Chapter 7

# Multivariable Systems

## 7.1 Stability in Nonlinear Differential Equations

In the previous chapter, we used our knowledge of linear algebra to give us insights into linear differential equations. The key to this approach is that the differential equation is viewed as a vector field, that is, as a function

$$V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

Therefore, since a linear differential equation is a linear vector field, which is a linear function, we used the eigenvalues and eigenvectors of the linear function to completely classify the stability of the equilibrium point of the corresponding vector field.

Now we can go on to nonlinear systems. What can we find out about them? First of all, we know that we can find the equilibrium points. As we saw in Chapter 3, we find the equilibrium points of the vector field

$$X' = f(X, Y)$$
$$Y' = g(X, Y)$$

by setting $f = g = 0$ and solving for the resulting pairs $(X^*, Y^*)$.

And in $n$ dimensions, the vector field

$$\begin{aligned} X_1' &= f_1(X_1, X_2, \ldots, X_n) \\ X_2' &= f_2(X_1, X_2, \ldots, X_n) \\ &\vdots \qquad \vdots \\ X_n' &= f_n(X_1, X_2, \ldots, X_n) \end{aligned}$$

or in vector notation

$$\mathbf{X}' = V(\mathbf{X})$$

has equilibrium points $(X_1^*, X_2^*, \ldots, X_n^*)$ whenever $X_1' = X_2' = \cdots = X_n' = 0$.

Now we want to find their stability. *The purpose of this chapter is to develop a general method for determining the stability of an equilibrium point of an n-dimensional vector field.* Previously, the only technique we had was simulation: pick a large number of initial conditions around the equilibrium point, simulate the system, and see where the points go as time evolves.

In order to grasp the general strategy, we will first revisit a section from Chapter 3 in which we introduced a technique for determining the stability of equilibrium points in one dimension.


A. Garfinkel et al., *Modeling Life*,
DOI 10.1007/978-3-319-59731-7_7

Since a vector field in one dimension is a function from $\mathbb{R}$ into $\mathbb{R}$, we could graph it in two dimensions (Figure 7.1). As can be seen, there are two equilibrium points in this system, $X = 0$ and $X = k$.

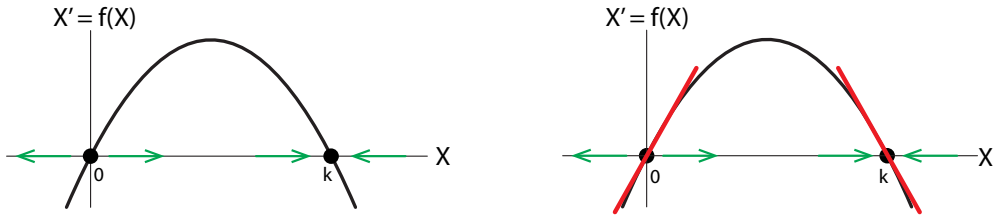



Figure 7.1: The linear approximations (red) to $X' = f(X)$ at the two equilibrium points $X = 0$ and $X = k$ give the stability of those equilibrium points.

We then argued that the stability of the equilibrium points at $X = 0$ and $X = k$ can be determined by the slope of the tangent to $f(X)$ at the two points, that is, by the derivative: if the derivative was positive, the equilibrium point was unstable, and if it was negative, the equilibrium point was stable. As can be seen, the slope of the tangent at $X = 0$ is positive, and so the equilibrium point at $X = 0$ is unstable. On the other hand, the slope of the tangent at $X = k$ is negative, and therefore, the equilibrium point at $X = k$ is stable.

**Exercise 7.1.1** What happens when the slope is zero?

**Exercise 7.1.2** Find the equilibria of $X' = X^3 - X$ and use this method to determine their stability.

As we mentioned, this was an application in one dimension of the *Hartman–Grobman theorem*: the stability of an equilibrium point of a nonlinear vector field is determined by the slope of the linear approximation to the nonlinear function at the equilibrium point.

The key to this theorem is the fact that the derivative *is* the linear approximation to a function at a point, as we saw in Chapter 2.

We will now use the same Hartman–Grobman principle in higher dimensions: the stability of an equilibrium point in a nonlinear vector field is given by the slope (except in this case, it is slopes) of its linear approximation at that point, that is, by the $n$-dimensional derivative at that point. So now we need to develop the $n$-dimensional concept of derivative.

We now need to know the following:

(1) What does a linear function look like in $n$ dimensions?
(2) How do we find the linear function that is the linear approximation to a nonlinear function in $n$ dimensions? In other words, what is the derivative in $n$ dimensions?

## 7.2 Graphing Functions of Two Variables

We will now be looking at functions of several variables, and it is important to understand what these functions look like geometrically. As usual, we will consider the case of two variables as our example.

First, let's take a linear case. Let's begin by considering the linear function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

given by

$$Z = f(X, Y) = -0.5X + Y$$

If we choose a pair $(X, Y)$ at random (Figure 7.2, gray dot), we can plot its corresponding $Z$ value calculated by $Z = f(X, Y)$ (black point). If we plot many points in this way, we get a point cloud of $Z$-values (black dots) corresponding to the $(X, Y)$ points (gray). The black dots are the thousand $Z$ values, and they all lie exactly on the green plane, which is the set of *all* $Z$ values for *all* $(X, Y)$ pairs in the $XY$ plane.

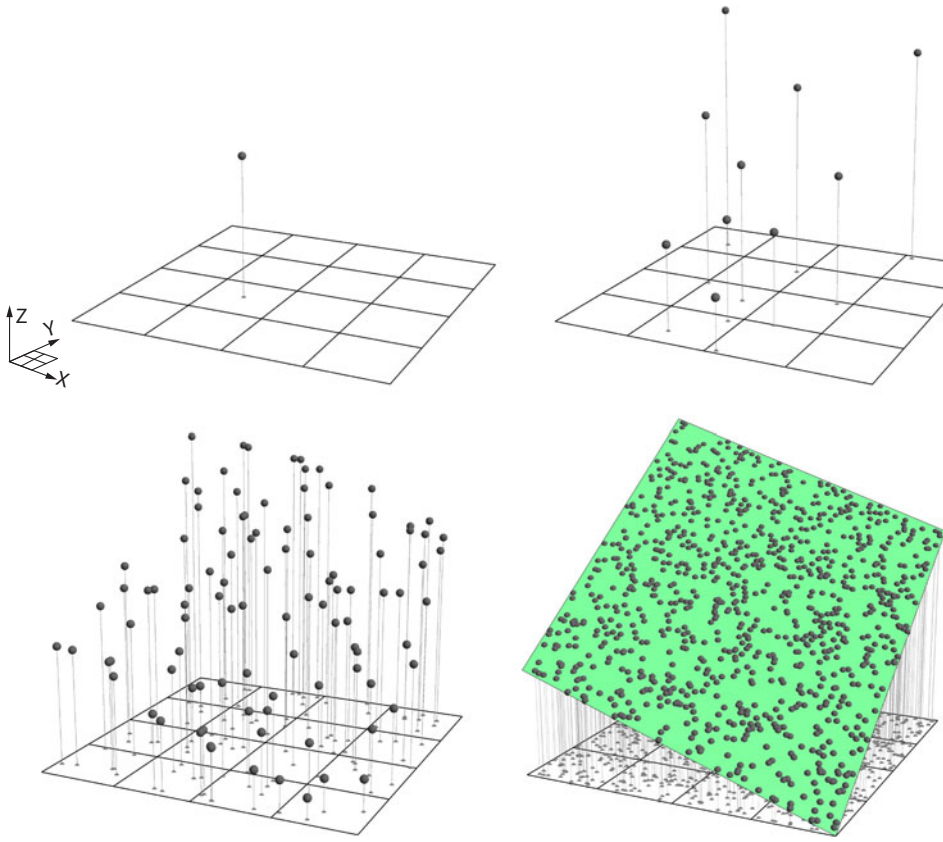



Figure 7.2: Points satisfying $Z = -0.5X + Y$. Shown are 1, 10, 100, and finally, 1000 points superimposed on the plane $Z = -0.5X + Y$.

A linear function $f : \mathbb{R}^2 \to \mathbb{R}$ is represented by a **plane** over the $(X, Y)$ plane.

For a nonlinear example

$$Z = f(X, Y)$$

let's use

$$Z = f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

If we choose a random pair $(X,Y)$ (Figure 7.3, gray dot) and plot the respective $Z$ value (black dot), we get a point in 3D space. If we plot many such points, the resulting point cloud begins to suggest a surface. Indeed, the points lie exactly on the curved surface, which is the graph of *all* $Z$ values corresponding to *all* $(X,Y)$ points in the square.

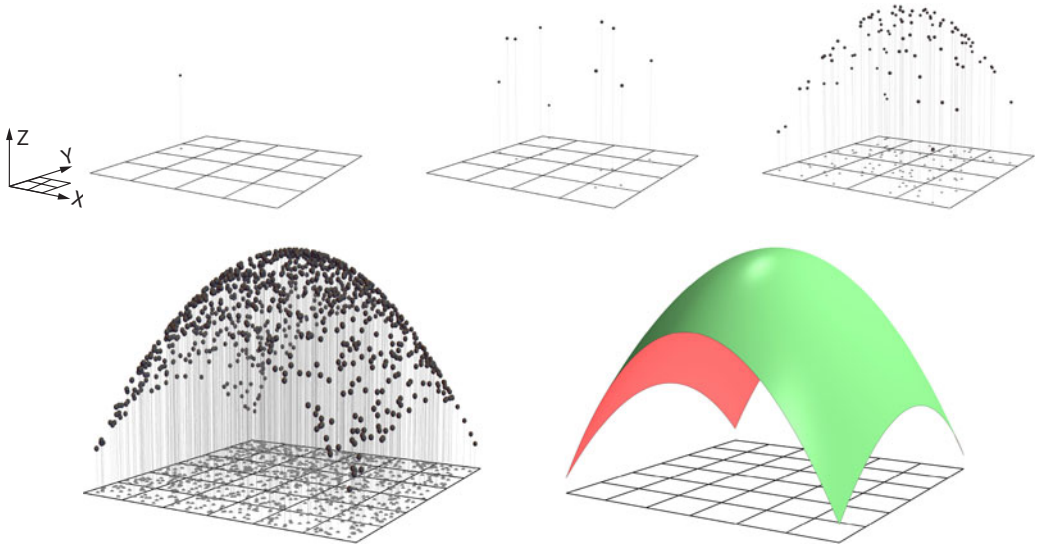



Figure 7.3: Top row: 1, 10, and 100 random $(X,Y)$ pairs (gray dots) give rise to corresponding $Z$ values (black dots) according to the equation $Z = 5 - \frac{X^2}{2} - \frac{Y_2}{4}$. Bottom left: a thousand random $(X,Y)$ pairs (gray dots) with their corresponding $Z$ values (black dots). Bottom right: the corresponding surface is the set of all $Z$ values for every $(X,Y)$ in the square.

This is true in general: the graph of a function $\mathbb{R}^2 \to \mathbb{R}$ is a surface over the $\mathbb{R}^2$ plane. These functions are sometimes called *height functions*, because you can look at them as a terrain map, with $Z$ representing the height of the terrain at the point $(X,Y)$.

A nonlinear function $f : \mathbb{R}^2 \to \mathbb{R}$ is represented by a **surface** over the $(X,Y)$ plane.

**Exercise 7.2.1** Why is the graph of a function $f : \mathbb{R}^2 \to \mathbb{R}$ a surface rather than, say, two surfaces? In other words, why can't we have a point that lies directly above another point?

**Exercise 7.2.2** Compute $f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$ for four points in the $(X,Y)$ plane. Then, use the `list_plot3d` command in SageMath to plot these points.

**Exercise 7.2.3** Do the same thing for another function of your choice. Then, use the `plot3d` command to plot the function on the same graph as the points. (The command `plot3d` works just like plot, except that you have to specify plotting ranges for two variables, not just one.)

## 7.3 Linear Functions in Higher Dimensions

We know from Chapter 6 what linear functions in $n$ dimensions look like algebraically. Now we want to look at them geometrically.

Let's start with an example in two dimensions.

A linear function $V : \mathbb{R}^2 \to \mathbb{R}^2$,

$$V : (X, Y) \longrightarrow (Z, W)$$

can be represented as

$$\begin{aligned} Z &= f(X, Y) = aX + bY \\ W &= g(X, Y) = cX + dY \end{aligned} \tag{7.1}$$

The first problem we face is visualization: the graph of a function $\mathbb{R}^2 \to \mathbb{R}^2$ that takes $(X, Y)$ to $(Z, W)$ would have to have four dimensions. So we use the technique of looking at the two $\mathbb{R}^2 \to \mathbb{R}$ component functions one by one, decomposing $V$ into the component functions $f$ and $g$. Recalling that $(X, Y)$ is the vector $\begin{pmatrix} X \\ Y \end{pmatrix}$, we can write

$$\begin{pmatrix} Z \\ W \end{pmatrix} = V\left(\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \begin{pmatrix} f(X, Y) \\ g(X, Y) \end{pmatrix}$$

For simplicity, in the rest of the chapter we will drop the vector notation and write

$$\begin{aligned} f &: (X, Y) \longrightarrow (Z) \qquad \text{and} \\ g &: (X, Y) \longrightarrow (W) \end{aligned}$$

both of which are $\mathbb{R}^2 \to \mathbb{R}$. So $f$ gives us the first coordinate, $Z$, and $g$ gives us the second coordinate, $W$.

These component functions *are* graphable (Figure 7.4).

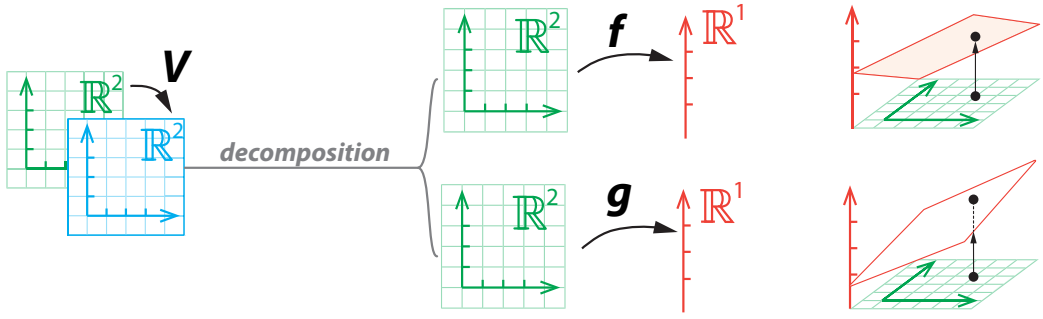



Figure 7.4: Decomposition of a 2D linear function $\mathbb{R}^2 \to \mathbb{R}^2$ into two linear functions $\mathbb{R}^2 \to \mathbb{R}$.

Let's begin by considering the first linear function

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}$$

given by

$$Z = f(X, Y) = -0.5X + Y$$

which, as we just saw, is a plane over $X$-$Y$ space (Figure 7.5).

In general, a plane is tilted with respect to both the $XZ$ and $YZ$ axes. If the plane passes through the origin, as the graph of a linear function must, knowing what the slopes are tells us exactly what the plane is.

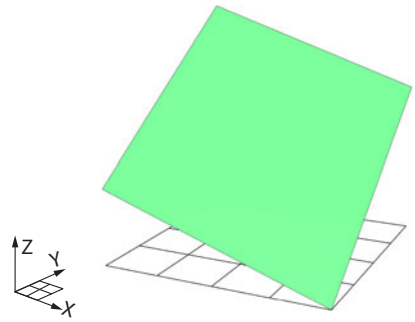



Figure 7.5: $Z = -0.5X + Y$. The green plane is the set of all such $Z$ values for $(X, Y)$ lying in the square.

**Exercise 7.3.1** Why does the graph of a linear function have to pass through the origin?

In order to calculate the tilt, we will visualize it using the cutting planes $X = 0$, which is the $YZ$ plane, and $Y = 0$, which is the $XZ$ plane (Figure 7.6).

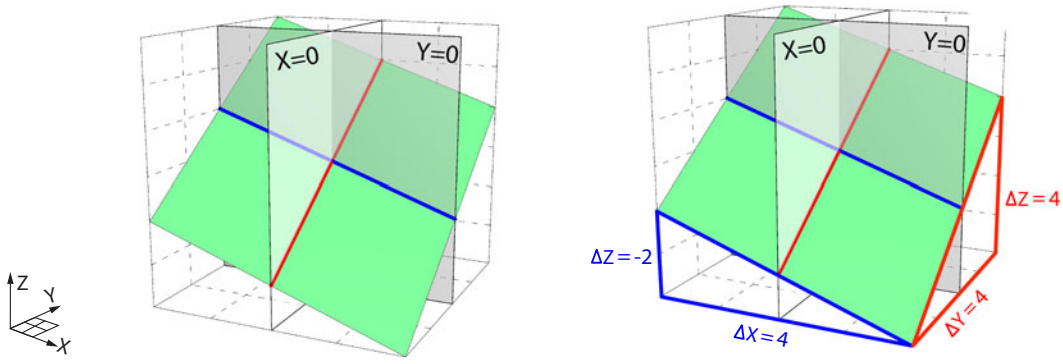



Figure 7.6: The plane $Z = -0.5X + Y$ has two slopes, revealed by the two gray cutting planes.

First, let's look at the gray cutting plane $X = 0$. The intersection of the green plane with the $X = 0$ cutting plane is the red line. The slope of the red line is

$$\frac{\Delta Z}{\Delta Y} = \frac{4}{4} = 1$$

**Exercise 7.3.2** In the right panel of Figure 7.6:

a) Where are the cutting planes?

b) What is the significance of the red triangle?

c) How do we know that the two red lines have the same slope?

d) How do we know the values of $\Delta Y$ and $\Delta Z$ (other than reading the labels)?

Now let's look at the other gray cutting plane, $Y = 0$. The intersection of the green plane with the $Y = 0$ cutting plane is the blue line.

**Exercise 7.3.3** Compute the slope of the blue line.

These two slopes determine the plane. Notice that the original plane was

$$Z = -0.5X + Y$$

What we have just seen is that the two slopes are $\frac{\Delta Z}{\Delta X} = -0.5$ and $\frac{\Delta Z}{\Delta Y} = 1$. In other words, the slope of the green plane along the $YZ$ axis is $\frac{\Delta Z}{\Delta X}$, which is the coefficient of the $X$ term. Similarly, the slope of the green plane along the $XZ$ axis is $\frac{\Delta Z}{\Delta Y}$, which is the coefficient of the $Y$ term.

In general, if $Z = aX + bY$ is a plane, then its slopes are given by

$$\frac{\Delta Z}{\Delta X} = a \qquad \text{and} \qquad \frac{\Delta Z}{\Delta Y} = b$$

This completes our analysis of the first component function $f$. By exactly similar reasoning, we can consider the second component function

$$W = g(X, Y) = cX + dY$$

whose slopes are

$$\frac{\Delta W}{\Delta X} = c \qquad \text{and} \qquad \frac{\Delta W}{\Delta Y} = d$$

When we put the two component functions $f$ and $g$ back together, we get the linear function

$$\begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ (X, Y) & \longrightarrow & (Z, W) \end{array}$$

which is given by the matrix

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

And the original linear equation (7.1) is represented by

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Z \\ W \end{pmatrix}$$

### *n* Dimensions

In $n$ dimensions, a linear function

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is decomposable into $n$ component functions, $f_1, f_2, \ldots, f_n$, where each component function

$$f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$$

has the form

$$f_i(X_1, X_2, \ldots, X_n) = a_{1i}X_1 + a_{2i}X_2 + \ldots a_{ni}X_n$$

so that the overall function is represented by the matrix

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

By analogy with the plane defined by a linear function $\mathbb{R}^2 \longrightarrow \mathbb{R}$, we say that each function $f_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ defines a *hyperplane*

$$Z = f_i(X_1, X_2, \ldots, X_n) = a_{1i}X_1 + a_{2i}X_2 + \ldots + a_{ni}X_n$$

The hyperplane has $n$ slopes given by $a_{1i}, a_{2i}, \cdots, a_{ni}$, so that the plane can also be written

$$Z = \frac{\Delta Z}{\Delta X_1}X_1 + \frac{\Delta Z}{\Delta X_2}X_2 + \ldots + \frac{\Delta Z}{\Delta X_n}X_n$$

**Further Exercises 7.3**

1. Write the equation for the plane passing through the origin that has the slopes below:
   a) $\frac{\Delta Z}{\Delta Y} = 3$ and $\frac{\Delta Z}{\Delta X} = 5$
   b) $\frac{\Delta Z}{\Delta X} = 4$ and $\frac{\Delta Z}{\Delta Y} = 1.5$
   c) $\frac{\Delta Z}{\Delta X} = -3$ and $\frac{\Delta Z}{\Delta Y} = -1$
2. Find $\frac{\Delta Z}{\Delta X}$ and $\frac{\Delta Z}{\Delta Y}$ for the planes specified by the equations below:
   a) $Z = 7X + 25Y$
   b) $Z = 3Y - 2X$
   c) $Z = \pi Y + 16X$

## 7.4 Nonlinear Functions in Two Dimensions

Recall that our goal is to find the linear vector field that is an approximation to a nonlinear one at an equilibrium point.

As usual, we will look at the vector field as a function

$$\begin{array}{rccc} V & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & & (X,Y) & \longrightarrow & (Z,W) \end{array}$$

We'll use the same technique as above and split $V$ into the two component functions (Figure 7.7)

$$\begin{array}{rccc} f & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & & (X,Y) & \longrightarrow & (Z) \end{array} \qquad \text{and } g : \begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ (X,Y) & \longrightarrow & (W) \end{array}$$

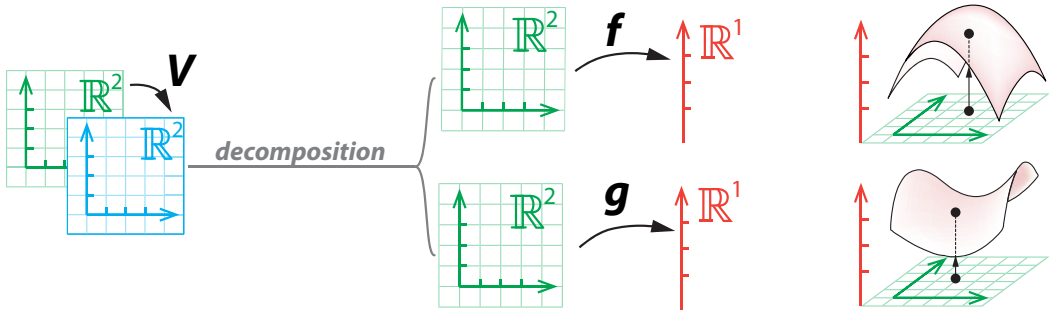



Figure 7.7: Decomposition of a nonlinear function $V : \mathbb{R}^2 \to \mathbb{R}^2$ into two component functions $f$ and $g$, $\mathbb{R}^2 \to \mathbb{R}$.

### First Component Function $f$

Let's consider the first component function: $Z = f(X,Y)$. We will start with the example

$$Z = f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

(Figure 7.8).

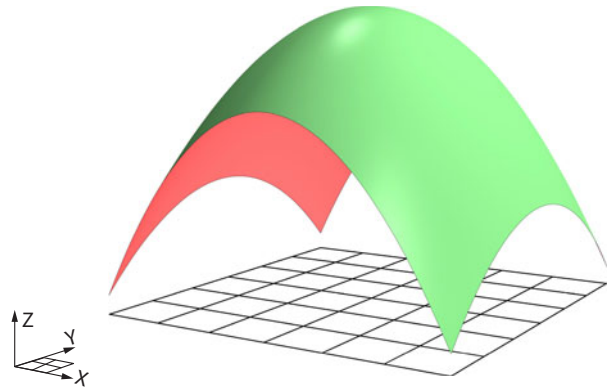



Figure 7.8: $Z = f(X,Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$. The corresponding surface is the set of all $Z$ values for every $(X,Y)$ in the square.

