

Chapter 2 of Calculus⁺⁺

Applications of differential calculus with several variables

by

Eric A Carlen
Professor of Mathematics
Georgia Tech

Overview

In this section we put our knowledge of linear and quadratic approximations to work on two important problems: solving nonlinear equations and constrained optimization problems.

Section 1: Iterative solution of nonlinear systems of equations

1.1 Newton's method

Consider the following system of non linear equations:

$$\begin{aligned}x^2 + 2yx &= 4 \\ xy &= 1 .\end{aligned}\tag{1.1}$$

Any system of two equations in two variables can be written in the form

$$\begin{aligned}f(x, y) &= 0 \\ g(x, y) &= 0 .\end{aligned}\tag{1.2}$$

In this case we define $f(x, y) = x^2 + 2xy - 4$ and $g(x, y) = xy - 1$. All you have to do is to take whatever is on the right hand side of each equations, and subtract it off of both sides, leaving zero on the right. Just so we can standardize our methods, we shall always assume our equations have zero on the right hand side. If you run into one that doesn't, you know what to do as your first step: Cancel off the right hand sides.

Next, introduce $F(\mathbf{x}) = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$ we can write this as a single vector equation

$$\mathbf{F}(\mathbf{x}) = 0 .\tag{1.3}$$

In the case of (1.1), we have

$$\mathbf{F}(x, y) = \begin{bmatrix} x^2 + 2yx - 4 \\ xy - 1 \end{bmatrix} .\tag{1.4}$$

In this case, we can solve (1.3) by algebra alone. For this purpose, the original formulation is most convenient. Using the second equation in (1.1) to eliminate y , the first equation becomes $x^2 = 2$. Hence $x = \pm\sqrt{2}$. The second equation says that $\frac{y=1}{x}$ and so we have two solutions

$$(\sqrt{2}, 1/\sqrt{2}) \quad \text{and} \quad (-\sqrt{2}, -1/\sqrt{2}) .$$

In general, it may be quite hard to eliminate either variable, and algebra alone cannot deliver solutions.

There is a way forward: Newton's method is a very effective algorithm for solving such equations. This is a "successive approximations method". It takes a starting guess for the solution \mathbf{x}_0 , and iteratively improves the guess. The iteration scheme produces an *infinite sequence* of approximate solutions $\{\mathbf{x}_n\}$. Under favorable circumstances, this sequence will converge *very rapidly* toward an exact solution. In fact, the number of correct digits x_n and y_n will more or less double double at each step. If you have one digit right at the

outset, you may expect about a million correct digits after 20 iterations – more than you are ever likely to want to keep!

To explain the use of Newton’s method, we have to cover three points:

- (i) How one picks the starting guess \mathbf{x}_0 .
- (ii) How the iterative loop runs; i.e., the rule for determining \mathbf{x}_{n+1} given \mathbf{x}_n .
- (iii) How to break out of the iterative loop – we need a “stopping rule” that ensures us our desired level of accuracy has been achieved when we stop iterating.

We begin by explaining (ii), the nature of the loop. Once we are familiar with it, we can better understand what we have to do to start it and stop it.

The basis of the method is the linear approximation formula for \mathbf{F} at \mathbf{x}_0 :

$$\mathbf{F}(\mathbf{x}) \approx \mathbf{F}(\mathbf{x}_0) + J_{\mathbf{F}}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) . \tag{1.5}$$

Using this, we replace (1.3) with the approximate equation

$$\mathbf{F}(\mathbf{x}_0) + J_{\mathbf{F}}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = 0 . \tag{1.6}$$

