** of a function of two variables in a specified domain.

- You should be able to sketch a plausible set of contour lines for a function whose gradient vecctor field is given; you should be able to sketch a plaausible gradient vector field for a ffunction whose contour plot is given.
- You should be able to find the critical points of a function of two variables, and you should be able to determine whether a critical point is an extreme by inspecting a graph or a contour plot.
- You should be able to find a local maximum of a function of two variables by the method of **steepest descent**.
- You should be able to find an extreme of a function of two variables subject to a constraint either by inspecting a graph or contour plot, or by the method of **Lagrange multipliers**.