### The Tangent Plane

Our next task is to find the linear function $\mathbb{R}^2 \to \mathbb{R}$ that approximates the surface $f$ at the point $(X_0, Y_0)$. What is this linear function? As we saw above, a linear function $\mathbb{R}^2 \to \mathbb{R}$ defines a *plane*.

To visualize this plane, remember that in one dimension, we zoomed in on a 1D curve to visualize the 1D tangent line. Here we are going to zoom in on a 2D surface (Figure 7.9). We see that *as we zoom in on the 2D surface, it begins to resemble a 2D plane*. This plane is called the *tangent plane to* $\boldsymbol{f}$ *at the point* $(\boldsymbol{X_0, Y_0})$.

The linear approximation to the 2D surface $Z = f(X, Y)$ at the point $(X_0, Y_0)$ is called the **tangent plane** to $f$ at the point $(X_0, Y_0)$.

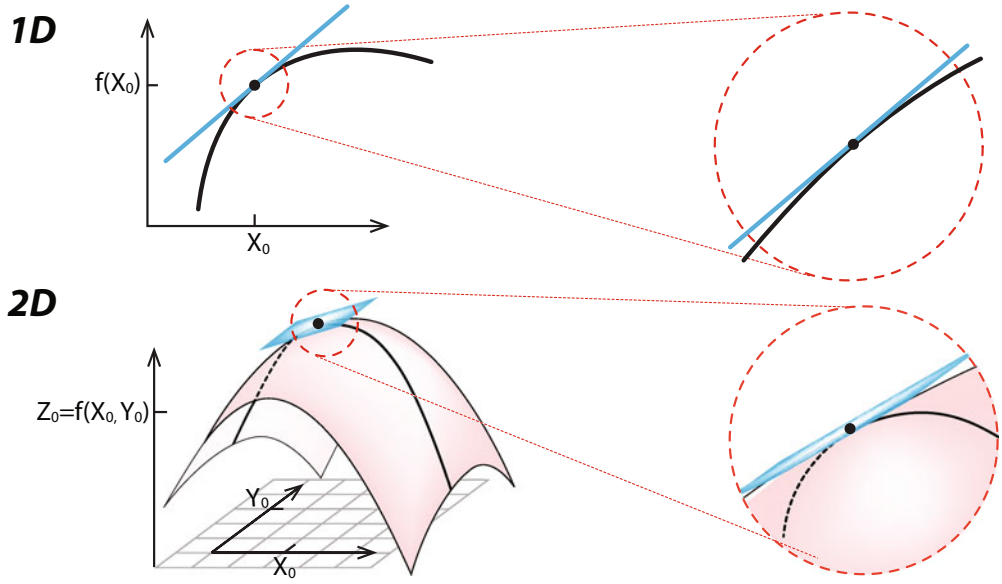



Figure 7.9: Just as zooming in on a 1D curve gives a 1D straight line, zooming in on a 2D surface gives a plane.

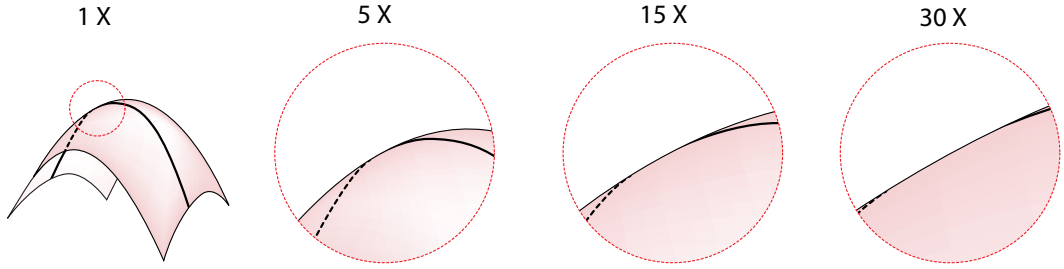



Figure 7.10: Zooming in on a smooth 2D surface makes the surface look flatter and flatter.

It makes sense that the linear approximation in 2D should be a plane, because the linear approximation must be a linear function, and we just saw that the linear functions $\mathbb{R}^2 \to \mathbb{R}$ are defined by planes.

**Exercise 7.4.1** In SageMath, plot a function of two variables. Pick a point on the function and zoom in on it. What do you observe?

### Calculating the Tangent Plane

The tangent plane is a plane, and we saw earlier that a plane is defined by two slopes. We now need to calculate the two slopes that determine the tangent plane.

To do this, we will make another critical decomposition: at each point on the 2D surface, we will split a small patch of surface around that point into two 1D functions using a new method: we will use 2D *cutting planes*.

The cutting plane construction allows us, in any given patch of surface, to turn the $\mathbb{R}^2 \to \mathbb{R}$ function into two $\mathbb{R} \to \mathbb{R}$ functions.

The $XZ$ cutting planes are exactly the planes $Y = constant$. And $YZ$ cutting planes are exactly the planes $X = constant$. If we look at the $XY$ and $XZ$ cutting planes, we see that *the 2-dimensional surface $f$ always intersects the cutting plane in a 1-dimensional curve.*

For example, the $YZ$ cutting plane at $X = 1$ intersects the green surface $Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$ in the black curve, shown in Figure 7.11. The equation for this black curve can be found easily by plugging $X = 1$ into the $Z$ equation

$$Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

which gives us

$$\begin{aligned} & Z = 5 - \frac{1^2}{2} - \frac{Y^2}{4} \\ \implies \quad & Z = 4.5 - \frac{Y^2}{4} \end{aligned}$$

which is a curve in the $YZ$ plane (Figure 7.11, right).

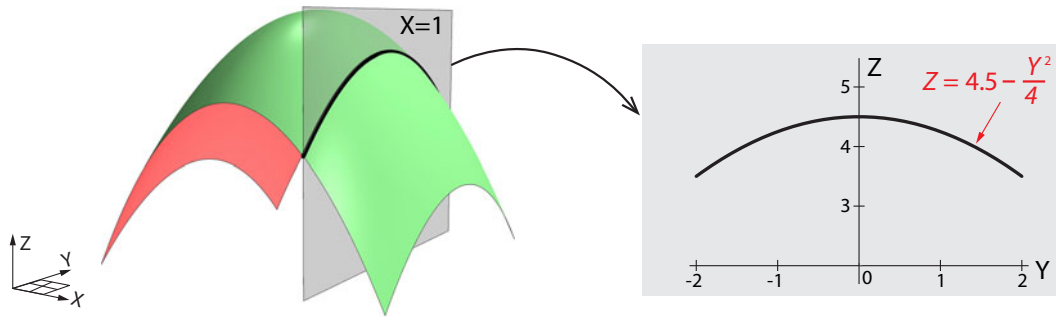



Figure 7.11: The $YZ$ cutting plane at $X = 1$ intersects the surface in the black curve.

**Exercise 7.4.2** Give an example of an $XZ$ cutting plane.

**Exercise 7.4.3** Find the equation of the curve that results from intersecting the surface $Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$ with the cutting plane $Y = 2$. Plot this curve in SageMath.

**Method of Cutting Planes**

To calculate the intersection of a 2D surface with a cutting plane $X = constant$, just plug the value of the cutting plane into the equation for the 2D surface. This gives a 1D function giving $Z$ as a function of $Y$, obtained by "holding $X$ constant."

Similarly, to calculate the intersection of a 2D surface with a cutting plane $Y = constant$, just plug the value of the cutting plane into the equation for the 2D surface. This gives a 1D function giving $Z$ as a function of $X$, obtained by "holding $Y$ constant."

Since the function $Z = f(X,Y)|_{X=1} = 4.5 - \frac{Y^2}{4}$, which gives $Z$ as a function of $Y$, is just a function of one variable, it has a derivative

$$\frac{dZ}{dY}\Big|_{Y=Y_0} \text{ at any point } Y_0$$

This derivative $\frac{dZ}{dY}$ can be thought of and calculated as the derivative of a 1-dimensional function $\mathbb{R} \to \mathbb{R}$, which is of course the subject of classical calculus as developed in Chapter 2. In this case, using classical calculus techniques, the curve

$$Z = 4.5 - \frac{Y^2}{4}$$

is seen to have as its derivative function

$$\frac{dZ}{dY} = -\frac{2}{4}Y$$

So for example,

$$\frac{dZ}{dY}\Big|_{Y=-1} = -\frac{2}{4} \times (-1) = 0.5$$

which means that the linear approximation to the curve is the function (Figure 7.12)

$$\Delta Z = 0.5\ \Delta Y$$

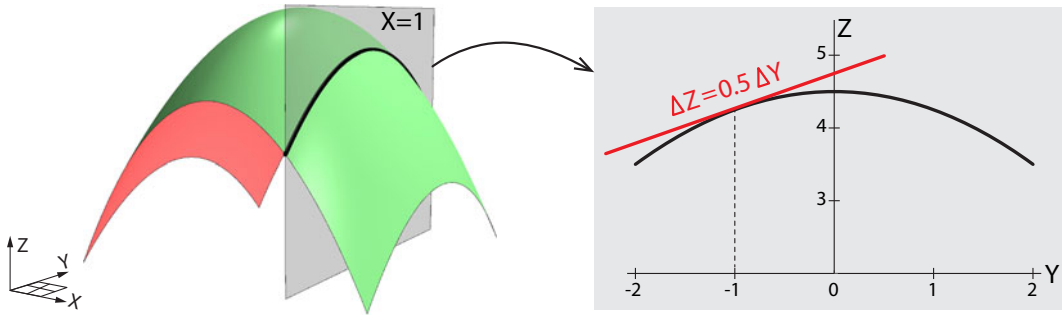



Figure 7.12: The tangent to the 1D curve produced by the intersection of the cutting plane and the original surface is shown at the point $Y = -1$.

**Notation**

We just defined "$\frac{dZ}{dY}$." But when we are dealing with functions of several variables, like $f(X,Y) = Z$, the derivative of $f$ with respect to one of the variables is written using a new symbol. Instead of writing $\frac{dZ}{dY}$ or $\frac{df}{dY}$, we use the symbol $\partial$ and write

$$\frac{\partial Z}{\partial Y} \text{ or } \frac{\partial f}{\partial Y}$$

to indicate that $Y$ is one of several variables that determine $Z$. This is called the **partial derivative of Z with respect to Y**.

**Exercise 7.4.4** Find the linear approximation to $Z = 4.5 - \dfrac{Y^2}{4}$ at $Y = 3$.

Note that we calculated the partial derivative $\frac{dZ}{dY}$ by looking at the function $Z = f(X,Y)$ and taking the derivative of this function while holding everything other than $Y$ constant. This is the algebraic equivalent of the method of cutting planes: the cutting plane is the geometric picture of holding the other variable constant. For example, using the $YZ$ cutting plane amounts to taking $X = \textit{constant}$. Similarly, using the $XZ$ cutting plane amounts to taking $Y = \textit{constant}$.

If $Z = f(X,Y)$, then the partial derivative of $Z$ with respect to $Y$ is calculated by holding all variables other than $Y$ constant and then calculating the 1-dimensional derivative of the resulting function.

So the linear approximation to $Z = f(X,Y)\Big|_{X=constant}$ is

$$\Delta Z = \frac{\partial f}{\partial Y} \cdot \Delta Y \quad \text{or} \quad \Delta Z = \frac{\partial Z}{\partial Y} \cdot \Delta Y$$

We have now answered half of our original question: what is the linear approximation to the 2-dimensional surface $Z = f(X,Y)$ at the point $(X_0, Y_0)$? We have found that one of the two slopes is

$$\frac{\partial f}{\partial Y}\Big|_{Y=Y_0}$$

What about the other slope?

By similar reasoning, we use a $Y = \textit{constant}$ cutting plane to find $Z$ as a function of $X$ (Figure 7.13). Here we use $Y = -1$.

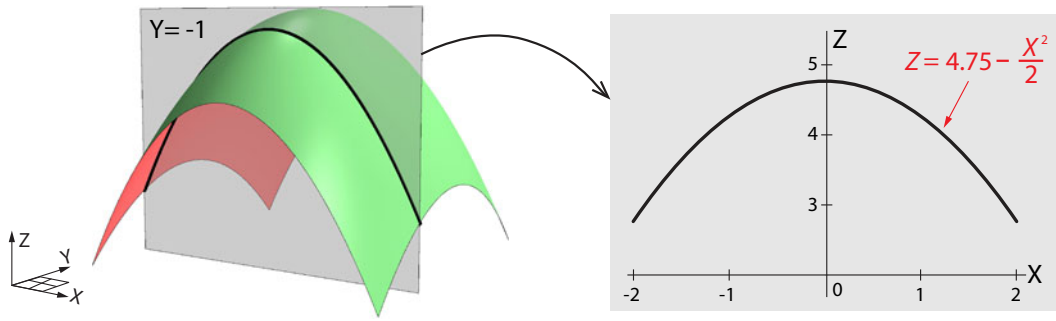



Figure 7.13: The $XZ$ cutting plane at $Y = -1$ intersects the surface in the black curve.

To find the equation for the black curve, we plug $Y = -1$ into the $Z$ equation

$$Z = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

to get

$$Z = 5 - \frac{X^2}{2} - \frac{(-1)^2}{4}$$

$$\implies \quad Z = 4.75 - \frac{X^2}{2}$$

and as with $Y$, the linear approximation to $Z$ as a function of $X$ is

$$\frac{\partial Z}{\partial X} = -X$$

At the point $X = 1$,

$$\frac{\partial Z}{\partial X}\Big|_{X=1} = -1$$

which means that the linear approximation to the curve at the point $X = 1$ is the linear function (Figure 7.14)

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{X=1} \cdot \Delta X = -1 \cdot \Delta X$$

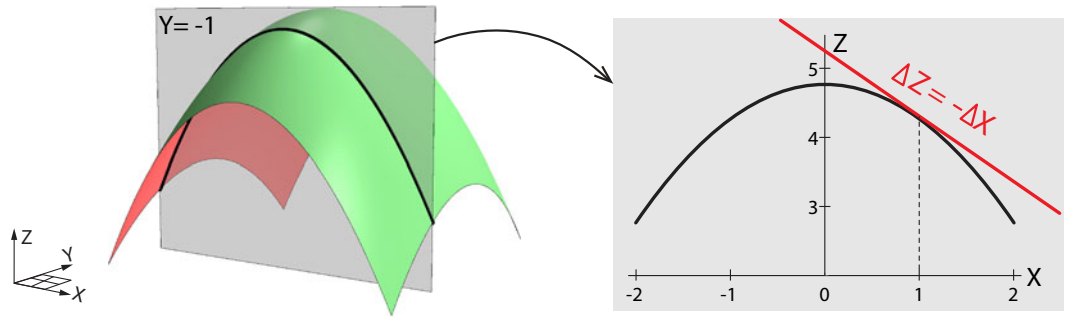



Figure 7.14: The tangent to the 1D curve produced by the intersection of the cutting plane and the original surface is shown at the point $X = 1$.

**Exercise 7.4.5** Find the linear approximation to the curve you computed in Exercise 7.4.3 on page 376 at $X = 1$.

We have now found the second slope, and we can now write the equation for the tangent plane. Since the tangent plane to $Z = f(X, Y)$ at the point $(X_0, Y_0, f(X_0, Y_0))$ has two slopes, $\frac{\partial Z}{\partial X}\big|_{(X_0,Y_0)}$ and $\frac{\partial Z}{\partial Y}\big|_{(X_0,Y_0)}$. It follows that the equation for the tangent plane is

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial Z}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

This is also the linear approximation to the curve $f$ at the point $(X_0, Y_0)$ (Figure 7.15).

If $Z = f(X, Y)$ is a surface over the 2D plane $XY$, then the linear approximation to $f$ at the point $(X_0, Y_0)$ is the linear function

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial Z}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

This function defines the **tangent plane** to $f$ at the point $(X_0, Y_0, f(X_0, Y_0))$.

Note that the tangent plane is a plane and is therefore not part of the curved surface. The plane and the surface have only one point in common.

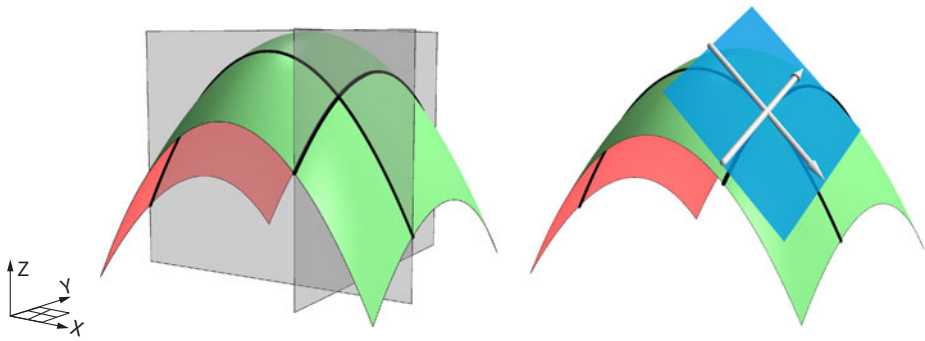



Figure 7.15: Tangent plane to the surface $Z = f(X, Y)$ at the point $(X_0, Y_0, f(X_0, Y_0))$, when $(X_0, Y_0) = (1, -1)$.

Every point on the surface has its own tangent plane (Figure 7.16). At this degree of magnification, it may look as though the blue tangent planes are lying in the green surface, but they aren't.

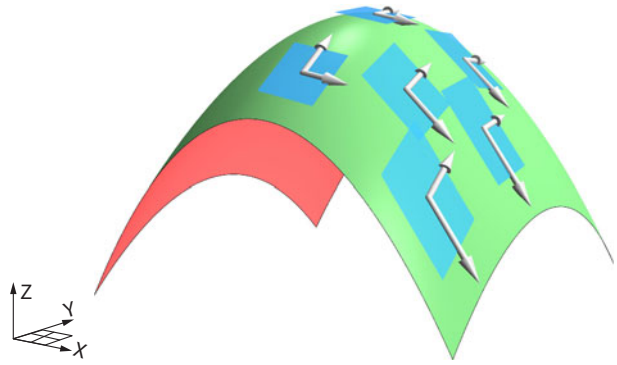



Figure 7.16: Representative tangent planes to the surface $Z = f(X, Y)$.

### Second Component Function $g$

Recall that we considered the function

$$\begin{array}{rccc} V & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ & & (X,Y) & \longrightarrow & (Z,W) \end{array}$$

and split $V$ into the two component functions $f$ and $g$:

$$\begin{array}{rccc} f: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (X,Y) & \longrightarrow & (Z) \end{array} \qquad \text{and} \qquad \begin{array}{rccc} g: & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & (X,Y) & \longrightarrow & (W) \end{array}$$

We have completed the analysis of the first component function $f$. We now need to consider the second $\mathbb{R}^2 \to \mathbb{R}$ component function

$$W = g(X, Y)$$

By methods exactly similar to those of the previous section, we use the method of cutting planes to extract the partial derivatives $\frac{\partial g}{\partial X}$ and $\frac{\partial g}{\partial Y}$, or in other words, $\frac{\partial W}{\partial X}$ and $\frac{\partial W}{\partial Y}$. We can then say that the linear approximation to $W = g(X, Y)$ at the point $(X_0, Y_0, g(X_0, Y_0))$ is

$$\Delta W = \frac{\partial W}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial W}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

or

$$\Delta g = \frac{\partial g}{\partial X}\Big|_{(X_0,Y_0)} \cdot \Delta X + \frac{\partial g}{\partial Y}\Big|_{(X_0,Y_0)} \cdot \Delta Y$$

Here we will use the example (Figure 7.17)

$$W = g(X, Y) = 0.5(X^2 - Y^2)$$

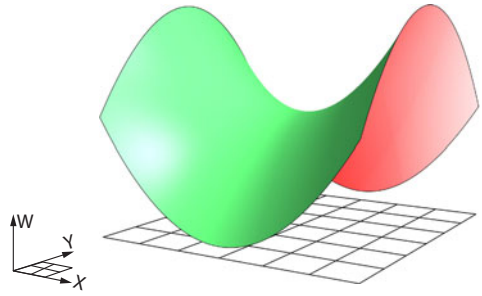



Figure 7.17: The surface $W = g(X, Y) = 0.5(X^2 - Y^2)$.

Since we have already found the approximation to the first component function $f$ at the point $(X_0, Y_0) = (1, -1)$, we will now study the second component function $g$ at the same point.

If we first consider the $YW$ cutting plane at $X = 1$, we get the black curve shown in Figure 7.18. We can easily calculate the equation for the black curve by plugging $X = 1$ into the equation for the surface:

$$\begin{aligned} & W = 0.5(1^2 - Y^2) \\ \Longrightarrow \quad & W = 0.5 - 0.5Y^2 \end{aligned}$$

At any point $Y_0$, this black curve has a 1-dimensional linear approximation. This is, of course, the derivative. The function $w = g(X, Y)\big|_{X=1}$ giving $W$ as a function of $Y$ "holding $X$ constant" has a derivative

$$\frac{dW}{dY}\Big|_{X=X_0}$$

at every point $X_0$.

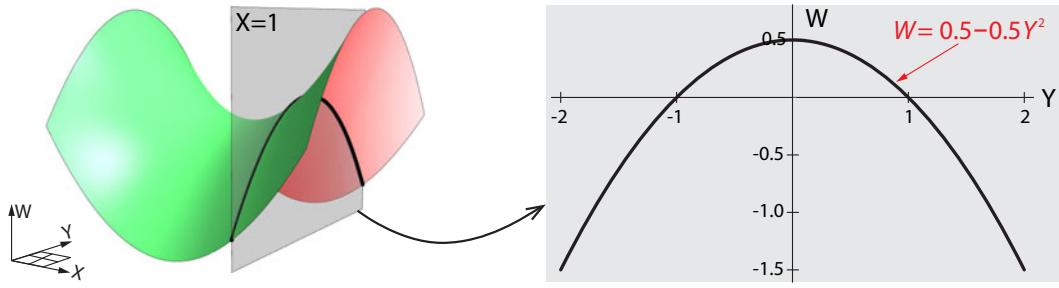



Figure 7.18: The $YW$ cutting plane at the point $X = 1$ intersects the original surface in the black curve.

This derivative $\frac{dW}{dY}$ can be calculated as before using classical calculus techniques.

The function

$$W = 0.5 - 0.5Y^2$$

has as its derivative function (Figure 7.19)

$$\frac{dW}{dY} = -0.5 \times 2Y = -Y$$

which at the point $Y = -1$ is given by

$$\left.\frac{dW}{dY}\right|_{Y=-1} = -0.5 \times 2(-1) = 1$$

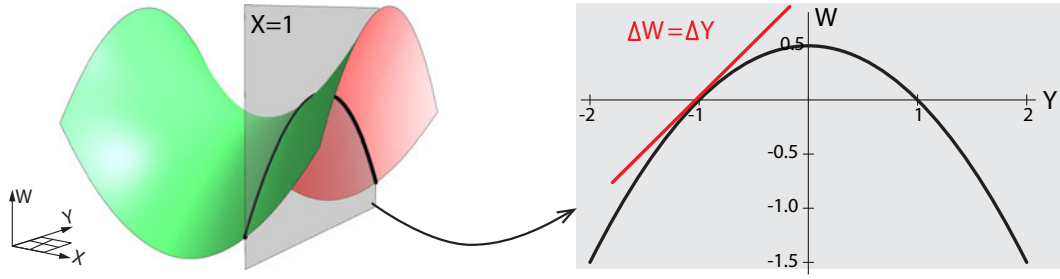



Figure 7.19: The linear approximation to the black curve is shown at the point $Y = -1$.

So the linear approximation to $W = g(X, Y)\big|_{X=constant}$ is

$$\Delta W = \frac{\partial g}{\partial Y} \cdot \Delta Y \quad \text{or} \quad \Delta W = \frac{\partial W}{\partial Y} \cdot \Delta Y$$

At the point $(X_0, Y_0) = (1, -1)$, the linear approximation is

$$\Delta W = 1 \times \Delta Y$$

We have now answered half of our original question: what is the linear approximation to the 2-dimensional surface $W = g(X, Y)$ at the point $(X_0, Y_0) = (1, -1)$? We have found that one of the two slopes is

$$\left.\frac{\partial g}{\partial Y}\right|_{X=X_0} = 1$$

What about the other slope?