Don’t let the notation obscure the simplicity of this: $\mathbf{F}(\mathbf{x}_0)$ is just a constant vector in \mathbb{R}^2 and $J_{\mathbf{F}}(\mathbf{x}_0)$ is just a constant 2×2 matrix. Using the shorter notation $\mathbf{F}(\mathbf{x}_0) = \mathbf{b}$ and $J_{\mathbf{F}}(\mathbf{x}_0) = A$, we can rewrite (1.6) as

$$A(\mathbf{x} - \mathbf{x}_0) = -\mathbf{b} .$$

We know what to do with this! We can solve this by row reduction. In fact, if A is invertible, we have $\mathbf{x} - \mathbf{x}_0 = A^{-1}\mathbf{b}$, or, what is the same thing,

$$\mathbf{x} = \mathbf{x}_0 - A^{-1}\mathbf{b} .$$

Writing this out in the full notation, we have a formula for the solution of (1.6)

$$\mathbf{x} = \mathbf{x}_0 - (J_{\mathbf{F}}(\mathbf{x}_0))^{-1} \mathbf{F}(\mathbf{x}_0) . \tag{1.7}$$

we now define \mathbf{x}_1 to be this solution. To get \mathbf{x}_2 from \mathbf{x}_1 , we do the same thing starting from \mathbf{x}_1 . In general, we define \mathbf{x}_{n+1} to be the solution of

$$\mathbf{F}(\mathbf{x}_n) + J_{\mathbf{F}}(\mathbf{x}_n)(\mathbf{x} - \mathbf{x}_n) = 0 . \tag{1.8}$$

If $J_{\mathbf{F}}(\mathbf{x}_n)$ is invertible, this gives us

$$\mathbf{x}_{n+1} = \mathbf{x}_n - (J_{\mathbf{F}}(\mathbf{x}_n))^{-1} \mathbf{F}(\mathbf{x}_n) . \tag{1.9}$$

Now let’s run through an example.

Example 1 (Using Newton’s iteration) Consider the system of equations $\mathbf{F}(\mathbf{x}) = 0$ where \mathbf{F} is given by (1.4). We will choose a starting point so that at least one of the equations in the system is satisfied,

and the other is not *too* far off. This seems reasonable enough. Notice that with $x = y = 1$, $xy - 1 = 0$, while $x^2 - 2xy - 4 = -1$. Hence with $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ we have

$$\mathbf{F}(\mathbf{x}_0) = \begin{bmatrix} -1 \\ 0 \end{bmatrix} .$$

Now let's write our system in the form $F(x, y) = 0$. We can do this with

$$F(\mathbf{x}) = \begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix} = \begin{bmatrix} x^2 + 2yx - 4 \\ xy - 1 \end{bmatrix} .$$

Computing the Jacobian, we find that

$$J_F(\mathbf{x}) = \begin{bmatrix} 2x + 2y & 2x \\ y & x \end{bmatrix} , \tag{1.10}$$

and hence

$$J_F(\mathbf{x}_0) = \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix} , \tag{1.11}$$

Hence (1.9) is

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ 0 \end{bmatrix} .$$

Since

$$\begin{bmatrix} 4 & 2 \\ 1 & 1 \end{bmatrix}^{-1} = \frac{1}{2} \begin{bmatrix} 1 & -2 \\ -1 & 4 \end{bmatrix} ,$$

we find

$$\mathbf{x}_1 = [3/2, 1/2] .$$

Notice that \mathbf{x}_1 is indeed considerably closer to the exact solution $[\sqrt{2}, 1/\sqrt{2}]$ than \mathbf{x}_0 . Moreover,

$$F(\mathbf{x}_1) = -\frac{1}{4} \begin{bmatrix} 1 \\ 1 \end{bmatrix} .$$

This is a better approximate solution; it is much closer to the actual solution. If you now iterate this further, you will find a sequence of approximate solutions converging to the exact solution $(\sqrt{2}, 1/\sqrt{2})$. You should compute \mathbf{x}_2 and \mathbf{x}_3 and observe the speed of convergence.