By similar reasoning, we use a $Y = constant$ cutting plane to find $W$ as a function of $X$ (Figure 7.20). Again we use $Y = -1$.

Plugging $Y = -1$ into the $W$ surface equation

$$W = 0.5(X^2 - Y^2)$$

we get the equation for the black curve (Figure 7.20),

$$\begin{aligned} & W = 0.5(X^2 - (-1)^2) \\ \Longrightarrow \quad & W = 0.5X^2 - 0.5 \end{aligned}$$

and as before, the function giving $W$ as a function of $X$ has a linear approximation at every point $X_0$. This linear approximation is given by

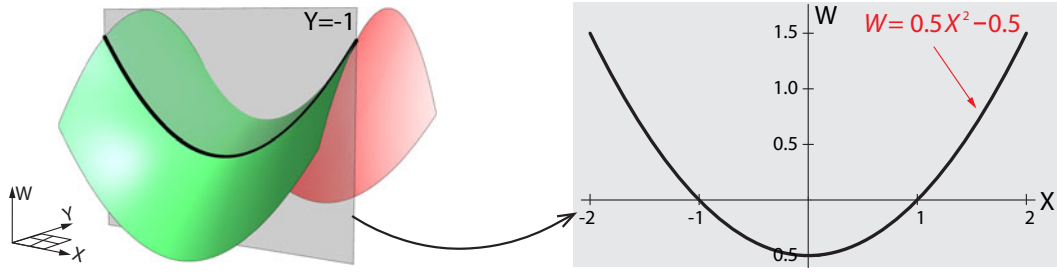



Figure 7.20: The $XW$ cutting plane $Y = -1$ intersects the original surface, yielding the black curve.

$$\Delta W = \frac{\partial g}{\partial X} \cdot \Delta X \quad \text{or} \quad \Delta W = \frac{\partial W}{\partial X} \cdot \Delta X$$

Using classical calculus techniques, we obtain

$$\frac{\partial W}{\partial X} = 0.5 \times 2X = X$$

At the point $(X_0, Y_0) = (1, -1)$, this gives us the approximation (Figure 7.21)

$$\Delta W = 1 \times \Delta X$$

We have now found the second slope; it is

$$\frac{\partial g}{\partial X}\Big|_{Y=Y_0} = 1$$

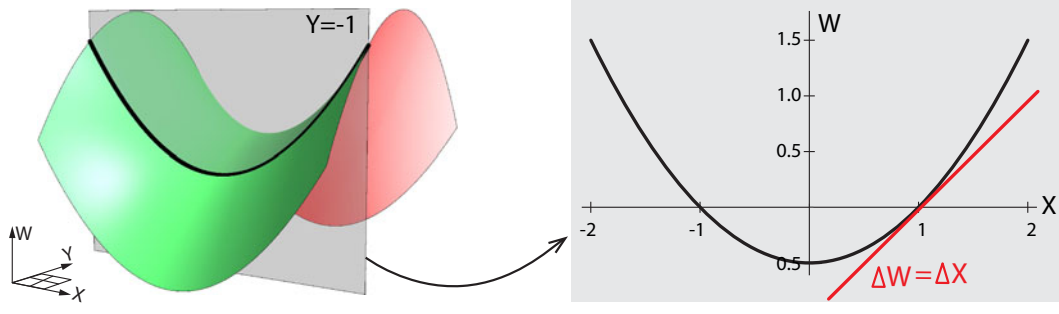



Figure 7.21: The linear approximation to the black curve at the point $X = 1$.

We now have found the equation for the tangent plane (Figure 7.22 left) to

$$W = g(X, Y)$$

at a point $(X_0, Y_0)$. It is

$$\Delta W = \frac{\partial g}{\partial X}\Big|_{(X_0,Y_0)} \Delta X + \frac{\partial g}{\partial Y}\Big|_{(X_0,Y_0)} \Delta Y$$

This is the linear approximation to $g$ at the point $(X_0, Y_0)$. In the example of $W = 0.5(X^2 - Y^2)$ at (1,-1), it is

$$\Delta W = \Delta X + \Delta Y$$

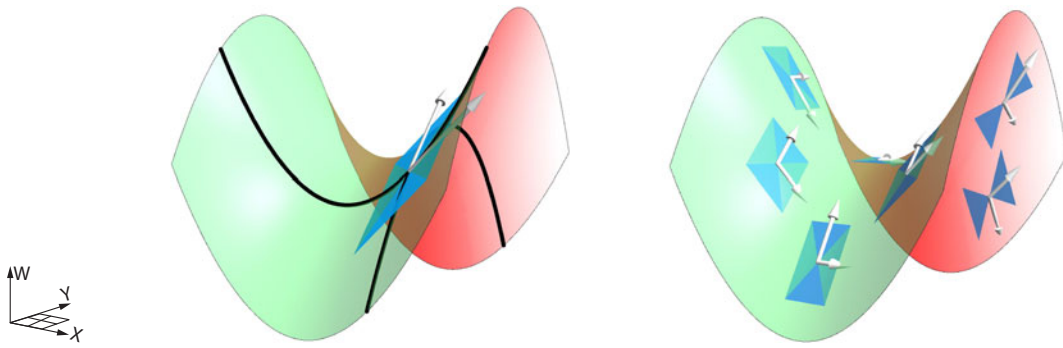



Figure 7.22: Left: Tangent plane to the original $W = g(X, Y)$ surface at the point $(X_0, Y_0, g(X_0, Y_0))$ where $(X_0, Y_0) = (1, -1)$. Right: Each point on the surface has its own tangent plane.

As we saw previously with the first component function $f$, there is a unique tangent plane to $g$ at every point $(X_0, Y_0)$ (Figure 7.22, right).

**Exercise 7.4.6** Find the tangent plane to $W = 0.5(X^2 - Y^2)$ at the point $\Big(1, 3, g(1, 3)\Big)$.

### Putting the Two Component Functions $f$ and $g$ Together

We can now put the linear approximation to $f$ and the linear approximation to $g$ back together again to produce a linear approximation to the original function $V : \mathbb{R}^2 \to \mathbb{R}^2$.

Since

$$\begin{aligned} V(X, Y) &= \big( \; f(X, Y) \; , \; g(X, Y) \; \big) \\ &= \big( \quad Z \quad , \quad W \quad \big) \end{aligned}$$

is a function $\mathbb{R}^2 \to \mathbb{R}^2$, the linear approximation to $V$ at the point $(X_0, Y_0)$ must be a *linear* function $\mathbb{R}^2 \to \mathbb{R}^2$. This is the function

$$(\Delta X, \Delta Y) \longrightarrow (\Delta Z, \Delta W)$$

whose first component is

$$\Delta Z = \frac{\partial Z}{\partial X}\Big|_{(X_0, Y_0)} \cdot \Delta X + \frac{\partial Z}{\partial Y}\Big|_{(X_0, Y_0)} \cdot \Delta Y$$

and whose second component is

$$\Delta W = \frac{\partial W}{\partial X}\Big|_{(X_0, Y_0)} \cdot \Delta X + \frac{\partial W}{\partial Y}\Big|_{(X_0, Y_0)} \cdot \Delta Y$$

Therefore, the composite linear function

$$(\Delta X, \Delta Y) \longrightarrow (\Delta Z, \Delta W)$$

is

$$\Big(\Delta X, \Delta Y\Big) \longrightarrow \left(\frac{\partial Z}{\partial X}\Delta X + \frac{\partial Z}{\partial Y}\Delta Y, \quad \frac{\partial W}{\partial X}\Delta X + \frac{\partial W}{\partial Y}\Delta Y\right)_{(X_0,Y_0)}$$

Notice that we have stopped writing $|_{(X_0,Y_0)}$ next to each of the partial derivatives; instead, we write it just once to indicate that it applies to the whole expression.

This 2D linear function is therefore represented by the matrix

$$\begin{bmatrix} \dfrac{\partial Z}{\partial X} & \dfrac{\partial Z}{\partial Y} \\ \dfrac{\partial W}{\partial X} & \dfrac{\partial W}{\partial Y} \end{bmatrix}_{(X_0,Y_0)} \tag{7.2}$$

which is called the *Jacobian matrix* or just the Jacobian. It acts on $(\Delta X, \Delta Y)$ to produce $(\Delta Z, \Delta W)$. The matrix equation is therefore

$$\begin{bmatrix} \dfrac{\partial Z}{\partial X} & \dfrac{\partial Z}{\partial Y} \\ \dfrac{\partial W}{\partial X} & \dfrac{\partial W}{\partial Y} \end{bmatrix}_{(X_0,Y_0)} \begin{pmatrix} \Delta X \\ \Delta Y \end{pmatrix} = \begin{pmatrix} \Delta Z \\ \Delta W \end{pmatrix}$$

If $V = (f, g)$, then the matrix

$$\begin{bmatrix} \dfrac{\partial f}{\partial X} & \dfrac{\partial f}{\partial Y} \\ \dfrac{\partial g}{\partial X} & \dfrac{\partial g}{\partial Y} \end{bmatrix}_{(X_0,Y_0)}$$

represents the linear approximation to $V$ at the point $(X_0, Y_0)$. It is called the Jacobian matrix of $V$ at the point $(X_0, Y_0)$.

**Exercise 7.4.7** Find the Jacobian of the function developed in this section at $X = 1$, $Y = 1$.

### $n$ Dimensions

In $n$ dimensions, if

$$V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

is an arbitrary function,

$$V(X_1, X_2, \ldots, X_n) = (Y_1, Y_2, \ldots, Y_n)$$

where

$$Y_i = f_i(X_1, X_2, \ldots, X_n) = a_{1i}X_1 + a_{2i}X_2 + \cdots + a_{ni}X_n$$

then the linear approximation to $V$ at the point $(X_1, X_2, \ldots, X_n)_0$ is given by the Jacobian matrix

$$\begin{bmatrix} \dfrac{\partial f_1}{\partial X_1} & \dfrac{\partial f_1}{\partial X_2} & \cdots & \dfrac{\partial f_1}{\partial X_n} \\ \dfrac{\partial f_2}{\partial X_1} & \dfrac{\partial f_2}{\partial X_2} & \cdots & \dfrac{\partial f_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial f_n}{\partial X_1} & \dfrac{\partial f_n}{\partial X_2}, & \cdots, & \dfrac{\partial f_n}{\partial X_n} \end{bmatrix}_{(X_1, X_2, \ldots, X_n)_0}$$

### Further Exercises 7.4

1. The Earth is round, but in everyday life, we get along fine acting as though it were flat. Why is this possible?

2. Compute the following partial derivatives:
   1. $f(X, Y) = X^3 - Y^3 + 2XY$. Compute $\dfrac{\partial f}{\partial Y}$.
   2. $u(s, t) = 5s^2 - 3st + 6t^3 + 8$. Compute $\dfrac{\partial u}{\partial s}$.
   3. $r(N, P) = 3P(1 + NP) + \log(3N) + e^{2P}$. Compute $\dfrac{\partial r}{\partial P}$.

3. Compute both partial derivatives of $f(X, Y) = 5X^2 + 2Y^3 - 4X^3Y^5$.

4. Compute all three partial derivatives of $g(X, Y, Z) = (X^2 - Y^2)(4X + 2Z) - \dfrac{YZ^3}{X + Z^3}$.

5. Compute the Jacobian matrix of the function
$$g(u, v) = \left(u^2 + v^3 - 2, \frac{u}{v}\right)$$

6. Let $f(X, Y) = 5XY - 3X^2 - Y^2$.
   1. Compute both partial derivatives of $f$.
   2. Compute $\left.\dfrac{\partial f}{\partial X}\right|_{(1,2)}$ and $\left.\dfrac{\partial f}{\partial Y}\right|_{(1,2)}$. That is, plug $(X, Y) = (1, 2)$ into your answer from part (a).
   3. Write down the linear approximation to $f(X, Y)$ at $(X, Y) = (1, 2)$ in the form
$$\Delta f \approx m \cdot \Delta X + n \cdot \Delta Y$$
   4. Expand your answer from part (c) by rewriting $\Delta f$ as $f(X, Y) - f(1, 2)$ and replacing $\Delta X$ and $\Delta Y$ as in problem 1 above, then solving for $f(X, Y)$.
   5. What is $f(0.97, 2.06)$, approximately?
   6. Use your answer from part (d) to write down the equation for the tangent plane to the graph of $f(X, Y)$ at $(X, Y) = (1, 2)$.

7. From chemistry, you may recall that the ideal gas law states that for $n$ moles of an ideal gas,
$$PV = nRT$$
where $R = 0.082$, and $P$, $V$, and $T$ are the pressure (in atmospheres), volume (in liters), and temperature (in kelvins), respectively. Suppose you have one mole of an ideal gas, so that its volume is
$$V = \frac{0.082T}{P}$$
Suppose the current pressure is 1 atm, and the current temperature is 300 K. Use a linear approximation to estimate how much the volume of the gas will *change* if the pressure increases by 0.1 atm *and* the temperature increases by 3 K.

8. The force of gravity exerted on Earth by the Moon is responsible for many phenomena that have a significant impact on biological systems, such as the level and frequency of high and low tides. This force is
$$f(M, R) = 398600\frac{M}{R^2}$$
where $M$ is the mass of the Moon and $R$ is its distance from Earth. Currently, $M = 73480 \times 10^{18}$ kg and $R = 384400$ km (on average), and these haven't changed much in several million years. But suppose an asteroid of mass $250 \times 10^{18}$ kg collides with the Moon, causing its mass to increase by that amount and shifting the Moon's orbit so that it is 400 km closer to Earth! Using a linear approximation to estimate how much the Moon's gravitational pull on Earth will change.

## 7.5 Linear Approximations to Multivariable Vector Fields

We can now return to our actual goal: using linearization to learn about the stability of equilibria of nonlinear differential equations. We did this for one-variable systems earlier and will now develop a way to do it for multivariable systems. First, however, we need some assurance that this can, in fact, be done. This assurance comes in the form of the *Hartman–Grobman theorem*: near an equilibrium point, a vector field behaves like its linear approximation. We already used this principle, the principle of linearization, in one dimension, but it holds in any number of dimensions.

As a technical note, we have to keep in mind that here, as in all applications of the Hartman–Grobman theorem, we have to assume that the real part of the eigenvalue is not zero. Cases in which $\lambda = 0$ or $\lambda = \pm \boldsymbol{i}$ are atypical and fragile: their behavior is qualitatively altered by even the tiniest perturbation. So in general, cases in which the real part of the eigenvalue is zero have to be dealt with by special handling; we can't directly infer the quality of the nonlinear equilibrium point from the linearization. There are some exceptions to this, which we will use in our discussions of the shark–tuna model and the pendulum.

Please note that the condition of this theorem is that we are *near* an equilibrium point. The condition that linearization works only near an equilibrium point is critical. Far from an equilibrium point, all bets are off, and we have only simulation as a tool to study the system's behavior.

So how do we go about finding a linear approximation to a vector field at a point? We have already seen that the linear approximation to an $n$-dimensional function

$$V : \mathbb{R}^n \longrightarrow \mathbb{R}^n$$

at a point $(X_1, X_2, \ldots, X_n)_0$ is given by the Jacobian

$$\begin{bmatrix} \frac{\partial f_1}{\partial X_1} & \frac{\partial f_1}{\partial X_2} & \cdots & \frac{\partial f_1}{\partial X_n} \\ \frac{\partial f_2}{\partial X_1} & \frac{\partial f_2}{\partial X_2} & \cdots & \frac{\partial f_2}{\partial X_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial X_1} & \frac{\partial f_n}{\partial X_2}, & \cdots, & \frac{\partial f_n}{\partial X_n} \end{bmatrix}_{(X_1, X_2, \ldots, X_n)_0}$$

where $f_1, f_2, \ldots, f_n$ are the $n$ component functions of the vector field $V$, each of which is a function $\mathbb{R}^n \to \mathbb{R}$.

This Jacobian defines a linear function $\mathbb{R}^n \to \mathbb{R}^n$, which gives us a linear vector field. We call this vector field

$$\boldsymbol{D}_{V(X_1, X_1, \ldots, X_n)_0}$$

which we read as "the ***D***erivative of $V$ at the point $(X_1, X_1, \ldots, X_n)_0$."

Let's call this linear vector field $\boldsymbol{D}$ for short.

As we saw in Chapter 6, we determine the stability of the equilibrium point by finding the *eigenvalues* of $\boldsymbol{D}$, which are the solutions to

$$\left|\boldsymbol{D} - \lambda \mathbb{I}\right| = 0$$

Recall that the eigenvalues decompose $\boldsymbol{D}$ into subspaces along which $\boldsymbol{D}$ acts like a

- stable equilibrium point ($\lambda < 0$) (1D subspace) or
- unstable equilibrium point ($\lambda > 0$) (1D subspace) or
- stable spiral ($\lambda = -a \pm b\boldsymbol{i}$) (2D subspace) or
- unstable spiral ($\lambda = +a \pm b\boldsymbol{i}$) (2D subspace).

Therefore, we know how to find the linear approximation to $V$, and we know how to find the stability of a linear vector field. Now we can put the two together:

To determine the stability of an equilibrium point of a vector field $V : \mathbb{R}^n \to \mathbb{R}^n$:
(1) Find the linearization of $V$ at the equilibrium point, which is the Jacobian.
(2) Determine the stability of this linear function, using the method of eigenvalues.
(3) Provided no eigenvalue is zero or has zero real part, conclude that the equilibrium point of the nonlinear system is qualitatively similar to that of its linearization.

This is the *Hartman–Grobman theorem* in $n$ dimensions.

**Exercise 7.5.1** Why didn't we need to compute the Jacobian when we were working with linear systems?

### Example: The Rayleigh Oscillator

Recall the Rayleigh vector field from Chapter 4:

$$X' = V$$
$$V' = -X - (V^3 - V)$$

It has a single equilibrium point, at $(X, V) = (0, 0)$. Let's determine the stability of that equilibrium point.

First, we calculate the Jacobian matrix and evaluate it at the point $(0, 0)$:

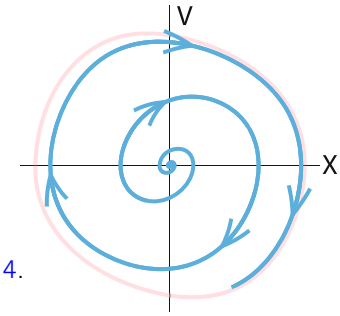
$$\begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial V} \\ \frac{\partial V'}{\partial X} & \frac{\partial V'}{\partial V} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & -3V^2 + 1 \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix}$$

Then we determine the stability of this linear function by calculating the eigenvalues,

$$det\left( \begin{bmatrix} 0 & 1 \\ -1 & 1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} 0 - \lambda & 1 \\ -1 & 1 - \lambda \end{vmatrix} = 0$$

This implies

$$\lambda^2 - \lambda + 1 = 0$$




which gives us

$$\lambda = \frac{+1 \pm \sqrt{1-4}}{2} = +\frac{1}{2} \pm \sqrt{3}\, i$$

So the linear approximation is an unstable spiral! This confirms the results of our simulations of the Rayleigh oscillator in Chapter 4.

**Exercise 7.5.2** Repeat this analysis for a situation in which the clarinet player is blowing harder, modeled by the equation

$$X' = V$$
$$V' = -X - (0.4V^3 - V)$$

### Example: Can Two Species Coexist?

As another example of this procedure, let's look at the second deer–moose competition model from Chapter 3, where $D$ = deer population and $M$ = moose population. We want to know whether the two species can coexist, or in other words, whether the equilibrium point at which both species have nonzero populations is stable.

The model describing the system is

$$D' = 3D - 2MD - D^2$$
$$M' = 2M - DM - M^2$$

Recall from Chapter 3 that the nontrivial equilibrium point of this system is

$$(D, M) = (1, 1)$$

The Jacobian of this system evaluated at the point $(1,1)$ is

$$\begin{bmatrix} \frac{\partial D'}{\partial D} & \frac{\partial D'}{\partial M} \\ \frac{\partial M'}{\partial D} & \frac{\partial M'}{\partial M} \end{bmatrix}_{(1,1)} = \begin{bmatrix} 3-2D-2M & -2D \\ -M & 2-D-2M \end{bmatrix}_{(1,1)} = \begin{bmatrix} -1 & -2 \\ -1 & -1 \end{bmatrix}$$

The eigenvalues are the solutions to

$$det\left( \begin{bmatrix} -1 & -2 \\ -1 & -1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -1-\lambda & -2 \\ -1 & -1-\lambda \end{vmatrix} = \lambda^2 + 2\lambda - 1 = 0$$

which gives

$$\lambda = -1 \pm \sqrt{2} \quad \Longrightarrow \quad \lambda = +0.41,\ \lambda = -2.41$$

indicating that the equilibrium point is an unstable saddle point. Therefore, *with these parameter values*, the two species cannot coexist.

**Exercise 7.5.3** Find and classify the other equilibrium points of this system.

**Exercise 7.5.4** Another deer–moose competition model we studied in Chapter 3 was

$$\begin{aligned} D' &= 3D - MD - D^2 \\ M' &= 2M - 0.5MD - M^2 \end{aligned} \tag{7.3}$$

Determine whether the deer and moose can coexist with these parameter values.

### When Linearization Fails: The Zero Eigenvalue

We've been using the very powerful tool that is the Hartman–Grobman theorem. It gives us the right to take a nonlinear system at an equilibrium point, find its linearization, study it, and then determine the stability of the original nonlinear equilibrium point.

However, there are two technical conditions that must be met before we can apply the theorem.

The first is that none of the eigenvalues of the linearized system is zero. Suppose this were not so, that is, suppose we had a system with two eigenvalues $\lambda_1$ and $\lambda_2$. Let's say $\lambda_1$ is $-a$ $(a > 0)$ and $\lambda_2$ is 0. This means that the 2D system can be split into two 1D axes, $\mathbf{U}$ and $\mathbf{V}$, with the system acting like $\mathbf{U}' = -a\mathbf{U}$ along $\mathbf{U}$ and $\mathbf{V}' = 0\mathbf{V} = 0$ along $\mathbf{V}$.

This means that there is an axis along which the state point is not changing ($\mathbf{V}' = 0$) and another one along which it is shrinking ($\mathbf{U}' = -a\mathbf{U}$).

**Exercise 7.5.5** Sketch a diagram of this situation.

A typical case is

$$\begin{aligned} X' &= X - 2Y \\ Y' &= 3X - 6Y \end{aligned}$$

represented by the matrix

$$M = \begin{bmatrix} 1 & -2 \\ 3 & -6 \end{bmatrix}$$

The eigenvalues of this matrix are solutions to the characteristic equation

$$\lambda^2 + 5\lambda = 0$$

which gives us

$$\lambda_1 = 0, \lambda_2 = -5$$

The first eigenvector is found by solving

$$M\mathbf{U} = \lambda_1 \mathbf{U}$$

$$M\mathbf{U} = \begin{bmatrix} 1 & -2 \\ 3 & -6 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X - 2Y \\ 3X - 6Y \end{pmatrix} = \lambda_1 \mathbf{U} = 0 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{cases} X - 2Y = 0 & \Longrightarrow \quad Y = 0.5X \\ 3X - 6Y = 0 & \Longrightarrow \quad Y = 0.5X \end{cases}$$

So the first eigenvector is any vector on the line $Y = 0.5X$, for example, $(X, Y) = (2, 1)$. The second eigenvector is found by solving

$$M\mathbf{V} = \lambda_2 \mathbf{V}$$

$$M\mathbf{V} = \begin{bmatrix} 1 & -2 \\ 3 & -6 \end{bmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X - 2Y \\ 3X - 6Y \end{pmatrix} = \lambda_2 \mathbf{V} = -5 \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} -5X \\ -5Y \end{pmatrix}$$

$$\begin{cases} X - 2Y = -5X & \Longrightarrow \quad Y = 3X \\ 3X - 6Y = -5Y & \Longrightarrow \quad Y = 3X \end{cases}$$

So the second eigenvector is any vector on the line $Y = 3X$, for example, the vector $(X, Y) = (1, 3)$. The resulting phase portrait is as follows (Figure 7.23):

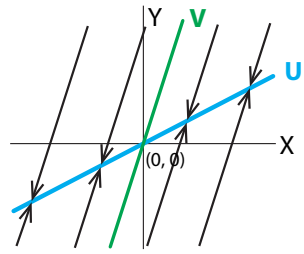



Figure 7.23: A dynamical system that has a zero eigenvalue and a negative eigenvalue will converge toward the eigenvector corresponding to the zero eigenvalue (**U** axis). In this system, every point on the **U** axis is an equilibrium point.