1.2 Choosing a starting point for Newton's method

With two variables, we can use what we know about generating plots of implicitly defined curves to locate good starting points. In fact, we can use such plots to determine the number of solutions. To do this, write \mathbf{F} in the form

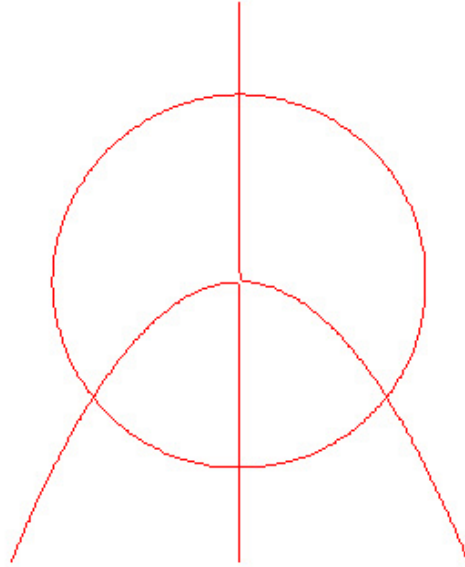
$$\mathbf{F}(x, y) = \begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix} .$$

Each of the equations

$$f(x, y) = 0 \quad \text{and} \quad g(x, y) = 0$$

is an implicit definition of a curve. Points where the two curves intersect are points belonging to the solution set of both equations; i.e., to the solution set of $\mathbf{F}(\mathbf{x}) = 0$.

Example 2 (Using a graph to find a starting point for Newton's iteration) Consider the system $\begin{bmatrix} f(x, y) \\ g(x, y) \end{bmatrix} = 0$ where $f(x, y) = x^3 + xy$, and $g(x, y) = 1 - y^2 - x^2$. This is non linear, but simple enough that we can easily plot the curves. The equation $g(x, y) = 0$ is equivalent to $x^2 + y^2 = 1$, which is the equation for the unit circle. Since $f(x, y) = x(x^2 + y)$, $f(x, y) = 0$ is and only if $x = 0$, which is the equation of the y axis, or $y = -x^2$, which is the equation of a parabola. Here is a graph showing the intersection of the implicitly defined curves:



The axes have been left off since one branch of the second curve is the y axis. Since one curve is the unit circle though, you can easily estimate the coordinates of the intersections anyway. As you see, there are exactly 4 solutions. Two of them are clearly the exact solutions $(0, \pm 1)$. The other two are where the parabola crosses the circle. Carefully measuring on the graph, you could determine (axes would now help) that $y \approx -0.618$ and $x \approx \pm 0.786$. This would give us two good approximate solutions. applying Newton's method, we could improve them to compute as many digits as we desire of the exact solution.

If you have more than two variables, graphs become harder to use. An alternative to drawing the graph is to evaluate $\mathbf{F}(\mathbf{x})$ at all of the points in some grid, in some limited range of the variables. Use whichever grid points give $\mathbf{F}(\mathbf{x}) \approx 0$ as your starting points.

1.3 Geometric interpretation of Newton's method

Newton's method is based on the tangent plane approximation, and so it has a geometric interpretation. This will help us to understand why it works when it does, and how we can reliably stop it.

Here is how this goes for the system

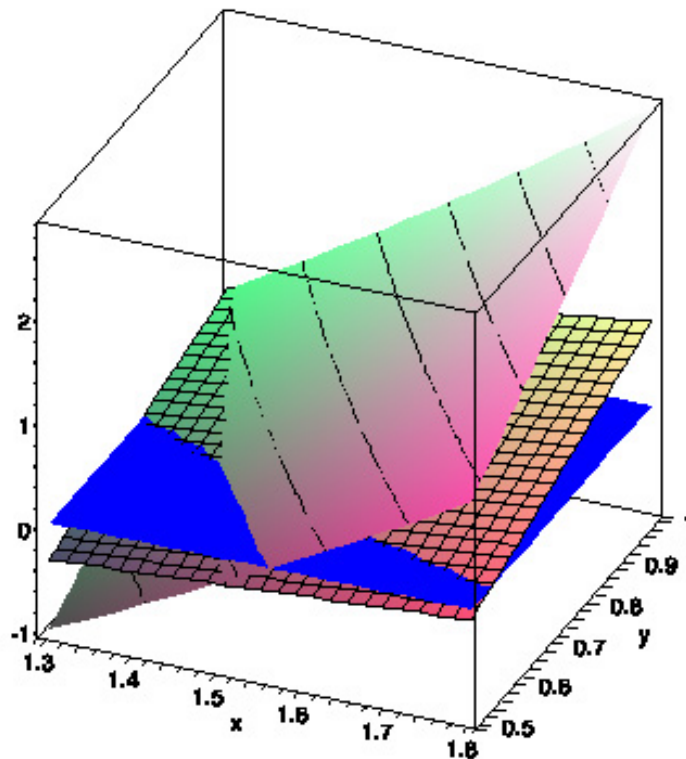
$$\begin{aligned} f(x, y) &= 0 \\ g(x, y) &= 0 . \end{aligned} \tag{1.12}$$