We see that the system does not have an isolated equilibrium point; instead, it has a line of equilibrium points: every point on the line $Y = 0.5X$ (the blue **U** eigenvector) is an equilibrium point.

This is a situation we have not seen before. There is what some writers call an "absorbing final state": every initial condition will approach some definite final state, but the final state depends on the initial condition.

**Exercise 7.5.6** Simulate this system for at least three different initial conditions and plot the trajectories. (You may want to overlay them.) Describe what happens.

The problem with systems like this is that they are not *robust*: adding even the tiniest, vanishingly small additional forces will yield qualitatively different systems. For example, let's add a tiny additional factor $\epsilon$ (epsilon) to the vector field to make it

$$X' = X - 2Y$$
$$Y' = (3 - \epsilon)X - 6Y$$

represented by the matrix

$$\mathbf{M} = \begin{bmatrix} 1 & -2 \\ 3 - \epsilon & -6 \end{bmatrix}$$

Note that the addition of the factor $\epsilon$ changed the nature of the point to either an unstable saddle or a stable node, depending on the sign of $\epsilon$ (Figure 7.24).

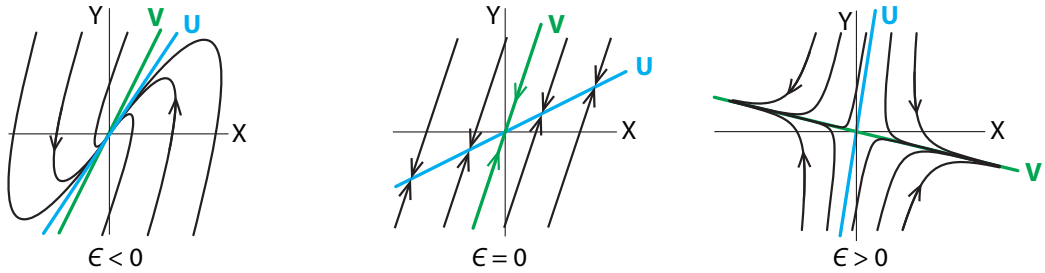



Figure 7.24: In the system $X' = X - 2Y, Y' = (3 - \epsilon)X - 6Y$, the equilibrium point changes from a stable spiral to a saddle point when the parameter $\epsilon$ goes from slightly negative to slightly positive.

Robust systems are called "structurally stable," and some writers suggest that every mathematical model of a natural system must be structurally stable (Abraham and Marsden 1978).[1] Note that this is a new concept of stability: structural stability means that the *vector field* is stable, not that points are stable.

The important thing to remember is that when a system is qualitatively susceptible to tiny changes in the dynamics, all bets are off when it comes to determining the stability of the nonlinear system. When the linearization is not even locally robust, a locally tiny difference between the system near its equilibrium point and the linearized version can result in qualitatively different dynamics. When you are faced with such a system in real life, consult a specialist for the technical math, and realize that we can always rely on simulation of the full nonlinear system, taking care to use very small time steps $\Delta t$, because the system is very sensitive to slight changes.

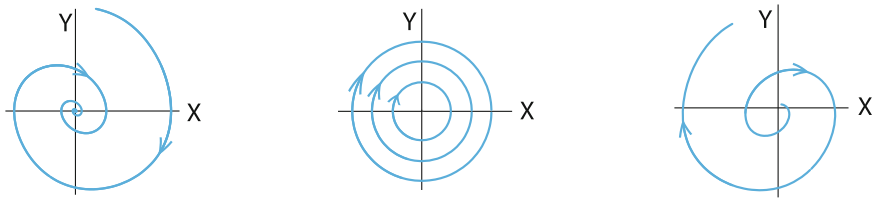
[1] Philosophers of science have also argued for the idea that a good explanation must be stable under small perturbations of its assumptions. It appears explicitly in the writings of the early twentieth-century philosopher Pierre Duhem (see the discussion in Garfinkel (1981)) and was used by philosophers in the later twentieth century to argue against certain kinds of reductionist explanations (see Putnam (1975) and Garfinkel (1981)).

### When Linearization Fails: Purely Imaginary Eigenvalues

The second type of case in which linearization fails occurs when the eigenvalues of the linear approximation are purely imaginary, $\lambda = \pm k\boldsymbol{i}$ (we will let $k = 1$ for convenience).

We know what this linearization looks like: it is a *center*.

The problem with a center is similar to the problem of the zero eigenvalue above: neither of these vector fields is structurally stable, and the tiniest additional force will turn the center into a spiral.




Just as in the case of the zero eigenvalue, the fact that the linearized system is not robust means that all bets are off when it comes to deciding the character of the equilibrium point of the nonlinear system.

**Exercise 7.5.7** Hartman–Grobman fail. Here's a pathological example in which linearization fails to give the right answer, because the eigenvalues are purely imaginary. Let

$$f(X,Y) = \frac{X^2+Y^2}{1+X^2+Y^2}$$

and consider the differential equation

$$X' = -Y + f(X,Y)\cdot X$$
$$Y' = X + f(X,Y)\cdot Y$$

a) Plot some trajectories for this vector field and show that $(0,0)$ is an unstable spiral equilibrium point.

b) Then calculate the linear approximation to this vector field, that is, the Jacobian

$$\mathbf{M}_{(0,0)} = \begin{bmatrix} \dfrac{\partial X'}{\partial X} & \dfrac{\partial X'}{\partial Y} \\ \dfrac{\partial Y'}{\partial X} & \dfrac{\partial Y'}{\partial Y} \end{bmatrix}_{(0,0)}$$

and show that it predicts that $(0,0)$ is a center.

However, there is one special class of nonlinear systems in which we can conclude that the equilibrium point *is* a center. It will help us solve both the pendulum and shark–tuna models.

The special class is the case of *conservative systems*. A system is said to be *conservative* if there is some continuous quantity $H$ that is constant on every trajectory, so that $H$ does not change over time ($\frac{dH}{dt} = 0$).

If there is such a conserved quantity in a given system, the consequences for the dynamics of the system are very strong. As Strogatz points out (Strogatz 2014), *conservative systems cannot have stable equilibrium points or limit cycle attractors*. They can have only centers and saddle points.

We said back in Chapter 4 that what we wanted in a model of a biological oscillation was that the oscillation be robust, that is, that it have a limit cycle attractor. *Conservative systems cannot have limit cycle attractors, and therefore they are not good models for biological systems.*

Yet even though conservative systems violate the *axiom of stability* that we mentioned in the previous section, they can be useful models for some purposes. But we have to be careful with them.

The major fact about conservative systems is that for such systems, we can sometimes prove that a nonlinear equation has a center, in spite of the inapplicability of the Hartman–Grobman theorem.

There's a helpful theorem.[2] Let $V(X,Y)$ be a two-dimensional vector field, and let $(X_0,Y_0)$ be an isolated equilibrium point of $V$. Suppose $V$ is a conservative system, that is, that there is some function $H(X,Y)$ that is constant on trajectories. If $(X_0,Y_0)$ is a local minimum (or maximum) of $H$ (see Section 7.7 for the notion of local maxima and minima), then $(X_0,Y_0)$ is a center equilibrium, and all orbits in a neighborhood around $(X_0,Y_0)$ are closed.

**Exercise 7.5.8** When could an equilibrium point not be isolated?

We will now apply this principle to two fundamental examples: the shark–tuna model and the frictionless pendulum.

### Example: Shark–Tuna

The shark–tuna vector field

$$S' = ST - S$$
$$T' = -ST + T$$

has two equilibrium points, $(S,T) = (0,0)$ and $(S,T) = (1,1)$. The linearization of the shark–tuna vector field is

$$\begin{bmatrix} \frac{\partial S'}{\partial S} & \frac{\partial S'}{\partial T} \\ \frac{\partial T'}{\partial S} & \frac{\partial T'}{\partial T} \end{bmatrix} = \begin{bmatrix} T-1 & S \\ -T & -S+1 \end{bmatrix}$$

Evaluated at the point $(0,0)$, this gives us the matrix

$$\begin{bmatrix} T-1 & S \\ -T & -S+1 \end{bmatrix}_{(0,0)} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

The eigenvalues are the solutions to

$$det\left( \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -1-\lambda & 0 \\ 0 & 1-\lambda \end{vmatrix} = \lambda^2 - 1 = 0$$

which gives us

$$\lambda = \pm 1$$

This is an unstable saddle point at $(0,0)$. Calculating the eigenvectors corresponding to these eigenvalues, we see that the eigenvector corresponding to the positive eigenvalue is the $T$-axis, which is $S = 0$, and the eigenvector corresponding to the negative eigenvalue is the $S$-axis. The equilibrium point $(0,0)$ is stable in the $S$-axis and unstable in the $T$-axis.

[2] Theorem 6.5.1 in Strogatz (2014).

**Exercise 7.5.9** Why does this make biological sense?

At the second equilibrium point $(S, T) = (1, 1)$, the Jacobian is

$$\begin{bmatrix} T-1 & S \\ -T & -S+1 \end{bmatrix}_{(1,1)} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

The eigenvalues are the solutions to

$$\lambda^2 + 1 = 0 \quad \Longrightarrow \quad \lambda = \pm i$$

Here, the equilibrium point $(1, 1)$ has eigenvalues that are purely imaginary. We recall that the condition of the Hartman–Grobman theorem is that for the theorem to apply, eigenvalues must *not* be purely imaginary. Therefore, we have to resort to other methods to show that $(1, 1)$ is a center.

Our theorem about conserved quantities comes to the rescue. The shark–tuna equations (whose formal name is the Lotka–Volterra equations) have a conserved quantity.

If we write the model as

$$S' = aST - dS$$
$$T' = cT - dST$$

then we can show that

$$H = c \ln S(t) - dS(t) - aT(t) + b \ln T(t)$$

is a conserved quantity and that $H$ has a maximum at the equilibrium point, which is $(S, T) = (\frac{c}{d}, \frac{b}{a})$ (Figure 7.25).

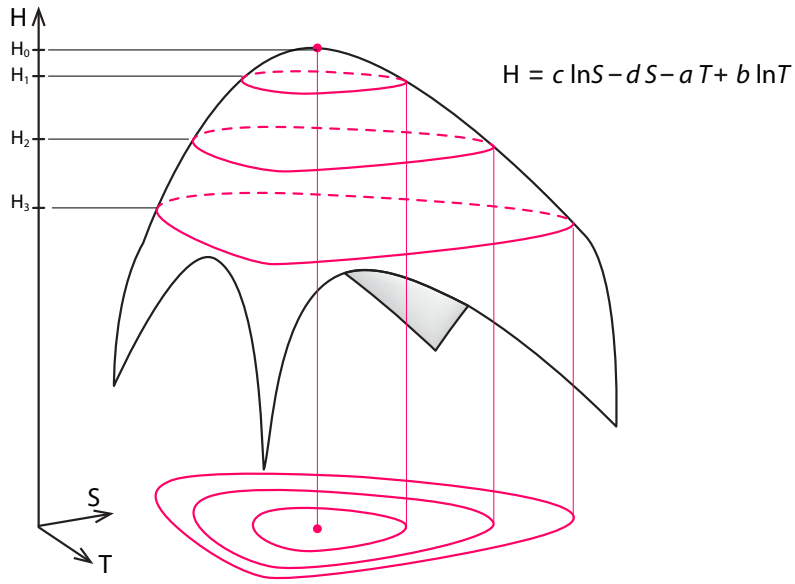


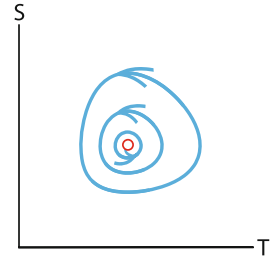

Figure 7.25: In the shark-tuna dynamical system, the quantity $H$ remains constant along all trajectories, meaning it is a conserved quantity. Since the graph has a local maximum $H_0$, the trajectories around it are closed.

**Exercise 7.5.10** Verify that $\dfrac{dH}{dt} = 0$. *(Hint: $\frac{d}{dx}\ln x = \frac{1}{x}$. You may also want to review the chain rule.)*

Therefore, the nonzero equilibrium point is a center surrounded by closed orbits.

Of course, this can be verified by simulations from initial conditions close to the equilibrium point.

We said that systems with conserved quantities are poor models for biological systems, and the Lotka–Volterra equations are no exception. Indeed, we already saw, in the discussion of the Holling–Tanner model in Chapter 4, that the Lotka–Volterra equations depended on unrealistic assumptions and that more realistic ones resulted in a system with a limit cycle attractor.




## Example: The Pendulum

The simple pendulum (Figure 7.26) gives us a great example of the power of nonlinear dynamics.

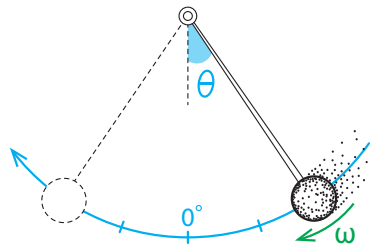



Figure 7.26: The pendulum. Its state variables are angular position $\theta$ and angular velocity $\omega$.

First of all, let's think about the essential dynamics. Since we are in the world of "mechanics," we can immediately write

$$X' = V$$
$$V' = -F$$

where, as usual in mechanics, $X$ is a physical space (position) variable and $V$ is a velocity variable. This is the form of "$F = ma$" stated in the language of differential equations.

But what are the correct $X$ and $V$ for the pendulum? The physical position of the pendulum is actually given not by a distance $X$, but by an *angle*, which is typically called by the Greek letter $\theta$ (theta).

Angle variables are very different from distance variables. Distances live on the real line $\mathbb{R}$. You can be one foot to the left or right of 0 (that is, $-1$ ft or $+1$ ft, or 50,000 miles to the left or right ($-50{,}000$ mi or $+50{,}000$ mi). The scale on $\mathbb{R}$ goes from $-\infty$ to $+\infty$, with each point, each value, representing a distinct position or state.

Not so for angles. The angle 360° *is* the angle 0°; the angle 370° *is* the angle 10°. So angles don't go on and on forever; they repeat after 360°.[3]

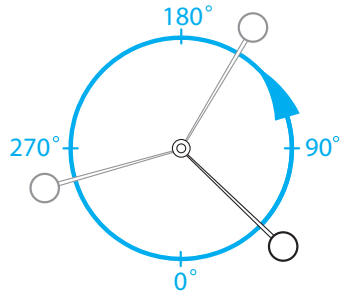



Figure 7.27: How the circle represents angles. The circle is referred to as $S^1$.

Therefore, the state space of angles has a different shape from that of the line that represents $\mathbb{R}$: it's a closed circle, not a line. Angles live on a circle, called $S^1$. This is our first example of a state space that is not $\mathbb{R}^n$ (Figure 7.27).

**Exercise 7.5.11** Come up with another example of a variable whose state space is a circle.

Then we need to find the state space for the velocity variable. This really is $\mathbb{R}$, since any positive value of velocity is possible, as is any negative value, and no two values are equivalent. Of course, the velocity here is angular velocity (speed and direction of rotation), typically called by the Greek letter $\omega$ (omega), so $\omega$-space is $\mathbb{R}$.

So now what is the joint state space for $(\theta, \omega)$? The angular position $\theta$ lives in $S^1$, and the angular velocity $\omega$ lives in $\mathbb{R}$, so the joint state space is $S^1 \times \mathbb{R}$, the set of all pairs $(\theta, \omega)$, where $\theta$ is in $S^1$ and $\omega$ is in $\mathbb{R}$. This is the same kind of construction that we used to make the state space for the spring, which is the set of all pairs $(X, V)$, where $X$ is the position and $V$ is the velocity. $S^1 \times \mathbb{R}$ is called the *Cartesian product* of $S^1$ and $\mathbb{R}$.

**Exercise 7.5.12** Give two examples of points in $S^1 \times \mathbb{R}$.

**Exercise 7.5.13** Give an example of two points in $S^1 \times \mathbb{R}$ that are actually the same point.

The space $S^1 \times \mathbb{R}$ looks like a cylinder (Figure 7.28). Notice that on the cylinder, specifying a point $\omega_0$ on the green $\omega$ axis and specifying an angle $\theta_0$ uniquely determines a point on the cylinder.

[3] Or if you prefer, $2\pi$ radians.

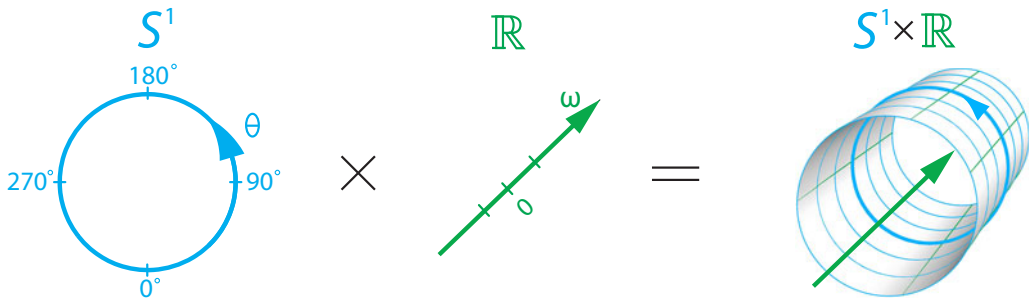



Figure 7.28: If one variable has a state space that is a circle and another variable has a state space that is a line, their joint state space is a cylinder.

That's our state space for the pendulum. Now let's go on to describe the dynamics by completing the differential equation. First, what is $F$ here? It's the force of gravity acting on the pendulum weight, which is of course equal to $mg$, where $m$ is the mass of the pendulum and $g$ is the acceleration due to gravity (its value is around 32 ft/sec$^2$).

However, the force of gravity is always acting straight down. Only part of that force is going to make the pendulum swing, and that is the part that is along the curve of movement, tangent to it, and perpendicular to the shaft of the pendulum. The other component, at right angles, is the part of the force that is acting along the line of the shaft, which is assumed to have no effect (Figure 7.29).

**Exercise 7.5.14** Briefly explain why this makes physical sense.

Therefore, the true force acting to change the angle is not $mg$ but rather $mg\sin\theta$.

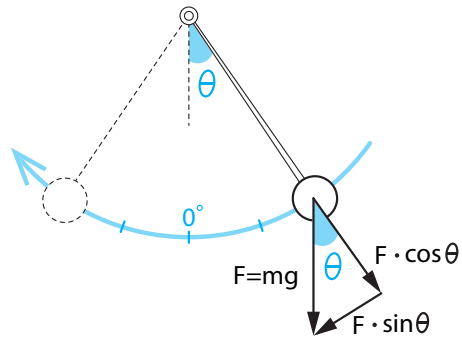



Figure 7.29: We can decompose the gravitational force F = mg into a component acting along the pendulum shaft ($F \cdot \cos\theta$) and a component acting perpendicular to the shaft ($F \cdot \sin\theta$).

**Exercise 7.5.15** Why $\sin\theta$? *(Hint: Think about vector addition and recall (or look up) basic trigonometry.)*

We now have our differential equation

$$\theta' = \omega$$

$$\omega' = -mg \sin\theta$$

By choosing a unit system in which $mg = 1$, the differential equation reduces to

$$\theta' = \omega$$
$$\omega' = -\sin\theta$$

Notice that this is highly nonlinear: it is certainly *not* the case that the sine function is linear; $\sin(X + Y)$ is definitely not $\sin X + \sin Y$, and $\sin(6 \times 30^\circ)$ is not equal to $6 \times \sin 30^\circ$.

**Exercise 7.5.16** Confirm these statements numerically.

So we have a nonlinear equation here, and paper and pencil methods are not going to solve it. Let's use the methods of this chapter to analyze this system.

*Finding equilibrium points.* The first step, as always, is to find the equilibrium points and determine their stability. If we set the right-hand side of the differential equation to 0, we get

$$0 = \omega$$
$$0 = -\sin\theta$$

Looking at the first equation, we see that every equilibrium point must have $\omega = 0$. This is intuitively clear, since it says that the pendulum must be at rest (angular velocity = $\omega$ = 0). Turning to the second equation, $-\sin\theta = 0$, what values of $\theta$ satisfy this? Looking at the graph of $\sin\theta$, we see two equilibrium points, $\theta = 0^\circ$ and $\theta = 180^\circ$. They have a physical meaning: $\theta = 0$ is rest at bottom dead center, and $\theta = 180^\circ$ means rest at top dead center. The two equilibrium points are therefore $(\theta, \omega) = (0, 0)$ and $(\theta, \omega) = (180^\circ, 0)$ (Figure 7.30).

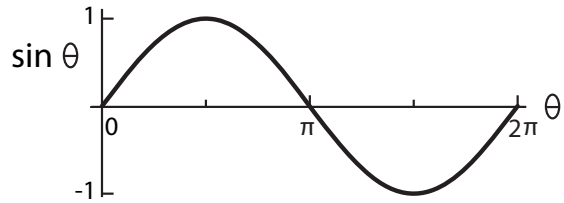



Figure 7.30: The function $\sin\theta$.

*Stability.* The next step is, as always, to determine the stability of these equilibrium points. Previously, we could only use simulation methods to determine stability in 2D or higher. Now we can use the methods of local linear approximation around the equilibrium point to analyze the stability of the equilibrium points.

In order to find the stability of the equilibrium point at $(\theta, \omega) = (0, 0)$, we begin by finding the Jacobian, the matrix of partial derivatives, that represents the linearization of the system at the point $(\theta, \omega) = (0, 0)$. From the definition of the Jacobian matrix (Equation 7.2 on page 385), the linear approximation is given by the matrix

$$\begin{bmatrix} \frac{\partial\theta'}{\partial\theta} & \frac{\partial\theta'}{\partial\omega} \\ \frac{\partial\omega'}{\partial\theta} & \frac{\partial\omega'}{\partial\omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}$$

If we evaluate this matrix at $(\theta, \omega) = (0, 0)$, we get

$$\begin{bmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \omega} \\ \frac{\partial \omega'}{\partial \theta} & \frac{\partial \omega'}{\partial \omega} \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

We find the eigenvalues of this matrix by solving

$$det\left( \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} 0-\lambda & 1 \\ -1 & 0-\lambda \end{vmatrix} = 0$$

The eigenvalues of this matrix are the solutions to $\lambda^2 + 1 = 0$, so the eigenvalues are purely imaginary: $\lambda = \pm \boldsymbol{i}$. Therefore, we can't directly apply the Hartman–Grobman theorem. However, we mentioned that there are certain cases in which we *can* say that the nonlinear system has a center when the linear system does.

Recall our discussion of the shark–tuna system at the equilibrium point $(1, 1)$: we said that the equilibrium point must be a center, because there is a *conserved quantity*, and the equilibrium point is a local maximum of that conserved quantity. The same thing is true of the frictionless pendulum at $(0, 0)$, only now the equilibrium point is a local minimum.

In the pendulum, which is a frictionless mechanical system, there is also a conserved quantity, called "energy." The physical principle of conservation of energy says that the sum of potential and kinetic energy must be a constant. But the kinetic energy is just $\frac{1}{2}\omega^2$, and the potential energy is $-\cos\theta$ (recall $m = 1$ here), so the quantity

$$H = \frac{1}{2}\omega^2 - \cos\theta = E$$

is a constant; hence the equilibrium point of the pendulum at the point $(0, 0)$ is a center.

**Exercise 7.5.17** Verify that $\dfrac{dH}{dt} = 0$.

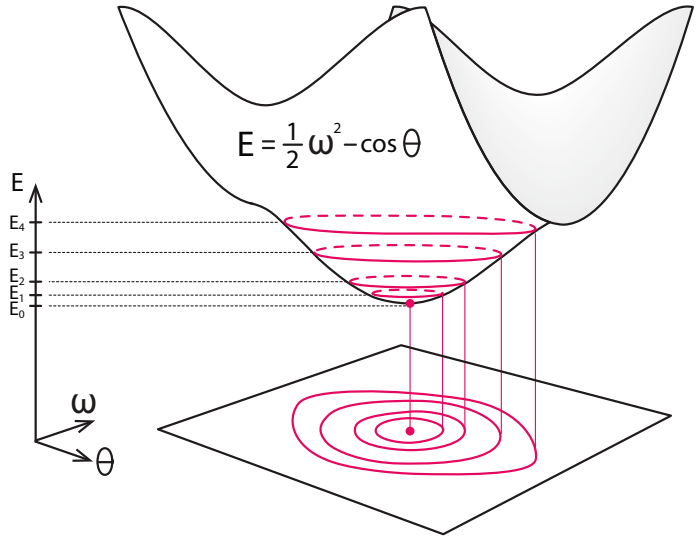



Figure 7.31: In the pendulum dynamical system, the quantity $E$ remains constant along all trajectories, meaning it is a conserved quantity. Since the graph has a local minimum $E_0$, the trajectories around it are closed.

It is easy to confirm that the point $(0, 0)$ is a local minimum of $E$, either using the minimization techniques of Section 7.7 or by plotting $E(\theta, \omega)$ as a surface over $(\theta, \omega)$ space (Figure 7.31).

We can confirm this by simulation using a few initial conditions in a small neighborhood of $(\theta, \omega) = (0, 0)$ (Figure 7.32).

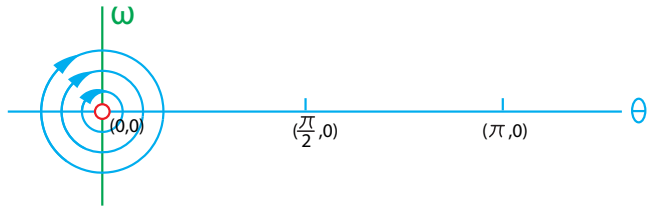



Figure 7.32: Pendulum behavior near the equilibrium point $(0, 0)$.

Clearly, the simulations confirm our calculation: $(\theta, \omega) = (0, 0)$ is a neutral equilibrium. Small perturbations do not go far away, nor do they return to the equilibrium point.

Let's go on to look at the equilibrium point $(\theta, \omega) = (180, 0)$, corresponding to the pendulum at rest at top dead center. You can guess physically what kind of equilibrium this is, but let's do it mathematically. Here the Jacobian is again

$$\begin{bmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \omega} \\ \frac{\partial \omega'}{\partial \theta} & \frac{\partial \omega'}{\partial \omega} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}$$

which, when evaluated at $(180, 0)$, gives us the matrix

$$\begin{bmatrix} \frac{\partial \theta'}{\partial \theta} & \frac{\partial \theta'}{\partial \omega} \\ \frac{\partial \omega'}{\partial \theta} & \frac{\partial \omega'}{\partial \omega} \end{bmatrix}_{(180,0)} = \begin{bmatrix} 0 & 1 \\ -\cos\theta & 0 \end{bmatrix}_{(180,0)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

The eigenvalues of this matrix are given by

$$det\left( \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} 0-\lambda & 1 \\ 1 & 0-\lambda \end{vmatrix} = 0$$

and the eigenvalues are therefore the solutions to $\lambda^2 - 1 = 0$, or $\lambda = \pm 1$. Two purely real eigenvalues, one positive and one negative. That's a saddle point.

Using a large number of simulations to assemble a phase portrait, we get the following picture (Figure 7.33):

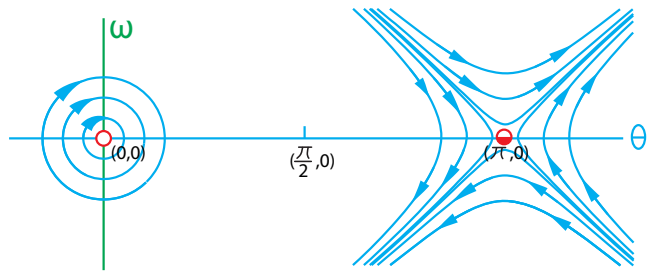



Figure 7.33: Pendulum behavior near the equilibrium points $(0, 0)$ and $(\pi, 0)$.

We have now figured out the behavior near the two equilibrium points. Far from equilibrium, linear approximation methods fail, and our only tool is numerical simulation. If we run a series of simulations to fill in the blank regions, we assemble the complete phase portrait of the pendulum (Figure 7.34). Here we are showing the phase portrait in a plane, using the technique of repeating $\theta$ over and over.

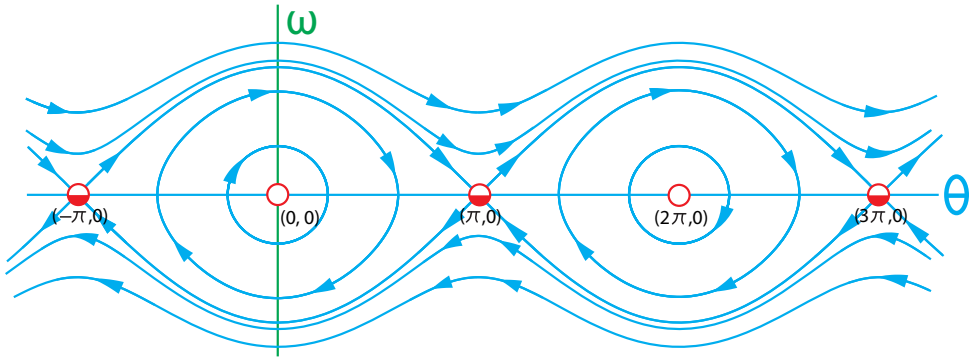



Figure 7.34: Phase portrait of the pendulum.

But really, as we said, the state space is a cylinder, and the true phase portrait looks like Figure 7.35. Figure 7.34 can be seen as the unrolled version of the cylinder in Figure 7.35.

Studying Figure 7.34, we see that there are two qualitatively different shapes of trajectories: the special trajectories that run from saddle point to saddle point form a shape like an eye. Inside the eye, trajectories are closed loops, which are round near the origin $(0, 0)$ and become more oval as they get nearer to the special trajectories that outline the eye. Outside the eye, they have a very different shape: they do not close, indeed, none ever cross the $\omega = 0$ axis.

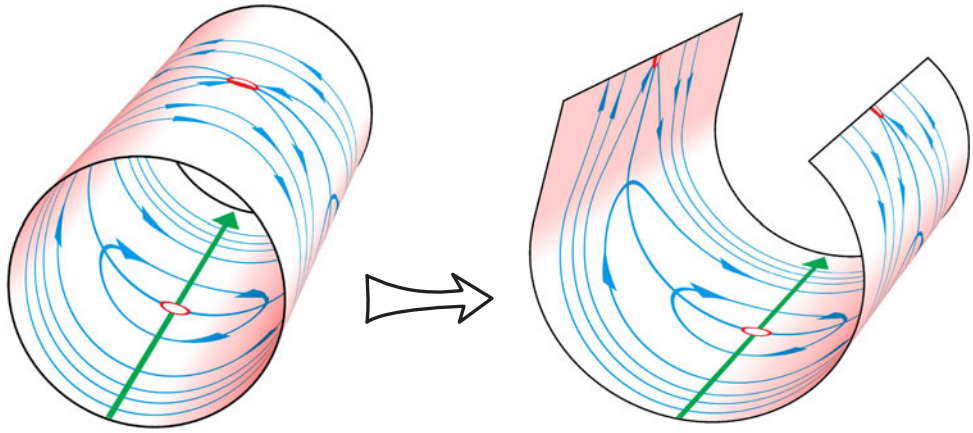


Figure 7.35: The cylindrical state space of the pendulum is best viewed unrolled.

The two types of trajectories represent two different forms of motion:

(1) Inside the eye, the closed loops represent back-and-forth motion of the pendulum around its bottom dead center. For half the cycle, the trajectory is in the positive $\omega$ half-plane: the pendulum is moving to the right. For the other half, the trajectory is below the $\theta$ axis, in negative $\omega$ territory, meaning that the pendulum is now moving back to the left. This motion repeats.

(2) But outside the eye, the trajectories don't cross the axis $\omega = 0$, meaning that the pendulum does not change its direction of motion. These trajectories correspond to motion that is always clockwise (positive $\omega$) or always counterclockwise (negative $\omega$). The pendulum in these cases is whirling around and around in one direction or the other. Not surprisingly, these correspond to higher angular velocities.

So the pendulum gives us an interesting example of a system having two very different forms of motion, depending on initial conditions. The phenomenon of multiple qualitatively different modes of behavior can be seen only in nonlinear systems.

**Exercise 7.5.18** It may seem strange that trajectories that don't seem to form closed loops represent periodic behavior. To understand what's actually happening, sketch Figure 7.34 on a piece of paper (standard-sized printer paper is fine) and wrap it around a cylinder. Describe what happens to the trajectories outside the eye and what this means in physical terms.

### Adding Friction

As we've said, the frictionless pendulum is an idealization. No real system can have zero energy loss. It is therefore interesting to ask what happens if we add a little friction. The model now becomes

$$\theta' = \omega$$
$$\omega' = -\sin\theta - k\omega$$

where $k$ is the friction coefficient. As we might expect, the system is no longer conservative, because energy is not conserved, and so the closed orbits disappear. The equilibrium point at $(0, 0)$ now becomes a stable spiral, and *all* trajectories approach it as $t \to \infty$ (Figure 7.36).

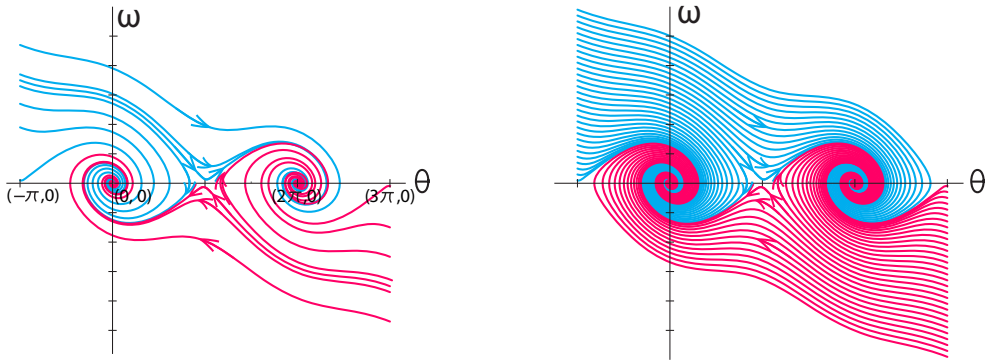



Figure 7.36: Adding friction to the pendulum model converts $(0, 0)$ into a stable equilibrium point.

**Exercise 7.5.19** Pick a few points at random on the phase portrait (Figure 7.36) and follow the trajectory through that point. What is happening to the pendulum as this trajectory is traced out?

Here we are using the technique of the unrolled cylinder representation of state space. The true state space is still the cylinder, and the trajectories now resemble the following figure (Figure 7.37):

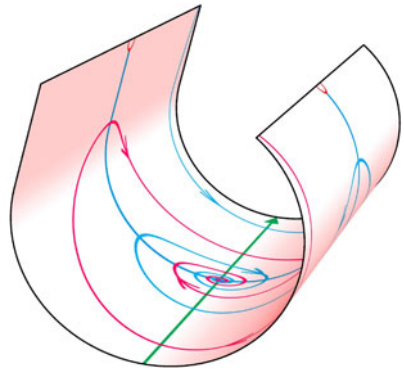


Figure 7.37: The cylindrical state space of the pendulum with friction, unrolled.

### The Linearized "Small-Angle" Pendulum

In some elementary physics and differential equations courses, this nonlinear behavior is considered "too advanced," and so a major simplifying assumption is made to make the system amenable to paper-and-pencil methods.

If we make the drastic assumption that the pendulum is restricted to very small motions, that is, that $\theta$ is close to 0, then we can replace the nonlinear $\sin(\theta)$ term in the $\omega'$ equation. For small angles, $\sin\theta$ is approximately equal to $\theta$ (Figure 7.38).

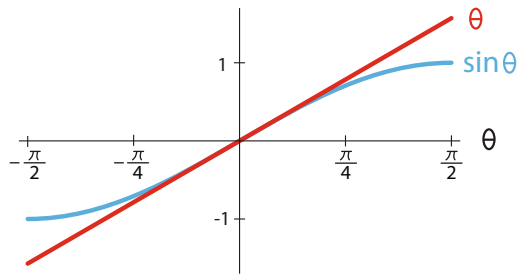



Figure 7.38: $\theta \approx \sin\ \theta$ for small angles $\theta$.

If we make the substitution of $\theta$ for $\sin\theta$, we get a linear differential equation:

$$\theta' = \omega$$
$$\omega' = -\theta$$

**Exercise 7.5.20** Where have we seen this equation (with different variable names) before?

For this simplified system, it is possible to find an explicit solution. In Chapter 2, when we learned about derivatives, we saw that the derivative of $\sin(x)$ is $\cos(x)$ and the derivative of $\cos(x)$ is $-\sin(x)$. Therefore, the equations

$$\theta = \sin(t)$$
$$\omega = \cos(t)$$

satisfy the requirement

$$\theta' = \frac{d}{dt}\sin(t) = \cos(t) = \omega$$
$$\omega' = \frac{d}{dt}\cos(t) = -\sin(t) = -\theta$$

So the functions

$$\theta = \sin(t) \quad \text{and} \quad \omega = \cos(t)$$

explicitly solve the linear differential equation above. The simplified small-angle pendulum is a linear system, and has an explicit solution, which simple calculus is able to provide.

But at what cost was this obtained? The simplified equations are incapable of showing the full behavior of the system. The entire equilibrium point at $(\theta, \omega) = (180, 0)$ has been lost, and with it, the possibility of multiple behaviors.

Many elementary calculus and physics courses make this move of drastic simplifications to make paper-and-pencil solutions possible, but we lose most of the interesting behaviors in this way. Using nonlinear dynamics and computer simulation, we have access to the full range of behaviors of systems in nature.

**Exercise 7.5.21** Sketch the phase portrait for the linear pendulum and compare it to the nonlinear one in Figure 7.34.

### Further Exercises 7.5

1. You and a friend are on a giant swing carnival ride. While you try to keep your lunch down, your friend asks, "Why does it feel like we're stopping as we go over the top?"
   a) Briefly explain what's happening.
   b) How would you explain this to your friend, who knows nothing about dynamics?

2. You have already seen a type of 1D vector field that wasn't structurally stable. What was it and why was it sensitive to changes in parameters? You'll probably want to use diagrams in your explanation.

3. Compute the Jacobian of the system of differential equations

$$X' = X(2 - Y) + XY^2$$
$$Y' = \frac{X + Y}{X - Y}$$

4. Consider the system of differential equations

$$N' = N^2 - 2NP$$
$$P' = P\left(1 - \frac{2P}{N}\right)$$

   a) Verify that $N = 2$, $P = 1$ is an equilibrium point of the system of differential equations.

   b) Find the Jacobian of this system *at this point*.

   c) Find the eigenvalues of this Jacobian.

   d) What kind of equilibrium point is this?

5. Let $D$ be the size of a population of deer, and $M$ the size of the population of moose in the same area. The Lotka–Volterra competition model for these species might look like the following:

$$D' = 0.3D - 0.05D^2 - 0.03DM$$
$$M' = 0.2M - 0.04M^2 - 0.02DM$$

   a) This system has four equilibrium points. Find them. (It might help to use a graphical method here, i.e., nullclines.)

   b) Classify each equilibrium point, using the eigenvalues of the Jacobian.

   c) What will happen to these two populations in the long run? Can they coexist?

6. In the Sonoran desert, kangaroo rats ($K$) compete with ants ($A$) for food, since both eat seeds. Suppose the competition is modeled by the equations

$$A' = 3A - 2A^2 - 2AK$$
$$K' = 2K - AK - 3K^2$$

   a) Find and classify all the equilibria for this system.

   b) What will happen to these species in the long run?

7. Consider the following model of Romeo, Juliet, and Juliet's nurse:

$$\begin{cases} R' = JN - \frac{8}{3}R \\ J' = 10(N - J) \\ N' = 28J - N - RJ \end{cases}$$

This system has three equilibrium points, at $(27, 6\sqrt{2}, 6\sqrt{2})$, $(27, -6\sqrt{2}, -6\sqrt{2})$, and $(0, 0, 0)$.

a) Compute the Jacobian of this system.

b) For each equilibrium point, plug the equilibrium point into the Jacobian and use Sage to find its eigenvalues. What type of equilibrium point is each one?

8. Recall the Holling–Tanner model,

$$N' = r_1 N(1 - \frac{N}{k}) - \frac{wN}{d+N}P$$
$$P' = r_2 P(1 - \frac{jP}{N})$$

Find and classify the biologically meaningful equilibria for this model, using the parameter values $r_1 = 1$, $r_2 = 0.1$, $k = 7$, $d = 1$, $j = 1$, and $w = 1$. Feel free to use SageMath to help with the algebra.

## 7.6 Hopf Bifurcation

Hopf bifurcation is the key to understanding oscillatory behavior. In Chapter 4, we said that a Hopf bifurcation occurs when a stable equilibrium point becomes unstable, and it gives way to a stable limit cycle attractor.

We can now study Hopf bifurcation analytically. Previously, we could use only experimental (simulation) methods: choose some parameter values and run multiple simulations. Now we can study Hopf bifurcation using the principle of linearization and the method of eigenvalues.

### The Rayleigh Model

Let's use the Rayleigh clarinet model as our example:

$$X' = V$$
$$V' = -X - c(V^3 - V)$$

We have inserted a parameter $c$ to be our control parameter.

By setting $X' = 0$ and $V' = 0$, we see that the only equilibrium point of this model is $(X, V) = (0, 0)$. Now let's determine its stability. We can leave the parameter $c$ in the model and work with it symbolically in the Jacobian.

The Jacobian of this vector field is

$$\begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial V} \\ \frac{\partial V'}{\partial X} & \frac{\partial V'}{\partial V} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & -c(3V^2 - 1) \end{bmatrix}$$

which evaluated at the equilibrium point $(0, 0)$ gives us

$$\begin{bmatrix} 0 & 1 \\ -1 & -c(3V^2 - 1) \end{bmatrix}_{(0,0)} = \begin{bmatrix} 0 & 1 \\ -1 & c \end{bmatrix}$$

The eigenvalues are therefore given by

$$det\left( \begin{bmatrix} 0 & 1 \\ -1 & c \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -\lambda & 1 \\ -1 & c - \lambda \end{vmatrix} = \lambda^2 - c\lambda + 1 = 0$$

which gives

$$\lambda = \frac{c \pm \sqrt{-4 + c^2}}{2}$$

Note that we have found $\lambda$ as a function of $c$, so it is easy to calculate the effect of $c$ on the eigenvalues.

First let's look at the case $c < 0$. Here we use $c = -0.5$. The eigenvalues are

$$\lambda|_{c=-0.5} = -0.25 \pm 0.97\, \boldsymbol{i}$$

These are complex conjugate eigenvalues with negative real part. Therefore, they represent a stable spiral. The phase portrait looks like Figure 7.39, left.

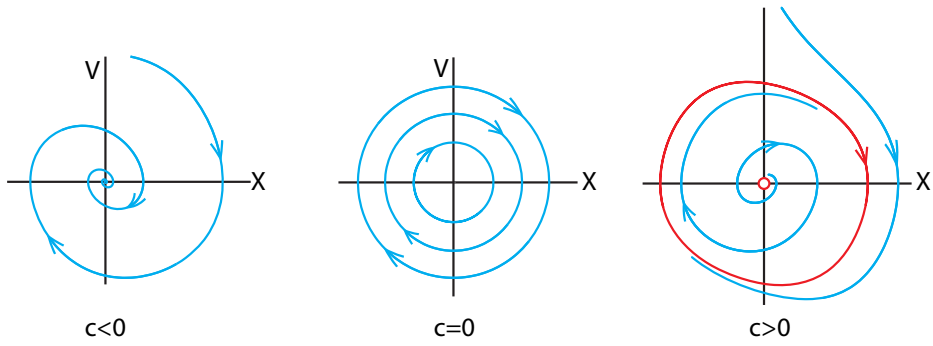



Figure 7.39: In the Rayleigh model, a Hopf bifurcation occurs when parameter $c$ passes from negative to positive.

Now let's look at the case $c > 0$. Here we choose $c = 0.5$. The eigenvalues are

$$\lambda|_{c=0.5} = 0.25 \pm 0.97\, \boldsymbol{i}$$

These are complex conjugate eigenvalues with positive real part. Therefore, they represent an unstable spiral. The phase portrait looks like Figure 7.39, right.

The special case $c = 0$ has a special set of trajectories. The eigenvalues are

$$\lambda|_{c=0} = \pm \boldsymbol{i}$$

which are purely imaginary, indicating a neutral center. The phase portrait looks like Figure 7.39, middle.

If we assemble a set of 2D phase portraits for varying values of $c$ and arrange them in order of their $c$ values, we get the bifurcation diagram for a Hopf bifurcation (Figure 7.40).

**Exercise 7.6.1** At what value of $c$ does the Hopf bifurcation occur?

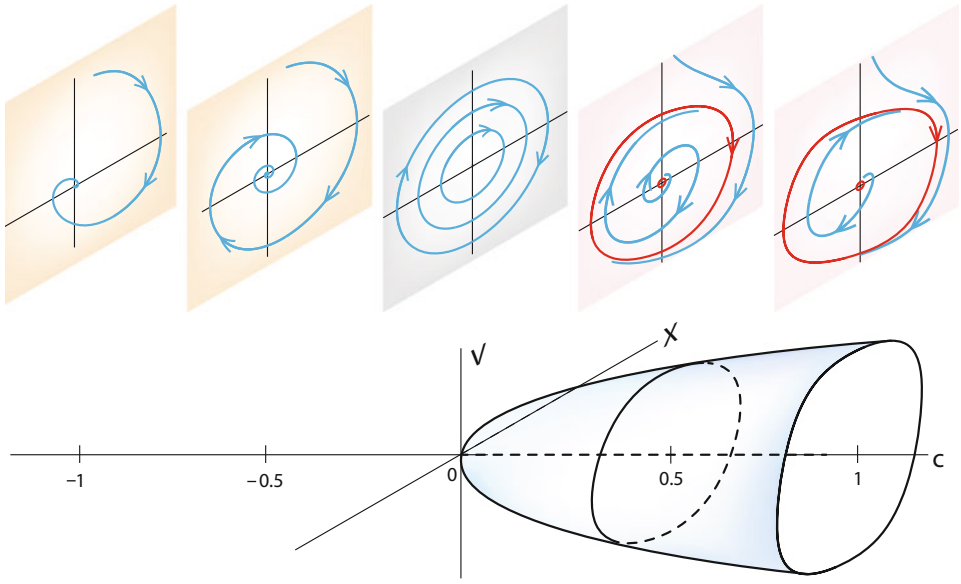



Figure 7.40: A 3D Hopf bifurcation diagram for the Rayleigh clarinet model.

**Hopf bifurcation theorem (approximately).** Consider an equilibrium point of a vector field that depends on a parameter. Let $J$ be the Jacobian matrix representing the linear approximation to the vector field at that equilibrium point. Suppose that a pair of conjugate eigenvalues of $J$, $\mathbf{a} \pm \mathbf{b}i$ passes from $\mathbf{a} < 0$ to $\mathbf{a} > 0$ as a parameter passes a critical value. In this case, the behavior changes from a stable equilibrium to an unstable equilibrium surrounded by a stable limit cycle attractor.

### Example: Glycolysis

In Chapter 4, we saw oscillations in metabolism in the energy-producing reactions of glycolysis. We studied the Selkov model

$$S' = v_0 - cSP^2$$
$$P' = cSP^2 - kP$$

Let's study the dynamics of this model analytically. We will set $V_0 = 1$ and $k = 1$. Our control parameter will be $c$.