Replace this by the equivalent system

$$\begin{aligned}z &= f(x, y) \\z &= g(x, y) \\z &= 0 .\end{aligned}\tag{1.13}$$

From an algebraic standpoint, we have taken a step backwards – we have gone from two equations in two variables to three equations in three variables. However, (1.13) has an interesting geometric meaning: The graph of $z = f(x, y)$ is a surface in \mathbb{R}^3 , as is the graph of $z = g(x, y)$. The graph of $z = 0$ is just the x, y plane – a third surface. Hence the solution set of (1.13) is given by the intersection of 3 surfaces.

For example, here you a plot of the three surfaces in (1.13) when $f(x, y) = x^2 + 2xy - 4$ and $g(x, y) = xy - 1$, as in Example 1. Here, we have plotted $1.3 \leq x \leq 1.8$ and $0.5 \leq y \leq 1$, which includes one exact solution of the system (1.12) in this case. The plane $z = 0$ is the surface in solid color, $z = f(x, y)$ shows the contour lines, and $z = g(x, y)$ is the surface showing a grid. You see where all three surfaces intersect, and that is the where the solution lies.



You also see in this graph that the tangent plane approximation is pretty good in this region, so replacing the surfaces by their tangent planes will not wreak havoc on the graph. So here is what we do: Take any point (x_0, y_0) so that the three surface intersect *near* $(x_0, y_0, 0)$. Then replace the surfaces $z = f(x, y)$ and $z = g(x, y)$ by their tangent planes at (x_0, y_0) , and compute the intersection of the tangent planes with the plane $z = 0$. This is a linear algebra problem, and hence is easily solved. Replacing $z = f(x, y)$ and $z = g(x, y)$

by the equations of their tangent planes at (x_0, y_0) amounts to the replacement

$$z = f(x, y) \quad \rightarrow \quad z = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

and

$$z = g(x, y) \quad \rightarrow \quad z = g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0)$$

where $\mathbf{x}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$. This transforms (1.13) into

$$\begin{aligned} z &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) \\ z &= g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) \\ z &= 0 . \end{aligned} \tag{1.14}$$

Now we can eliminate z , and pass to the simplified system

$$\begin{aligned} f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) &= 0 \\ g(\mathbf{x}_0) + \nabla g(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) &= 0 . \end{aligned} \tag{1.15}$$

Since $J_{\mathbf{F}}(\mathbf{x}_0) = \begin{bmatrix} \nabla f(\mathbf{x}_0) \\ \nabla g(\mathbf{x}_0) \end{bmatrix}$, this is equivalent to (1.6) by the usual rules for matrix multiplication.

We see from this analysis that how close we come to an exact solution in one step of Newton's method depends on, among other things, how good the tangent plane approximation is at the current approximate solution. We know that tangent plane approximations are good when the norm of the Hessian is not too large. We can also see that there will be trouble if $J_{\mathbf{F}}$ is not invertible, or even if ∇f and ∇g are nearly proportional, in which case $(J_{\mathbf{F}})^{-1}$ will have a large norm. There is a precise theorem, due to the 20th century Russian mathematician Kantorovich that can be paraphrased as saying that if \mathbf{x}_0 is not too far from an exact solution, $\|(J_{\mathbf{F}})^{-1}\|$ is not too large, and each component of \mathbf{F} has a Hessian that is not too large, Newton's method works and converges very fast. The precise statement makes it clear what "not too large" means. We will neither state it nor prove it here – it is quite intricate, and in practice one simply uses the method as described above, and stops the iteration when the answers stop changing in the digits that one is concerned with.