Setting $S' = P' = 0$, we see that the model has an equilibrium point at

$$(S, P) = (\frac{1}{c}, 1)$$

To study the stability of this equilibrium point, we calculate the Jacobian

$$\begin{bmatrix} \frac{\partial S'}{\partial S} & \frac{\partial S'}{\partial P} \\ \frac{\partial P'}{\partial S} & \frac{\partial P'}{\partial P} \end{bmatrix} = \begin{bmatrix} -cP^2 & -2cPS \\ cP^2 & 2cPS - 1 \end{bmatrix}$$

evaluated at $(\frac{1}{c}, 1)$,

$$\begin{bmatrix} -cP^2 & -2cPS \\ cP^2 & 2cPS - 1 \end{bmatrix}_{(\frac{1}{c},1)} = \begin{bmatrix} -c & -2 \\ c & 1 \end{bmatrix}$$

The eigenvalues are therefore given by

$$det\left( \begin{bmatrix} -c & -2 \\ c & 1 \end{bmatrix} - \lambda \mathbb{I} \right) = \begin{vmatrix} -c-\lambda & -2 \\ c & 1-\lambda \end{vmatrix} = \lambda^2 + (c-1)\lambda + c = 0$$

$$\lambda = \frac{1 - c \pm \sqrt{c^2 - 6c + 1}}{2}$$

If $c = 1.1$, the equilibrium point is $(S, P) = (0.91, 1)$, and the eigenvalues are

$$\lambda = -0.05 \pm 1.05\, \boldsymbol{i}$$

Therefore, the equilibrium point is a stable spiral.

If $c = 0.9$, the equilibrium point is $(S, P) = (1.11, 1)$, and the eigenvalues are

$$\lambda = 0.05 \pm 0.95\, \boldsymbol{i}$$

Therefore, the equilibrium point is an unstable spiral.

If $c = 1$, the equilibrium point is $(S, P) = (1, 1)$, and the eigenvalues at the mathematical bifurcation point are

$$\lambda = \pm\, \boldsymbol{i}$$

Therefore, at the bifurcation point, the equilibrium point is a center (Figure 7.41).

In summary, we can now say that the cause of oscillations in this model is a decrease in the reaction rate governed by the controller PFK, which is the $c$ parameter in the $cSP^2$ term.

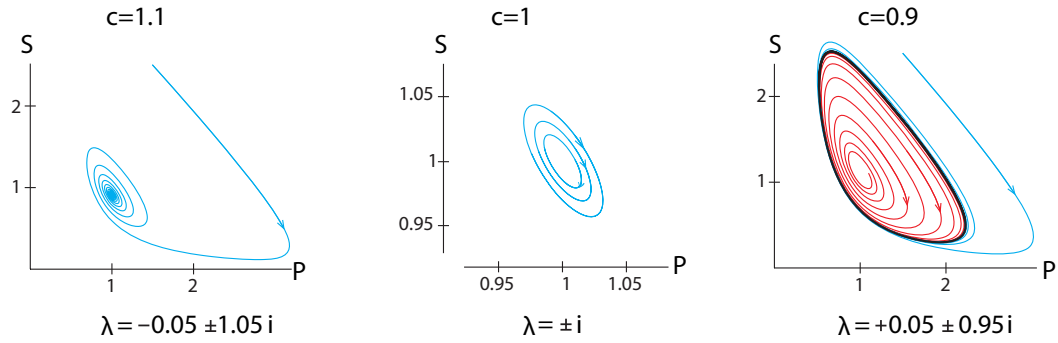



Figure 7.41: In the glycolysis model, decreasing the parameter $c$ past $c = 1$ creates a Hopf bifurcation.

**Exercise 7.6.2** Let $c = 1$ and calculate the value of $v_0$ at which the bifurcation occurs. You can use Sage to help with the algebra. *(Hint: What is $\lambda$ at the bifurcation point?)*

### Example: Oscillatory Gene Expression

As a final example of a Hopf bifurcation, let's consider the gene control oscillator we saw in Chapter 4. The genetic oscillator model consisted of a transcriptional factor $A$ and a transcriptional repressor $R$. The model by Smolen et al. was

$$A' = \frac{kA^2}{A^2 + 10(1 + \frac{R}{0.2})} - A + 0.4$$
$$R' = \frac{0.3A^2}{A^2 + 10(1 + \frac{R}{0.2})} - 0.2R$$

We will use $k$ as our control parameter. The Jacobian matrix can be expressed in terms of $A$, $R$, and $k$:

$$M = \begin{bmatrix} -2kA^3b^2 + 2kAb - 1 & -50A^2kb^2 \\ 0.6Ab - 0.6A^3b^2 & -15A^2b^2 - 0.2 \end{bmatrix} \quad \text{where } b = \frac{1}{A^2 + 10(1 + \frac{R}{0.5})}$$

Because of the complexity of this model, the only way to study the system is by plugging different $k$ values into the system and calculating the corresponding equilibrium points and the Jacobian matrix around that equilibrium point to determine its stability.

First of all, let's find the equilibrium points when $k = 9.5$. Solving $A' = R' = 0$, we get

$$(A, R)\big|_{k=9.5} = (1, 0.1)$$

Plugging in the $k$ value as well as the equilibrium point, we get the Jacobian matrix

$$M\big|_{k=9.5} = \begin{bmatrix} 0.14 & -1.9 \\ 0.036 & -0.26 \end{bmatrix}$$

The corresponding eigenvalues are solutions to

$$det\left(M\big|_{k=9.5} - \lambda\mathbb{I}\right) = 0 \Longrightarrow \lambda = -0.6 \pm 0.17\,\boldsymbol{i}$$

These are complex conjugate eigenvalues with negative real part. Therefore, this equilibrium point is a stable spiral (Figure 7.42).

Now let's consider the case $k = 10.5$. The equilibrium points can be found by setting $A' = R' = 0$. We get

$$(A, R)\big|_{k=10.5} = (2.5, 0.3)$$

Similarly, plugging in the $k$ value as well as the equilibrium point, we get the Jacobian matrix

$$M\big|_{k=10.5} = \begin{bmatrix} 0.34 & -3.4 \\ 0.038 & -0.3 \end{bmatrix}$$

And the corresponding eigenvalues are

$$det\left(M\big|_{k=10.5} - \lambda\mathbb{I}\right) = 0 \Longrightarrow \lambda = +0.024 \pm 0.16\,\boldsymbol{i}$$

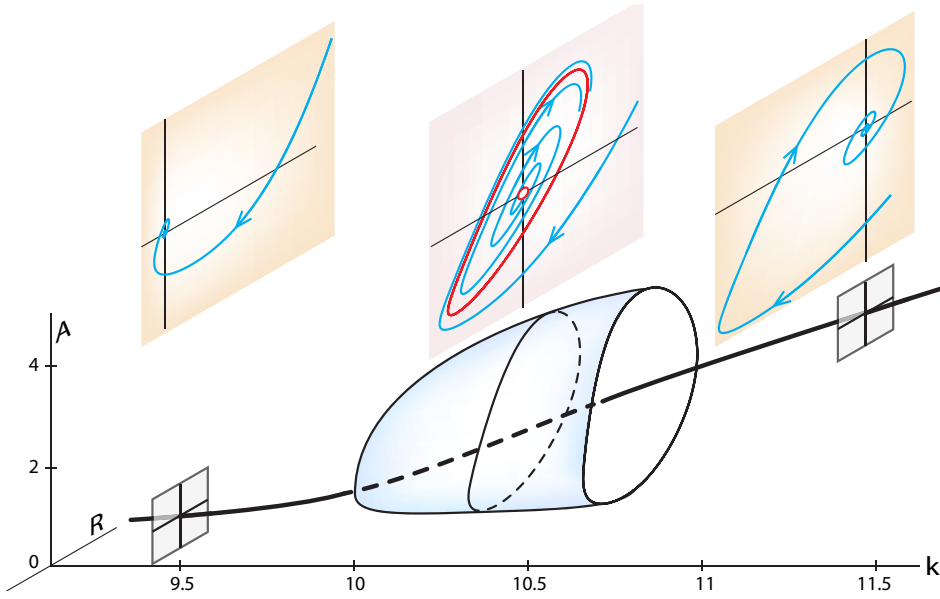



Figure 7.42: A 3D Hopf bifurcation diagram for the gene expression model.

which are complex conjugate eigenvalues with positive real part. Therefore, this equilibrium point when $k = 10.5$ is an unstable spiral. And by the Hopf bifurcation theorem, there is a stable limit cycle attractor surrounding the equilibrium point (Figure 7.42).

Lastly, we are going to consider the case $k = 11.5$. The equilibrium points are

$$(A, R)\big|_{k=11.5} = (4.5, 0.53)$$

As before, by plugging in the $k$ value as well as the equilibrium point, we get the Jacobian matrix

$$M\big|_{k=11.5} = \begin{bmatrix} 0.18 & -3.6 \\ 0.03 & -0.3 \end{bmatrix}$$

And the corresponding eigenvalues are

$$det\left(M\big|_{k=11.5} - \lambda\mathbb{I}\right) = 0 \Longrightarrow \lambda = -0.06 \pm 0.23\,\boldsymbol{i}$$

which are complex conjugate eigenvalues with negative real part. Therefore, this equilibrium point when $k = 11.5$ is a stable spiral (Figure 7.42).

By plugging in many $k$ values, making the same calculations of equilibrium points and stability analysis, and assembling them in order of $k$ value, we can get a *bifurcation diagram* for this model, as shown in Figure 7.42, lower panel.

**Exercise 7.6.3** Even if we can't compute the parameter value at which a Hopf bifurcation takes place, we can use SageMath to approximate it as closely as we want. Outline a procedure for doing so. (You don't have to code anything; just explain what the code would have to do.)

### A Technical Note on Hopf Bifurcation

We have characterized the Hopf bifurcation in two ways:

(1) A Hopf bifurcation is the birth of a stable oscillation from a stable equilibrium point as a parameter passes a critical point.

(2) A Hopf bifurcation occurs when a pair of complex conjugate eigenvalues has its real part pass from negative to positive.

These are, of course, deeply related. However, note that the premise of the theorem is that a pair of complex conjugate eigenvalues has its real part go from negative to positive. Based on our knowledge of eigenvalues, we can then easily say that the motion before the bifurcation will be a stable spiral (negative real part) changing into an unstable spiral (positive real part), while the critical value is a center (zero real part).

However, the conclusion of the Hopf bifurcation theorem tells us much more than that. It guarantees that there is a closed orbit that persists when the parameter is past the critical point, and it also guarantees that under minimal conditions, that closed orbit is an attractor. The math here is deep, and the courageous reader is pointed to technical treatments of Hopf bifurcation theory (Marsden and McCracken is the classic source).

#### Further Exercises 7.6

1. Let's look at a different parameterization of the Higgins–Selkov model,

$$S' = v_0 - 0.23SP^2$$
$$P' = 0.23SP^2 - 0.4P$$

   a) Regardless of the value of $V_0$, this system has one equilibrium point. Find its coordinates *in terms of* $v_0$.

   b) Find the Jacobian matrix of this system at the equilibrium point. Again, this will have to be in terms of $v_0$.

   c) In reality, the value of $v_0$ can vary from around 0.48 to 0.60. For some of these values of $v_0$, the system will exhibit oscillations (there will be a limit cycle attractor). At what exact value of $v_0$ does the Hopf bifurcation occur?

2. Recall the Holling–Tanner predator–prey model:

$$N' = r_1 N\left(1 - \frac{N}{3000}\right) - \frac{300N}{1000 + N}P$$
$$P' = 0.03P\left(1 - \frac{150P}{N}\right)$$

   a) Suppose first that $r_1$ (the natural growth rate of the prey species in the absence of predators) is 0.4. This system has an equilibrium point at about $(226.8, 1.512)$. (This is the only equilibrium point at which both populations are positive.) What type of equilibrium point is it?

b) Now suppose that due to some external factor, $r_1$ drops to 0.2. With these parameters, the equilibrium point is at about $(106.7, 0.712)$. Now what kind of equilibrium point is it?

c) Find the exact value of $r_1$ where the Hopf bifurcation occurred. *(Hint: For a $2 \times 2$ matrix $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$, if the eigenvalues are complex, then their real part is just $\frac{a+d}{2}$. (Why is this true?))*

## 7.7 Optimization

There are many occasions in biology in which we are looking for a maximum or a minimum value of some quantity. The process of finding maxima or minima is called *optimization*.

What kinds of quantities might we want to optimize? Here are a few examples.

- A foraging animal is interested in maximizing caloric intake and minimizing energy costs and exposure to predators. It must also optimize the time spent foraging versus time spent in the nest.
- We would expect organisms to evolve to maximize the number of surviving offspring they have. We will study this example, optimal clutch size, a little later. (Of course, animals don't consciously perform calculations, but we expect their behavior to evolve so as to optimize their overall fitness.)
- In ecology, a species might be trying to maximize its use of available resources or to find an optimal strategy against various competitors, predators, and prey.
- In evolutionary biology, theorists have proposed that different combinations of gene expression lead to different traits with varying amounts of "fitness." Evolution is seen as optimizing "fitness" for a given set of genes.
- In physiology, many of the body's processes are optimal solutions. For example, we can breathe very slowly and use little energy, but then we take in little oxygen, or we can breathe very fast and take in a lot of oxygen, but then we have to work very hard to breathe and spend a lot of energy. Physiological breathing rate is the optimum value.

Building mathematical models of reproduction or behavior and then analyzing what an organism should do if it is attempting to optimize a particular quantity can give us insight into the organism's biology. A mismatch between model predictions and the organism's observed behavior can be particularly revealing, since it indicates that something is wrong with our model.

In each of these cases, we are seeking the maximum or minimum values of some function. Let's now discuss how to find these maxima and minima.

### Maxima and Minima in One Dimension

Let's say that our variable is $X$, and the function to be maximized or minimized is $Y = f(X)$. The maximum value of $f(X)$ is the value that is greater than or equal to all other values $f(X)$ in the domain of $X$. This is what is called a *global maximum*. (There may be several $X$ values at which this maximum is reached.)

**Exercise 7.7.1** By the same logic, what is the minimum value of $f(X)$?

That's easy to say, but how do we find those points? The key step in finding the maxima or minima of $f$, and the values of $X$ at which it is reached, is to first find what are called *local maxima* (or *local minima*). A local maximum is an $f$ value that is greater than any other value *in its immediate neighborhood*.

**Exercise 7.7.2** What is a local minimum?

We can find these values using derivatives. To say that a value $f(X_0)$ is greater than any other value in its neighborhood is to say that to the left of $X_0$, the function is increasing, and to the right of $X_0$, the function is decreasing. But that just means that to the left of $X_0$, the derivative of $f$ with respect to $X$ is positive, and to the right of $X_0$, the derivative of $f$ with respect to $X$ is negative.[4] It follows that if the derivative is a continuous function, then its value at $X_0$ must be 0, because a continuous function can pass from positive to negative only by passing through zero (Figure 7.43).

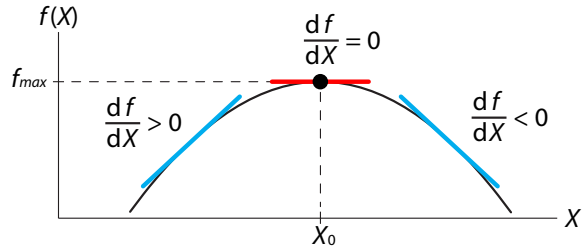



Figure 7.43: If $X_0$ is a local maximum of $f$, then $df/dX$ is positive to the left of $X_0$ and negative to the right.

Similarly, at a local minimum of $f$, the function must be decreasing to the left of $X_0$ and increasing to the right of $X_0$. Again, this implies that if the derivative of $f$ is a continuous function, it must have the value zero at $X_0$ (Figure 7.44).

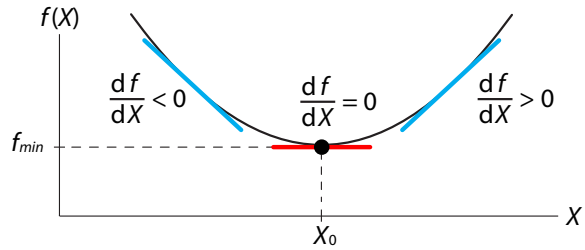



Figure 7.44: If $X_0$ is a local minimum of $f$, then $df/dX$ is negative to the left of $X_0$ and positive to the right.

[4] Technically, we should add the words "on average" after "the function is increasing (decreasing)" and "the derivative of $f$ is positive (negative)." This is to rule out some pathological examples, including functions that oscillate infinitely often in the neighborhood of the critical point.

**Exercise 7.7.3** Restate the conclusions of the previous two paragraphs in geometric terms.

**Exercise 7.7.4** Find the local maxima and minima of the following functions and determine whether they are maxima or minima. *(Hint: Use the definitions.)*

a) $f(X) = X^4 - 2X^2$

b) $f(X) = \dfrac{X^3}{3} - 2X^2 + 3X + 2$

c) $f(X) = 2X^3 - 9X^2 - 24X - 12$

There is a mathematical theorem that sums up all the possible ways for a local maximum or minimum to occur at $X_0$. First, we have to rule out the possibility that $X_0$ is an endpoint of the domain. If $X_0$ is an endpoint, $f(X_0)$ can be a local maximum or minimum even when the derivative of $f$ at that point does not equal zero (Figure 7.45).

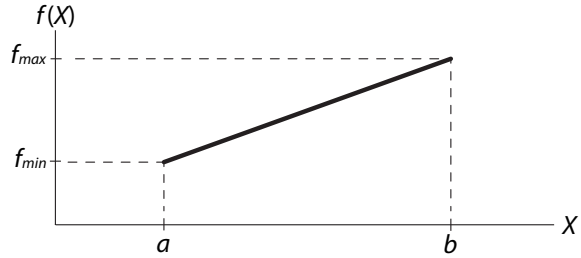



Figure 7.45: A function can have local minima and maxima at the endpoints of its domain even if $df/dX$ is not zero there.

The theorem, which is due to the seventeenth-century French mathematician Fermat, says that if $X_0$ is not an endpoint of the domain and $f(X_0)$ is a local maximum or minimum, then either

$$\left.\frac{df}{dX}\right|_{X_0} = 0$$

or $f$ is not differentiable at $X_0$.

The second clause has to be there because of functions like the absolute value function

$$f(X) = |X|$$

which has an obvious minimum at $X = 0$, although the derivative there is not equal to zero. Indeed, it's undefined (Figure 7.46).

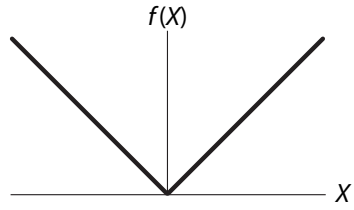



Figure 7.46: A function, such as $f(X) = |X|$, can have a local minimum or maximum at a point where it is not differentiable.

We can now make a definition. We will say that $f$ has a *critical point* at $X_0$ if $\frac{df}{dX}|_{X_0} = 0$ or is undefined. Now suppose $f$ has a critical point at $X_0$. How can we tell whether this is a local maximum, a local minimum, or neither?

Of course, we could just graph the function and look at the graph. This is easy with one variable, more difficult with two, and impossible with three or more variables. So we want to develop a method for classifying critical points that carries over to higher dimensions.

At a local maximum, the function changes from increasing to decreasing. The derivative of the function was positive and is now negative, and therefore the derivative has been decreasing. In other words, the derivative of the derivative, that is, the second derivative, must be negative (Figure 7.47). We write this as

$$\textbf{local maximum} \qquad \frac{d}{dX}\left(\frac{df}{dX}\right) = \frac{d^2f}{dX^2} < 0$$

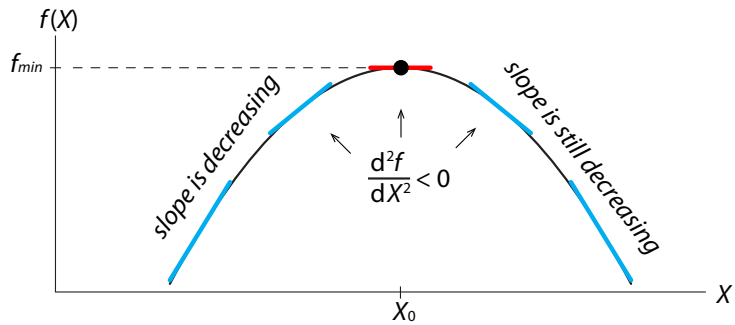



Figure 7.47: To the left of a local maximum of $f$, the slope of $f$ is decreasing. The slope continues to decrease (becomes more negative) to the right of the local maximum.

Similarly, let's look at a local minimum (Figure 7.48). To the left of the local minimum, the slope (first derivative) of $f$ is becoming less and less negative, that is, it is increasing. And to the right of the local minimum, the slope continues to increase, now into positive values. So the second derivative is positive everywhere at and around this minimum. We write this as

$$\textbf{local minimum} \qquad \frac{d}{dX}\left(\frac{df}{dX}\right) = \frac{d^2f}{dX^2} > 0$$

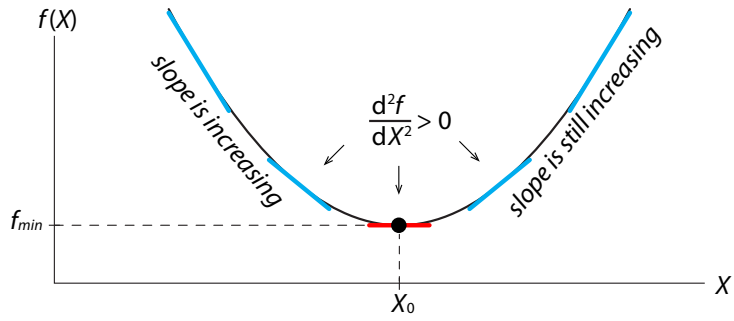



Figure 7.48: To the left of a local minimum of $f$, the slope of $f$ is increasing (becoming less negative). The slope continues to increase to the right of the local maximum.

If the function $f$ has a critical point at $X_0$ and the second derivative of $f$ is less than zero, then the critical point is a maximum.

Similarly, if the function $f$ has a critical point at $X_0$ and the second derivative of $f$ is greater than zero, then the critical point is a minimum.

$$\left.\begin{aligned}\left.\frac{df}{dX}\right|_{X_0} &= 0\\ \left.\frac{d^2f}{dX^2}\right|_{X_0} &< 0\end{aligned}\right\} \Longrightarrow \text{local maximum} \qquad \left.\begin{aligned}\left.\frac{df}{dX}\right|_{X_0} &= 0\\ \left.\frac{d^2f}{dX^2}\right|_{X_0} &> 0\end{aligned}\right\} \Longrightarrow \text{local minimum}$$

In the very special case in which the second derivative is equal to zero, the test is inconclusive. The critical point may be a maximum or a minimum, or it may be neither, such as an *inflection point* (Figure 7.49).

**Exercise 7.7.5** Consider the function $f(X) = X^4$. What is the character of the critical point at $X = 0$?

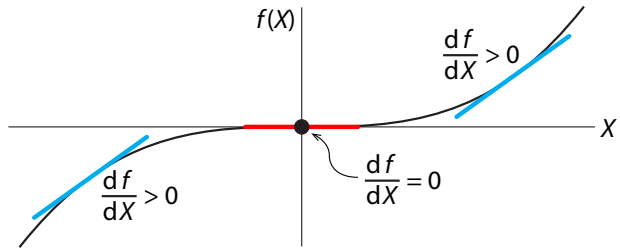



Figure 7.49: The function $f(X) = X^3$. There is an inflection point at $X = 0$.

As an example, let's look at the growth of the population in the logistic model $X' = rX(1-\frac{X}{K})$.

The growth rate starts out slow, then increases, then decreases again as the population approaches the carrying capacity $K$. At what point is the growth rate $X'$ at its maximum?