1.4 More variables

We have explained everything so far in the case of two equations in two variables. But if \mathbf{F} is a functions from \mathbb{R}^n to \mathbb{R}^m , so that $\mathbf{F}(\mathbf{x}) = 0$ is a system of m equations in n variables, the passage to the approximate linear system

$$\mathbf{F}(\mathbf{x}_0) + J_{\mathbf{F}}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = 0 . \tag{1.16}$$

by way of the linear approximation

$$\mathbf{F}(\mathbf{x}) \approx \mathbf{F}(\mathbf{x}_0) + J_{\mathbf{F}}(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) . \quad (1.17)$$

is just as valid. We may therefore recursively define the sequence $\{\mathbf{x}_n\}$ by

$$J_{\mathbf{F}}(\mathbf{x}_0)(\mathbf{x}_{n+1} - \mathbf{x}_n) = -\mathbf{F}(\mathbf{x}_0) . \quad (1.18)$$

Notice we do not write this in terms of the inverse of $J_{\mathbf{F}}$ any longer – indeed, when $m \neq n$, the Jacobian will not be square, and there will be no inverse. As long as there are more variables than equations though, we can hope that the system in (1.18) is underdetermined, and hence solvable. We can proceed as before.

Problems

1 Let $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$ where $f(x, y) = x^3 + xy$, and $g(x, y) = 1 - 4y^2 - x^2$. Let $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$.

a Compute $J_{\mathbf{F}}(\mathbf{x})$ and $J_{\mathbf{F}}(\mathbf{x}_0)$.

b Use \mathbf{x}_0 as a starting point for Newton's method, and compute the next approximate solution \mathbf{x}_1 .

c Evaluate $\mathbf{F}(\mathbf{x}_1)$, and compare this with $\mathbf{F}(\mathbf{x}_0)$.

d Draw graphs of the curves implicitly defined by $f(x, y) = 0$ and $g(x, y) = 0$. How many solutions are there of this non linear system?

2 Let $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$ where $f(x, y) = \sqrt{x} + \sqrt{y} - 3$, and $g(x, y) = x^2 + 4y^2 = 18$.

a Compute $\mathbf{F}(\mathbf{x}_0)$ for $\mathbf{x}_0 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}$. does this look like a reasonable starting point? Compute $J_{\mathbf{F}}(\mathbf{x}_0)$. What happens if you try to use \mathbf{x}_0 as your starting point for Newton's method?

b Draw graphs of the curves implicitly defined by $f(x, y) = 0$ and $g(x, y) = 0$. How many solutions are there of this non linear system? Find starting points \mathbf{x}_0 near each of them with integer entries.

c Let \mathbf{x}_0 be the starting point that you found in part (b) that is closest to the x -axis. Compute the next approximate solution \mathbf{x}_1 .

d Evaluate $\mathbf{F}(\mathbf{x}_1)$, and compare this with $\mathbf{F}(\mathbf{x}_0)$.

3 Let $\mathbf{F}(\mathbf{x}) = \begin{bmatrix} f(\mathbf{x}) \\ g(\mathbf{x}) \end{bmatrix}$ where $f(x, y) = \sin(xy) - x$, and $g(x, y) = x^2y - 1$. Let $\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

a Compute $J_{\mathbf{F}}(\mathbf{x})$ and $J_{\mathbf{F}}(\mathbf{x}_0)$.

b Use \mathbf{x}_0 as a starting point for Newton's method, and compute the next approximate solution \mathbf{x}_1 .

c Evaluate $\mathbf{F}(\mathbf{x}_1)$, and compare this with $\mathbf{F}(\mathbf{x}_0)$.

d How many solutions of this system are there in the region $-2 \leq x \leq 2$ and $0 \leq y \leq 10$? Compute each of them to 10 decimal places of accuracy – using a computer, of course.

Section 2: Optimization problems

2.1 What is an optimization problem?

A *optimization problem* in two variables is one in which we are given a function $f(x, y)$, and a set D in the plane of *admissible points*, and we are asked to find either the maximum or minimum value of $f(x, y)$ as (x, y) ranges over D .