This question is asking for the maximum value of the function

$$f(X) = rX(1 - \frac{X}{K})$$

Let's find it by differentiating

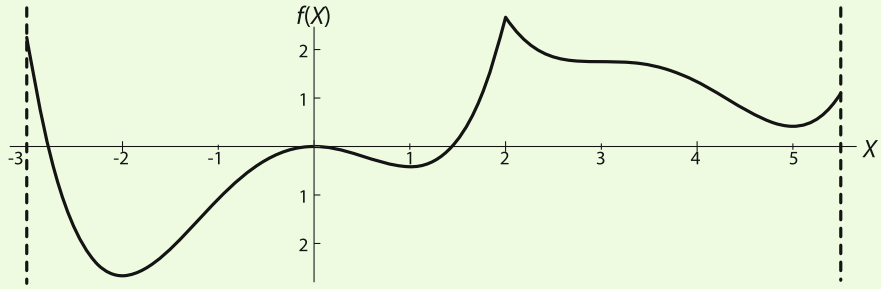
$$\frac{df}{dX} = \frac{df(X)}{dX} = r - \frac{2r}{K}X$$

This is the equation for a straight line with slope $-\frac{2r}{K}$ and $Y$-intercept $r$. Thus it is a perfectly well defined function, and there are no undefined points for $\frac{df}{dX}$.

Next, let's ask when $\frac{df}{dX} = 0$, and the answer is exactly once, when $X = \frac{K}{2}$. Therefore the function has either a unique maximum or a unique minimum at $X = \frac{K}{2}$. We find out which by looking at the second derivative, which is $-\frac{2r}{K}$ and is therefore always negative. Therefore, the point $X = \frac{K}{2}$ defines a maximum of the growth rate. If we plug $X = \frac{K}{2}$ into the function $f(X)$, we get the value $\frac{rK}{4}$, which is the maximum of the growth rate.

This calculation reveals an interesting feature of the logistic model: the maximum growth rate depends on the carrying capacity, a fact that is not obvious.

**Exercise 7.7.6** Consider the function whose graph is shown below:




a) Visually identify all critical points in this graph, identifying each as a maximum, a minimum, or neither. For each critical point, say why this point is a maximum, a minimum, or neither.

b) The function that has this graph is

$$f(X) = \begin{cases} \frac{1}{4}X^4 + \frac{1}{3}X^3 - X^2 & \text{if } -3 \le X \le 2 \\ \frac{1}{4}(X-3)^4 - \frac{2}{3}(X-3)^3 + 1.75 & \text{if } 2 < X \le 5.5 \end{cases}$$

Find the critical points of this function. Then use the second derivative $\frac{d^2f}{dX^2}$ to determine whether they are local maxima, minima, or neither.

**Exercise 7.7.7** Use second derivatives to find the local minima and maxima of the following functions:

a) $f(X) = X^3 - 3X^2 - 9X - 2$

b) $f(X) = 4X^4 - 5X^3 - 36X^2 - 60$

c) $f(X) = (X+2)^2(X-1)^2$

### Optimal Clutch Size

We expect organisms to evolve to maximize their number of surviving offspring. However, different species have vastly different numbers of young. Why does this happen? In birds, the question of optimal clutch size—the number of eggs a bird lays in its nest—has been studied particularly intensively. The contributors to a bird's annual breeding success can be expressed in the following word equation:

$$\begin{array}{c}\text{surviving}\\ \text{offspring}\\ \text{per year}\end{array} = \begin{array}{c}\text{nests}\\ \text{per year}\end{array} \times \begin{array}{c}\text{offspring}\\ \text{per nest}\end{array} \times \begin{array}{c}\text{probability of each}\\ \text{offspring surviving}\end{array}$$

If a bird lays only one nest of eggs per year, we can focus on the other two terms in the equation.

It makes sense that the probability of a baby bird surviving decreases with the number of young in its clutch. More young means more mouths to feed. This not only raises the possibility of

starvation but forces parents to spend more time away from the nest, increasing the chances that either the nest or a parent will be attacked by a predator. The optimal clutch size predicted by life history theory, however, depends on the precise relationship between clutch size and offspring survival.

For example, suppose that offspring survivorship, $S$, for a particular bird species decreases with clutch size, $C$, as

$$S = 1 - 0.1C$$

What is the optimal clutch size for this species?

If the bird lays only one clutch of eggs per year, we can express breeding success as the product of clutch size (number of eggs laid) and survivorship (probability of an egg hatching and maturing into an adult bird). Calling this quantity $y(C)$, we write

$$y(C) = CS = C(1 - 0.1C)$$

To find the maximum of this function, we first expand it to obtain

$$y(C) = C - 0.1C^2$$

and then differentiate with respect to $C$. This gives

$$\frac{dy}{dC} = 1 - 0.2C$$

To maximize this function, we set $\frac{dy}{dC}$ equal to zero and solve for $C$:

$$\frac{dy}{dC} = 0 = 1 - 0.2C$$

Therefore,

$$C = 5$$

The optimal clutch size for this species is five offspring.

**Exercise 7.7.8** (From Case.) Offspring survivorship, $S$, for another bird species decreases with clutch size, $C$, as $S = 0.5 - 0.1C$. What is the optimal clutch size for this species? Again, assume that the bird lays one clutch per year, regardless of how many eggs are in the clutch.

**Exercise 7.7.9** Find a symbolic expression for optimal clutch size in a species that has a survivorship–clutch size relationship of the form $S = a - bC$.

### The Lifeguard Problem

A lifeguard at point $A$ sees a swimmer struggling at point $B$ (Figure 7.50). The lifeguard knows not to run straight toward the swimmer and then continue swimming in the same straight line; running on sand is much faster than swimming in water. Therefore, in order to save time, it's better to spend more time on the sand and less time in the water. What path would get the lifeguard to the swimmer in the shortest possible time?

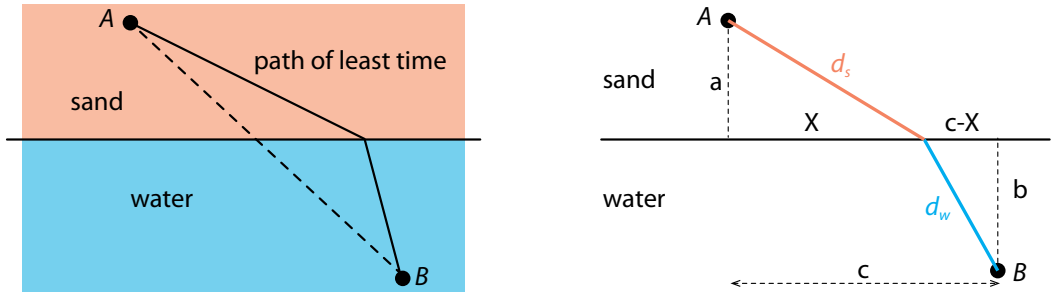



Figure 7.50: The lifeguard problem. The lifeguard runs a distance $d_s$ and then swims a distance $d_w$. We want to know what combination of $d_s$ and $d_w$ gets the lifeguard at $A$ to a struggling swimmer $B$ the fastest.

The lifeguard can run on sand at a speed $v_s$ and can swim in water at a speed $v_w$. Assume that we know how far the lifeguard is from the water ($= a$), how far the swimmer is from the shore ($= b$), and how far down the shore the swimmer is from the lifeguard ($= c$). We will let $X$ be the distance down the shoreline at which the lifeguard enters the water, while $d_s$ and $d_w$ are the distances covered by the lifeguard on the sand and in the water, respectively. We want to find the value of $X$ that minimizes the total time.

The total time is then the sum of the running time plus the swimming time, which in each case is the distance divided by the corresponding running speed:

$$\text{total time} = \frac{\text{distance covered on sand}}{\text{running speed on sand}} + \frac{\text{distance covered in water}}{\text{swimming speed in water}} = \frac{d_s}{v_s} + \frac{d_w}{v_w}$$

We can express $d_s$ and $d_w$ in terms of $X$ using the Pythagorean theorem:

$$d_s = \sqrt{a^2 + X^2} \qquad d_w = \sqrt{b^2 + (c - X)^2}$$

So the expression for the total time as a function of $X$ is

$$t_{\text{total}} = \frac{d_s}{v_s} + \frac{d_w}{v_w} = \frac{\sqrt{a^2 + X^2}}{v_s} + \frac{\sqrt{b^2 + (c - X)^2}}{v_w}$$

To find the entry point $X$ that gives the minimum value of $t_{\text{total}}$, we need to differentiate $t_{total}$ with respect to $X$, set the resulting expression equal to zero, and solve for $X$. But "first derivative = 0" guarantees only a critical point, not necessarily a minimum. To guarantee that a critical point is a minimum, we would need to evaluate the second derivative (see page 418).

If we try to solve this symbolically by hand, or even using SageMath or another computer algebra program, the result is a large, unpleasant fourth-order polynomial with many subterms. Much better is to assume particular values for $a$, $b$, $c$, $v_s$, and $v_w$; then the process is straightforward and can be solved numerically.

Let's say $a = 20\,\text{m}$, $b = 50\,\text{m}$, $c = 100\,\text{m}$, $v_s = 6\frac{\text{m}}{\text{sec}}$, and $v_w = 3\frac{\text{m}}{\text{sec}}$. Then let's plot $t_{total}$ as a function of $X$ (Figure 7.51):

$$t_{\text{total}} = \frac{\sqrt{X^2 + 400}}{6} + \frac{\sqrt{(100 - X)^2 + 2500}}{3}$$

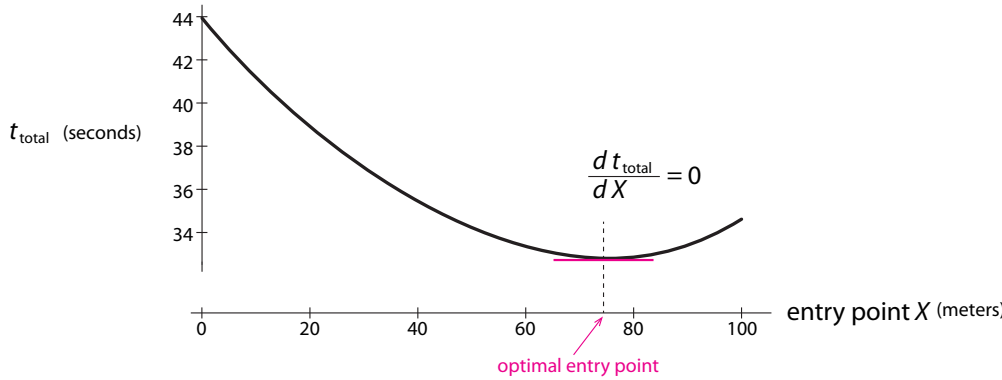



Figure 7.51: Total time needed to reach the swimmer as a function of the entry point $X$.

We see that the function has a unique minimum between 60 m and 80 m. So we don't need to calculate the second derivative; we can see from the graph that the critical point is indeed a minimum. We can find the exact value of the optimal entry point by setting the derivative to zero and solving in SageMath. The SageMath code finds the answer to be $X = 75.38$ m down the shore.

```
>>> a=20  # distance from A to water
>>> b=50  # distance from B to shore
>>> c=100 # distance along the shore between A and B
>>> vs=6  # running speed on sand
>>> vw=3  # swimming speed in water
>>> t_total=1/va*(a^2+x^2)^0.5+1/vw *((c-x)^2+b^2)^0.5 # total time consumed
>>> t_dev=t_total.derivative(x) # calculate the first derivative of t_total
>>> find_root(t_dev, 0, c)  # find the solution x that satisfied t_dev = 0

    SageMath output:
    75.38
```

## Optimization in $n$ Dimensions

We have taken care of the case that $f$ is a function of a single variable $X$.

The much more interesting case occurs when $f$ is a function of several variables, and we want to optimize $f$ over *all* the variables.

Let's consider the 2D case in which $f$ is a function of two variables,

$$Z = f(X, Y)$$

Now we can use our new toolbox of partial derivatives to optimize these functions.[5]

[5] In $n$ dimensions, just as in 1D, optima can occur at domain boundaries and at points where the derivative is undefined. We are not considering those cases here, focusing on the third category of optima, which are places where the derivative equals zero. The value-neutral mathematical term for "optima" is "extrema".

As we saw, a function $f(X,Y)$ can be interpreted as a surface over $X$-$Y$ space whose height at every point $(X_0, Y_0)$ is $Z_0 = f(X_0, Y_0)$. For example,

$$Z = f(X,Y) = e^{-X^2-Y^2}$$

is graphed here (Figure 7.52).

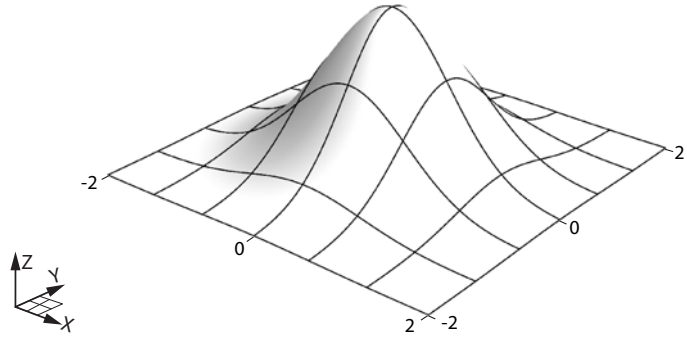



Figure 7.52: A function $Z = f(X,Y)$ gives rise to a 2D surface of $Z$ over the $X$-$Y$ plane.

How do we find optima in 2D? We said that in 1D, an optimum occurs when the tangent line to the graph is flat, that is,

$$\textbf{1D optimum} \quad \Longleftrightarrow \quad \frac{df}{dX} = 0$$

The generalization of Fermat's theorem to 2D is then as follows: a function $Z = f(X,Y)$ has an optimum if and only if the tangent plane to the function is flat, that is,

$$\textbf{2D optimum} \quad \Longleftrightarrow \quad \frac{\partial f}{\partial X} = \frac{\partial f}{\partial Y} = 0$$

This function has an obvious maximum at $(0,0)$. And note that the tangent plane to the surface is indeed flat at that point (Figure 7.53).

The slopes of the tangent plane are the two partial derivatives $\frac{\partial f}{dX}$ and $\frac{\partial f}{dY}$ (Figure 7.54).

It now remains only to calculate these points. The function generating the surface is

$$Z = f(X,Y) = e^{-X^2-Y^2}$$

so the derivative of $f$ with respect to $X$ is

$$\frac{\partial f}{\partial X} = -2Xe^{-X^2-Y^2}$$

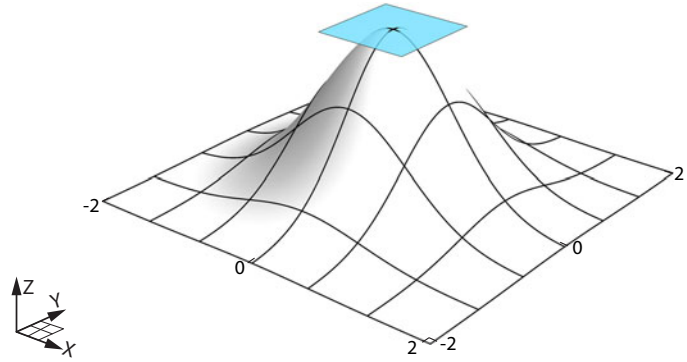



Figure 7.53: At a local maximum of $f$, the tangent plane (blue) to $f(X,Y)$ is horizontal.

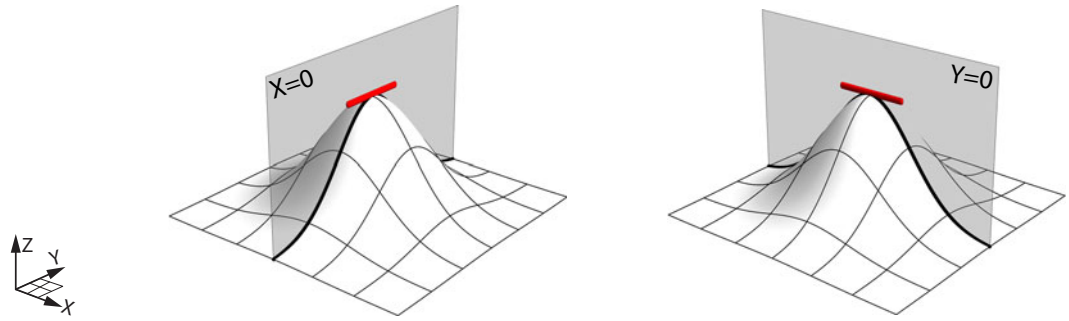



Figure 7.54: At a local maximum of $f$, both partial derivatives of $f$ (the slope of the red lines) are zero.

Setting it to zero gives

$$-2Xe^{-X^2-Y^2} = 0$$

But $e^{-X^2-Y^2}$ can never equal zero, so the only way that this expression can be zero is that

$$X = 0$$

Similarly, the derivative of $f$ with respect to $Y$ is

$$\frac{\partial f}{\partial Y} = -2Ye^{-X^2-Y^2}$$

Setting it to zero gives

$$-2Ye^{-X^2-Y^2} = 0 \quad \Longrightarrow \quad Y = 0$$

**Exercise 7.7.10** Find the critical points of the following functions:

a) $f(X,Y) = X^2 + Y^3 - 6Y$

b) $f(X,Y) = 2X^3 - 3Y^2 + XY$

c) $f(X,Y) = X^2 + 3X - 2Y^2 + 4Y$

So we have verified that $(X,Y) = (0,0)$ is a critical point. But what kind of critical point is it? We might think that it is a maximum or minimum. But in 2D and higher dimensions, there is a third possibility.

Look at the surface generated by the function (Figure 7.55)

$$Z = f(X,Y) = 0.5(X^2 - Y^2)$$

It resembles a saddle, and indeed, the point in the center is called a *saddle point*. Note that at that point, both derivatives are zero, $\frac{df}{dX} = 0$ and $\frac{df}{dX} = 0$, but the point is a maximum in $Y$ and a minimum in $X$. So this point is not an optimum in the two variables.

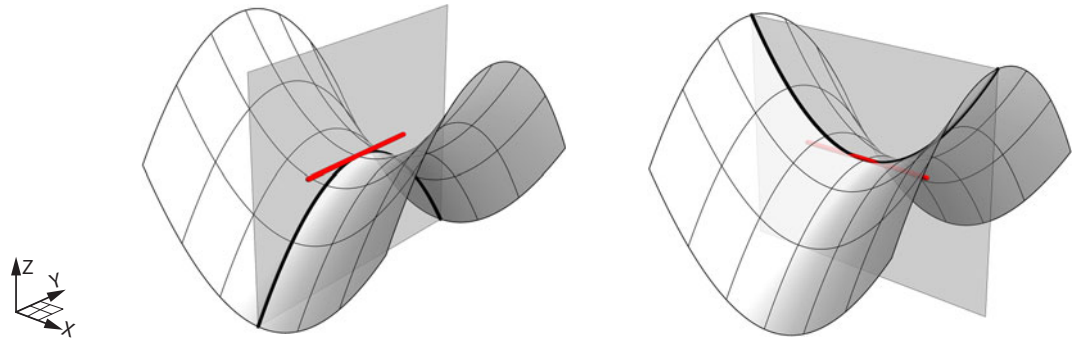



Figure 7.55: At a saddle point of $f$, both partial derivatives of $f$ (the slope of the red lines) are zero, but the point is not a local optimum (maximum or minimum).

**Exercise 7.7.11** By calculating partial derivatives, verify that $(X, Y) = (0, 0)$ is a critical point of the function $Z = f(X, Y) = 0.5(X^2 - Y^2)$.

So how do we classify critical points as maxima, minima, or saddle points? We will use the deep relationship between critical points of functions and equilibrium points of differential equations.

Given a function $Z = f(X, Y)$, we can define a new vector field on $(X, Y)$ space by

$$X' = \frac{dX}{dt} = \frac{\partial f}{\partial X} \qquad \text{and} \qquad Y' = \frac{dY}{dt} = \frac{\partial f}{\partial Y}$$

(Recall that $X'$ is the change of $X$ with respect to time, $\frac{dX}{dt}$.) This new vector field, derived from the function $Z = f(X, Y)$, is called the *gradient vector field* of $f$, called "*grad f*" and often written as $\nabla f$.

**Exercise 7.7.12** Compute $\nabla f$ for the functions in Exercise 7.7.10.

What are the equilibrium points of this vector field? By definition, they are points where $X' = 0$ and $Y' = 0$, that is, $\frac{\partial f}{\partial X} = 0$ and $\frac{\partial f}{\partial Y} = 0$.

But we just said that a critical point of the function $f$ is a point where $\frac{\partial f}{\partial X} = 0$ and $\frac{\partial f}{\partial Y} = 0$. *Therefore, the critical points of $f$ are exactly the equilibrium points of the vector field $\nabla f$.*

**Exercise 7.7.13** Verify that at the critical points you found in Exercise 7.7.10, $\nabla f = 0$.

If $Z = f(X, Y)$ is a height function, we can define the gradient vector field $\nabla f$ as

$$X' = \frac{\partial f}{\partial X}$$
$$Y' = \frac{\partial f}{\partial Y}$$

Critical points of $f$ (maxima, minima, saddles) exactly correspond to equilibrium points (stable, purely unstable, saddle) of the gradient vector field $\nabla f$.

We will now make the key connection that will enable us to identify critical points as maxima, minima, or saddles.

First, let's consider three simple height functions. We will plot the function $f$ and project it down onto the $X$ and $Y$ axes, where we have calculated and plotted the vector field $\nabla f$. The first example is a hill (Figure 7.56, left). The function is

$$Z = f(X, Y) = 5 - \frac{X^2}{2} - \frac{Y^2}{4}$$

The vector field $\nabla f$ is then

$$X' = \frac{\partial f}{\partial X} \qquad Y' = \frac{\partial f}{\partial Y}$$

So

$$X' = -X$$
$$Y' = -0.5Y$$

This is obviously a linear vector field that has a stable equilibrium point at $(0, 0)$.

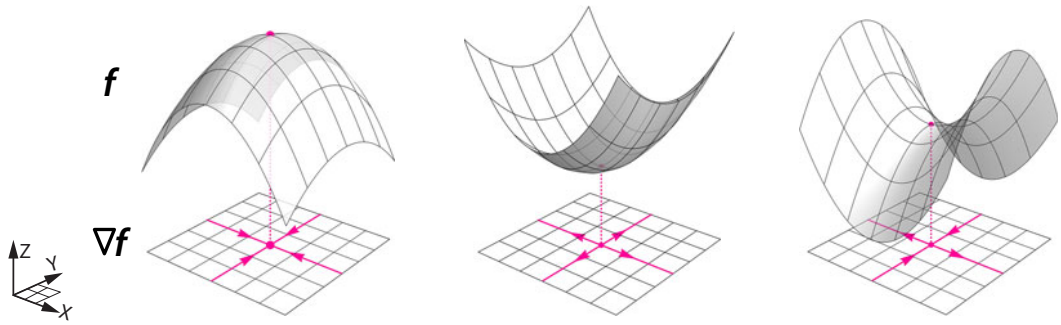



Figure 7.56: At a local maximum of $f$, its gradient vector field $\nabla f$ has a stable node. At a local minimum of $f$, $\nabla f$ has an unstable node. At a saddle point of $f$, $\nabla f$ has a saddle point.

The second example is a bowl (Figure 7.56, middle):

$$Z = f(X, Y) = \frac{X^2}{2} + \frac{Y^2}{4}$$

The vector field $\nabla f$ is then

$$X' = \frac{\partial f}{\partial X} = X$$
$$Y' = \frac{\partial f}{\partial Y} = 0.5Y$$

which again is a linear differential equation, with an unstable equilibrium point at $(0, 0)$.

The third example is a saddle (Figure 7.56, right):

$$Z = f(X, Y) = 0.5(X^2 - Y^2)$$

The vector field $\nabla f$ is then

$$X' = \frac{\partial f}{\partial X} = X$$
$$Y' = \frac{\partial f}{\partial Y} = -Y$$

which again is a linear differential equation, this time with a saddle point at $(0, 0)$.

So in this example:

- Maxima of $f$ correspond to stable equilibrium points (stable nodes) of $\nabla f$.
- Minima of $f$ correspond to purely unstable equilibrium points (unstable nodes) of $\nabla f$.
- Saddle points of $f$ correspond to saddle points of $\nabla f$.