It can be the case there is neither a maximum nor a minimum. Consider for example $f(x, y) = x + y$ and $D = \mathbb{R}^2$. Then

$$\lim_{t \rightarrow \infty} f(t, t) = \infty \quad \text{and} \quad \lim_{t \rightarrow \infty} f(-t, -t) = -\infty .$$

However, if D is a bounded closed domain, and if f is continuous, then there is always point (x_1, y_1) in D with the property that

$$f(x_1, y_1) \geq f(x, y) \quad \text{for all } (x, y) \text{ in } D , \quad (2.1)$$

and there is always a point (x_0, y_0) in D with the property that

$$f(x_0, y_0) \leq f(x, y) \quad \text{for all } (x, y) \text{ in } D . \quad (2.2)$$

Definition (Maximizer and minimizer) Any point (x_1, y_1) satisfying (2.1) is called a *maximizer of f in D* , and any point (x_0, y_0) satisfying (2.2) is called a *minimizer of f in D* . The value of f at a maximizer is the *maximum value of f in D* , and the value of f at a minimizer is the *minimum value of f in D* .

To solve an optimization problem is to find all maximizers and minimizers, if any, and the corresponding maximum and minimum values. Our goal in this section is to explain a strategy for doing this.

2.1 A strategy for solving optimization problems

Recall that if $g(t)$ is a function of the single variable t , and we seek to maximize it on the closed bounded interval $[a, b]$, we proceed in three steps:

(1) We find all values of t in (a, b) at which $g'(t) = 0$. Hopefully there are only finitely many of these, say $\{t_1, t_2, \dots, t_n\}$.

(2) Compute $g(t_1), g(t_2), \dots, g(t_n)$, together with $g(a)$ and $g(b)$. The largest number on this finite list is the maximum value, and the smallest is the minimum value. The maximizers are exactly those numbers from among $\{t_1, t_2, \dots, t_n\}$ together with a and b , at which f takes on the maximum values, and similarly for the minimizers.

The reason this works is that if t belongs to the open interval (a, b) , and $g'(t) \neq 0$, it is possible to move either “uphill” or “downhill” while staying within $[a, b]$ by moving a bit to the right or the left, depending on whether the slope is positive or negative. *Hence no such point can be a maximizer or a minimizer.* We are applying the Sherlock Holmes principle:

- *When you have eliminated the impossible, whatever else remains, however unlikely, is the truth.*

When all goes well, the elimination procedure reduces an *infinite* sets of suspects – all of the points in $[a, b]$ – to a *finite* list of suspects: $\{t_1, t_2, \dots, t_n\}$ together with a and b . Finding the minimizers and maximizers among a finite list of points is easy – just compute the value of f at each point on the list, and see which ones give the largest and smallest values.

We now adapt this to two or more dimensions, focusing first on two. Suppose that D is a closed bounded domain. Let U be the interior of D and let B be the boundary. For example, if D is the closed unit disk

$$D = \{\mathbf{x} : |\mathbf{x}| \leq 1\}$$

we have

$$U = \{\mathbf{x} : |\mathbf{x}| < 1\} \quad \text{and} \quad B = \{\mathbf{x} : |\mathbf{x}| = 1\} .$$

Notice that in this case, the boundary consists of infinitely many points.

- *In optimization problems, the big difference between one dimension and two or more dimensions is that in one dimension, the interval $[a, b]$ has only finitely boundary points – two – and there is no problem with throwing them into the list of suspects. But in two or more dimensions, there are the boundary will generally consist of infinitely many points, and if we throw them all onto the list of suspects, we make the list infinite, and therefore useless.*

We will therefore have to develop a “sieve” to filter the boundary B for suspects, as well as a “sieve” to filter the interior U for suspects. Here is the interior sieve:

If \mathbf{x} belongs to U , the interior of D , and $\nabla f(\mathbf{x}) \neq 0$, then \mathbf{x} is not a suspect. Recall the $\nabla f(\mathbf{x})$ points in the direction of steepest ascent. so if one moves a bit away from \mathbf{x} in the direction $\nabla f(\mathbf{x})$, one moves to higher ground. Likewise, if one moves a bit away from \mathbf{x} in the direction $-\nabla f(\mathbf{x})$, one moves to lower ground.