This is true in general, due to the definition of the gradient vector field. Since $Z = f(X, Y)$, we know that the change in $Z$ is given by

$$\Delta Z = \frac{\partial f}{\partial X} \cdot \Delta X + \frac{\partial f}{\partial Y} \cdot \Delta Y$$

But from the definition of $\nabla f$, we know that

$$\begin{cases} X' = \dfrac{\partial f}{\partial X} & \implies \quad \dfrac{\Delta X}{\Delta t} = \dfrac{\partial f}{\partial X} \implies \quad \Delta X = \dfrac{\partial f}{\partial X} \cdot \Delta t \\ Y' = \dfrac{\partial f}{\partial Y} & \implies \quad \dfrac{\Delta Y}{\Delta t} = \dfrac{\partial f}{\partial Y} \implies \quad \Delta Y = \dfrac{\partial f}{\partial Y} \cdot \Delta t \end{cases}$$

If we substitute these expressions for $\Delta X$ and $\Delta Y$ in the $\Delta Z$ equation, we get

$$\begin{aligned} \Delta Z &= \frac{\partial f}{\partial X} \cdot \Delta X + \frac{\partial f}{\partial Y} \cdot \Delta Y \\ &= \frac{\partial f}{\partial X} \cdot \frac{\partial f}{\partial X} \cdot \Delta t + \frac{\partial f}{\partial Y} \cdot \frac{\partial f}{\partial Y} \cdot \Delta t \\ &= \left(\frac{\partial f}{\partial X}\right)^2 \cdot \Delta t + \left(\frac{\partial f}{\partial Y}\right)^2 \cdot \Delta t \end{aligned}$$

Since

$$\Delta t > 0, \quad \left(\frac{df}{dX}\right)^2 > 0, \quad \text{and} \quad \left(\frac{df}{dY}\right)^2 > 0$$

the whole $\Delta Z$ expression is positive. Therefore, $Z$ will always increase following the gradient function $\nabla f$.

**Exercise 7.7.14** Work through this reasoning for $f(X, Y) = X^2 + Y^3 - 6Y$.

We can see this in an even simpler way, by realizing that the gradient vector field is

$$X' = \frac{dX}{dt} = \frac{df}{dX}, \quad Y' = \frac{dY}{dt} = \frac{df}{dY}$$

So if $\frac{df}{dX}$ is positive, this means that $f$ is increasing with respect to $X$. But then the vector field $X' = \frac{dX}{dt} = \frac{df}{dX}$ is positive, which means that $X$ is increasing with respect to time. Since $\Delta Z = \frac{\partial f}{\partial X} \cdot \Delta X + \frac{\partial f}{\partial Y} \cdot \Delta Y$, this increase of $X$ in time will increase $Z$, precisely because $\frac{df}{dX}$ is positive.

And if $\frac{df}{dX}$ is negative, then the vector field $X' = \frac{df}{dX}$ is negative, which means that $X$ will decrease with respect to time. This decrease of $X$, which is reflected in a negative value of $\Delta X$,

will also cause an increase in $Z$, since $\frac{df}{dX}$ and $\Delta X$ are both negative! So $Z$ will always increase. A similar argument for $Y$ shows that when moving under the gradient vector field, the quantity $Z$ will always increase.

The fact that $Z = f(X, Y)$ is always increasing when it follows the gradient vector field $\nabla f$ explains why maxima of $f$ correspond to stable equilibrium points of $\nabla f$. If $Z_0$ is a local maximum, then $Z$ cannot increase any further, so the process of increasing $Z$ must have come to a stable equilibrium point.

**Exercise 7.7.15** Using SageMath, plot the vector fields $\nabla f$ for the functions in Exercise 7.7.10. What do the equilibria look like?

In our earlier examples, we made our task easy: when we looked at the vector field $\nabla f$, it was obviously a very simple linear vector field, and so the stability of the equilibrium point was obvious by inspection.

In the general case, we will have to use our theory of the stability of nonlinear vector fields. Let's consider a different example:

$$f(X, Y) = X^2 + 2Y^2 - X^2Y$$

Now when we calculate the gradient vector field $\nabla f$, it is

$$\frac{\partial f}{\partial X} = 2X - 2XY$$
$$\frac{\partial f}{\partial Y} = 4Y - X^2$$

giving us the vector field

$$X' = 2X - 2XY$$
$$Y' = 4Y - X^2$$

It is far from obvious what the equilibrium points even are, let alone what their stability is. But we can use the method of this chapter to answer these questions.

First, let's find the equilibrium points of the vector field.

Setting $X' = 0$, we get

$$X' = 2X - 2XY = 0$$

which implies

$$X = 0 \quad \text{or} \quad Y = 1$$

Plugging $X = 0$ into the $Y' = 0$ equation, we get

$$Y = 0$$

Plugging $Y = 1$ into the $Y' = 0$ equation, we get

$$X = \pm 2$$

Therefore, there are exactly three equilibrium points in this vector field. They are

$$(X, Y) = (0, 0)$$
$$(X, Y) = (2, 1)$$

$$(X, Y) = (-2, 1)$$

Next, we will determine the stability of the vector field at these equilibrium points by the method of linearization. First, we find the Jacobian matrix

$$M = \begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial Y} \\ \frac{\partial Y'}{\partial X} & \frac{\partial Y'}{\partial Y} \end{bmatrix} = \begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}$$

Then we evaluate the Jacobian matrix at each equilibrium point to give us the linearization at that point, and then we use the method of eigenvalues to determine the stability of the linearization.

Let's do this for the three equilibrium points.

**(X, Y) = (0, 0).** The Jacobian matrix at this point is

$$\begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}_{(0,0)} = \begin{bmatrix} 2 & 0 \\ 0 & 4 \end{bmatrix}$$

This is a diagonal matrix with positive eigenvalues, indicating a purely unstable equilibrium point. Therefore, we conclude that the height function $f$ has a minimum at the point $(0, 0)$.

**(X, Y) = (2, 1).** The Jacobian matrix at this point is

$$M\big|_{(2,1)} = \begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}_{(2,1)} = \begin{bmatrix} 0 & -4 \\ -4 & 4 \end{bmatrix}$$

To calculate its eigenvalues, we solve

$$det\left(M\big|_{(2,1)} - \lambda \mathbb{I}\right) = \begin{vmatrix} 0-\lambda & -4 \\ -4 & 4-\lambda \end{vmatrix} = 0$$

$$\lambda^2 - 4\lambda - 16 = 0$$

$$\lambda = \frac{1 \pm \sqrt{5}}{2}$$

Since there is one positive eigenvalue and one negative one, we conclude that $(2, 1)$ is a saddle point, and the function $f$ has a saddle at this point.

**(X, Y) = (-2, 1).** The Jacobian matrix at this point is

$$M\big|_{(-2,1)} = \begin{bmatrix} 2-2Y & -2X \\ -2X & 4 \end{bmatrix}_{(-2,1)} = \begin{bmatrix} 0 & 4 \\ 4 & 4 \end{bmatrix}$$

and a similar calculation gives us

$$\lambda = \frac{1 \pm \sqrt{5}}{2}$$

So $(-2, 1)$ is also a saddle point equilibrium of the gradient vector field $\nabla f$, and it is a saddle point of the function $f$.

The general idea is that given a height function $f(X, Y)$, we can always define a dynamical system $\nabla f$ on the state space $(X, Y)$. The dynamical system $\nabla f$ defines a process of always increasing the value of $f$, and doing so by finding the steepest path on the hill and following it. If $f$ defines a field of hills and valleys, then $\nabla f$ is the command to climb as rapidly as possible.

We can see this by plotting the contours of $f$ in the $(X, Y)$ plane (Figure 7.57). This is similar to the technique of a contour map, in which lines of constant altitude are drawn on the 2D map surface.

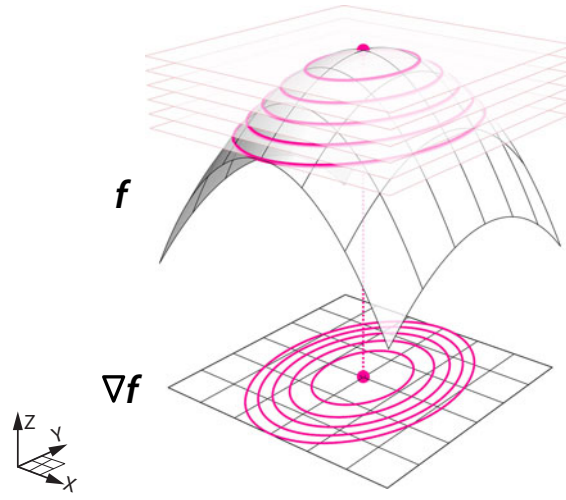



Figure 7.57: Plotting curves along which $f(X,Y)$ has a constant $Z$-value, and projecting these curves down onto the $X$-$Y$ plane, gives the equivalent of a contour map of the gradient vector field $\nabla f$ .

Let's plot a trajectory of the gradient vector field $\nabla f$ in the $(X,Y)$ plane and project this trajectory up onto the surface (Figure 7.58).

There are two features of this vector field:

(1) It is everywhere perpendicular to the contour lines.

(2) The trajectory is the path of steepest ascent, that is, it is the path that maximizes the change in $f$.[6]

(These last two principles are quickly shown using techniques of linear algebra that are outside the scope of this text and are easily found on the Internet.)

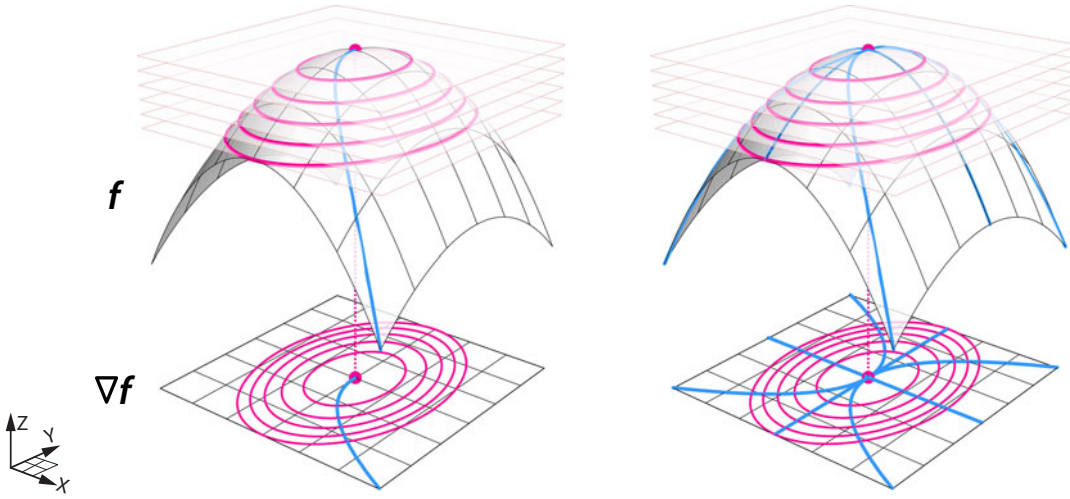



Figure 7.58: When $f$ has a local maximum, trajectories (shown in blue) that follow the gradient vector field $\nabla f$ will climb the hill defined by $f(X,Y)$ as rapidly as possible (the steepest ascent).

[6]For this reason, one of the authors thinks of $\nabla f$ as the rock climber's vector field.

**Exercise 7.7.16** Use this method to classify the critical points of the functions in Exercise 7.7.10 as local maxima, local minima, or saddle points.

To classify the critical points of a height function $Z = f(X, Y)$:
(1) Find the critical points by setting $\frac{\partial f}{\partial X}$ and $\frac{\partial f}{\partial Y}$ equal to zero, and find the points $(X_0, Y_0)$ that satisfy that equation.
(2) Form the gradient vector field $\nabla f$:

$$X' = \frac{\partial f}{\partial X} \quad \text{and} \quad Y' = \frac{\partial f}{\partial Y}$$

(3) Take the Jacobian of $\nabla f$.
(4) Use the method of eigenvalues to determine the stability of each equilibrium point. If the equilibrium point is
- stable, the function has a maximum.
- purely unstable, the function has a minimum.
- a saddle point, the function has a saddle point.

If we write out the Jacobian of $\nabla f$, we see that it takes a particularly simple form. In general, the Jacobian is

$$M = \begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial Y} \\ \frac{\partial Y'}{\partial X} & \frac{\partial Y'}{\partial Y} \end{bmatrix}$$

and here

$$X' = \frac{\partial f}{\partial X} \quad \text{and} \quad Y' = \frac{\partial f}{\partial Y}$$

so

$$M = \begin{bmatrix} \frac{\partial X'}{\partial X} & \frac{\partial X'}{\partial Y} \\ \frac{\partial Y'}{\partial X} & \frac{\partial Y'}{\partial Y} \end{bmatrix} = \begin{bmatrix} \frac{\partial(\frac{\partial f}{\partial X})}{\partial X} & \frac{\partial(\frac{\partial f}{\partial X})}{\partial Y} \\ \frac{\partial(\frac{\partial f}{\partial Y})}{\partial X} & \frac{\partial(\frac{\partial f}{\partial Y})}{\partial Y} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial X^2} & \frac{\partial^2 f}{\partial X \partial Y} \\ \frac{\partial^2 f}{\partial Y \partial X} & \frac{\partial^2 f}{\partial Y^2} \end{bmatrix}$$

The Jacobian of a gradient vector field $\nabla f$ is called the **Hessian** of $f$. It is the matrix of second partial derivatives.

Note the two nondiagonal terms in the Hessian. It is a theorem from multivariable calculus that if the two partial derivatives $\frac{\partial f}{\partial X}$ and $\frac{\partial f}{\partial Y}$ are both continuous, then the mixed partial derivatives are equal to each other:

$$\frac{\partial^2 f}{\partial X \partial Y} = \frac{\partial^2 f}{\partial Y \partial X}$$

Therefore, the Hessian matrix is always symmetric. We can therefore apply a theorem from linear algebra that says that a symmetric matrix can have only real eigenvalues. This has the consequence that there can be no spiraling in a gradient vector field: the state point must head straight upward by the steepest path.

Consequently, we can restate the main conclusion by saying that a critical point of $f$ is a maximum if the Hessian has all negative eigenvalues, is a minimum if the Hessian has all positive eigenvalues, and is a saddle point if eigenvalues are positive and negative.

### Evolution and the "Fitness Landscape"

The metaphor of the hills and valleys of a height function is a powerful one. We think of the gradient vector field $\nabla f$ as ascending to the heights of the hills; we can think of the motion of a helium-filled balloon lying under the surface of $Z = f(X, Y)$ and rising to the local maximum. Similarly, we can visualize the negative gradient, $-\nabla f$, as a solid ball, rolling downhill into the local valley.

This metaphor was very attractive to the evolutionary theorist and genetics pioneer Sewall Wright. In a famous paper of 1932, he proposed that we can imagine a "fitness landscape" (or "evolutionary landscape"), in which all possible combinations of expression levels of gene $X$ and expression levels of gene $Y$ are considered, and the height function $f(X, Y)$ then gives the level of "fitness" or "adaptability" of that combination (Wright 1932).

His image was then that evolution is a process of moving uphill on this evolutionary landscape, up the gradient of fitness (Figure 7.59).

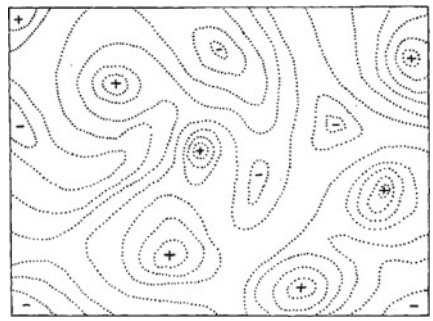


Figure 7.59: Contour lines of hypothetical evolutionary landscape (Wright 1932).

Subsequent work, including work by Wright himself, has raised several criticisms of the concept: the word "landscape" denotes a fixed topography. But the real environment is changing in time, leading to the concept of a "seascape" rather than a "landscape." For example, climate change is certainly a factor that is reshaping the evolutionary landscape.

Mathematical biologists have continued to work on the concept of the evolutionary landscape. (See, for example, the paper "Multiple Fitness Peaks on the Adaptive Landscape Drive Adaptive Radiation in the Wild," by Christopher H. Martin and Peter C. Wainwright (Martin and Wainwright 2013).

### From Local to Global Maxima and Minima

So far, we have restricted our attention to finding *local* maxima and minima. These are the types of points that can be found using derivative-based techniques, which is not surprising, since the derivative is a local concept.

You might object that what we are really interested in are *global* maxima and minima, not local ones. The point is well taken, but the problem is that there are no elegant techniques for finding a global optimum. All you can do is find all the local maxima or minima, including those at the boundaries and cusp points, and then choose the one with the largest (smallest) value (Figure 7.60).

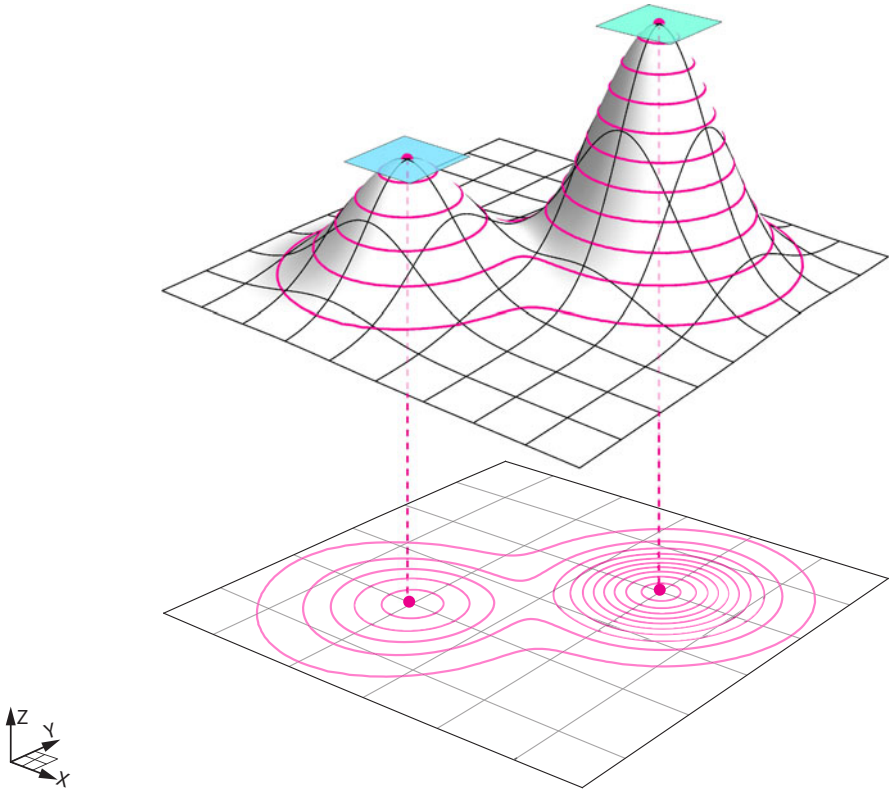



Figure 7.60: Contour lines and local maxima for a hypothetical fitness landscape.

If all we can do is follow the gradient vector field $\nabla f$, we will go to the local maximum. But what if the local maximum is not a global maximum? Then we are "stuck in a local maximum (or minimum)."

There are advanced mathematical techniques for getting out of local maxima and minima. The most popular technique is to add some noise to the system, to shake it up a little. Imagine a ball stuck in a local minimum of a topographic 3D surface lying on a table. If we shake the table a little, the ball will become dislodged from the local minimum and be free to seek other minima. (This technique is called "simulated annealing.")

In evolution, it is not so easy to back out of a local maximum or minimum; it may cause a catastrophic loss of fitness.

Consider the fact that mammalian eyes have a "blind spot," while those of cephalopods do not. Why is this? Because sight evolved a number of times. Both vertebrates and cephalopods have camera-type eyes, but they evolved independently. The cephalopod eye is built like you'd expect—the photoreceptors are in front of the optic nerve. But our eyes are backward—the optic nerve passes in front of the retina, causing a blind spot that the brain has to compensate

for. More seriously, it makes us vulnerable to retinal detachment, which cephalopods don't get. But we're stuck with this backward design because it would be too hard to undo and would have to pass through stages that are worse. Evolution can't go downhill in order to get to a higher peak later.

### Further Exercises 7.7

1. Find and classify the critical points of the following functions:
   a) $f(X) = X^2 + 10$
   b) $f(X) = \dfrac{6X}{X^2 + 36}$
   c) $f(X) = X^2 + \frac{16}{X}$
   d) $f(X) = 3X^4 - 4X^3 - 36X^2 + 60$

2. Bonnacons grow at a rate $g(t) = 8t^3 - 3t^2 - t + 4$, where $t$ is the time since the bonnacon's birth. At what value of $t$ is the bonnacon's growth rate minimized, and what is its value at that minimum?

3. For infants younger than nine months, the relationship between weight $W$ (in pounds) and the rate of growth (in pounds/month) is approximately
$$\frac{dW}{dt} = cW(21 - W)$$
for some constant $c$. At what weight is the infant growing fastest?

4. When a person coughs, their trachea narrows, speeding up air flow and increasing the force on the object that the cough is meant to expel. X-ray studies show that the radius of the trachea, which is circular, contracts to about $\frac{2}{3}$ of normal during a cough. The velocity, $v$, of the airstream is related to the radius, $r$, of the trachea by
$$v(r) = k(r_0 - r)r^2 \qquad \frac{1}{2}r_0 \le r \le r_0$$
where $r_0$ is the normal radius of the trachea and $k$ is a proportionality constant. The restriction on $r$ is due to the stiffening of the trachea as it narrows, which prevents the person from suffocating.
   a) The average radius of a human trachea is about 12.7 mm. Pick a value for $k$ and plot $v(r)$ on the interval $[0, r_0]$. What aspects of the graph remain the same regardless of the value of $k$?
   b) Find the value of $r$ on the interval $[\frac{1}{2}r_0, r_0]$ at which $v(r)$ is maximized. Give an expression for the value of $v$ at this point.

5. Termites live in a colony in which each individual (ignoring the king and queen) develops into one of two highly specialized castes: workers who forage for food and maintain the colony's nest, and soldiers who defend the colony from ants and other predators. Assume that $X$ represents the fraction of termites in a colony that are workers, and the rest $(1 - X)$ are soldiers. While studying a termite colony, you develop a function

describing the growth rate of the colony as a function of $X$:

$$f(X) = \sqrt{X^2 - 2X} - \frac{5}{4}X$$

When the growth rate is maximized, what fraction of the termites will be workers? (Note: Don't just find the critical point(s). Be sure to test whether each one is a maximum or minimum.)

6. Find and classify all the critical points of the following functions:
   a) $f(X, Y) = 10 - X^2 - Y^2$
   b) $f(X, Y) = 12X^2 + Y^3 - 12XY$
   c) $f(X, Y) = X^3 + Y^3 - 3XY + 4$
   d) $f(X, Y) = 3X^2Y + Y^3 - 3X^2 - 3Y^2 + 2$

7. You are studying the effect of two traits on the evolution of sparrows. Let $X$ represent the value of one trait, such as bill width, and let $Y$ represent the level value of the other trait, such as wingspan. You have found that the following function models the fitness of individuals born with any given level of $X$ and $Y$:

$$f(X, Y) = 9X^2 + 6Y^2 - 4X^3 - 2Y^3 - 3X^2Y^2$$

   This function has critical points at $(0, 0)$, $(0, 2)$, $(1, 1)$, and $(1.5, 0)$.
   a) Classify each critical point as a local maximum, local minimum, or saddle point.
   b) At what values of $X$ and $Y$ might you expect distinct species of sparrows to form?

8. Plants need nitrogen ($N$) and phosphorus ($P$) to grow, but both of these nutrients can become toxic at high concentrations. Suppose that the growth rate of a plant is given by

$$g(N, P) = 5 - (N - 3)^2 - (P - 2)^2$$

   Find the optimal nitrogen and phosphorus levels for this plant. Make sure to check that your critical point really is the maximum.