Since \mathbf{x} is in the interior U of D , it is possible to move some positive distance from \mathbf{x} in any direction and stay inside D . Hence, if \mathbf{x} belongs to U , and $\nabla f(\mathbf{x}) \neq 0$, there are nearby points at which f takes on strictly higher and lower values. Such a point \mathbf{x} cannot be a maximizer!

Apply the Sherlock Holmes principle: Eliminate all points \mathbf{x} in U at which $\nabla f(\mathbf{x}) \neq 0$, and the remaining points are the only valid suspects in U . There are exactly the critical points in U – the points in U at which $\nabla f(\mathbf{x}) = 0$.

- *The suspect list from the interior U consists exactly of the critical points in U*

This is the “sieve” with which we filter the interior: We filter out the non critical points.

Next, we need a “sieve” for the boundary. Here we need to make some assumptions. We will suppose first that the boundary B of D is the set of solutions of some equation

$$g(x, y) = 0 .$$

For example, when B is the unit circle, then we have

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

Here is the sieve in this case:

Theorem 1 *Suppose that f and g are two functions on \mathbb{R}^2 with continuous first order partial derivatives. Let B denote the level curve of g given by $g(x, y) = 0$. Let \mathbf{x}_0 be any point on B , and suppose that $\nabla g(\mathbf{x}_0) \neq 0$. Form the 2×2 matrix $[\nabla f(\mathbf{x}_0), \nabla g(\mathbf{x}_0)]$ whose first column is $\nabla f(\mathbf{x}_0)$, and whose second column is $\nabla g(\mathbf{x}_0)$.*

Then, if

$$\det([\nabla f(\mathbf{x}_0), \nabla g(\mathbf{x}_0)]) \neq 0 ,$$

\mathbf{x}_0 *neither maximizes nor minimizes f on B .*

The idea behind the proof is this: Imagine that B is a path in a hilly landscape, and you are walking along this path, and passing through \mathbf{x}_0 . Since B is a level curve, the direction in which you are moving at that instant is one of the two directions that are orthogonal to $\nabla g(\mathbf{x}_0)$. Let \mathbf{v} denote your velocity vector as you pass through \mathbf{x}_0 . We may assume that you are actually moving so $\mathbf{v} \neq 0$. We have

$$\mathbf{v} \cdot \nabla g(\mathbf{x}_0) = 0 . \tag{2.3}$$

Now, if as you move from \mathbf{x}_0 in the direction \mathbf{v} , if you are “cutting across” contour lines of f at any non zero angle, you are at that instant going either strictly uphill or strictly downhill, and there is points just behind and ahead of you on the path at which the ground is higher and lower. This means that \mathbf{x}_0 is neither a minimizer, nor a maximizer.

• *In our search for minimizers or maximizers along the “path” B , we can eliminate all points on the path at which it “cuts across” level sets of f .*

What remains are the points at which B runs along, instead of across, a level set of f . That is, what remains are points \mathbf{x}_0 at which the level curves of f and g have the same tangent direction. But they have the same tangent direction if and only if they have the same normal direction, and so $\nabla f(\mathbf{x}_0)$ and $\nabla g(\mathbf{x}_0)$ must be proportional. But in this case

$$\det([\nabla f(\mathbf{x}_0), \nabla g(\mathbf{x}_0)]) = 0 . \tag{2.4}$$

We will give a more formal proof at the end of this section. It is well worth going through it, but our first priority is to get to some examples. We now have a strategy for searching out maximizers and minimizers in a region D bounded by a level curve given by an equation of the form $g(x, y) = 0$.

(1) Find all critical points in U , the interior of D .

(2) Find all points on B , the boundary of D , at which (2.4) holds.

(3) The combined list of points found in (1) and (2) is a comprehensive list of suspected maximizers and minimizers. Hopefully it is a finite list. Now interrogate the suspects:

Evaluate f at each of them, and see which produce the largest and smallest values. Case closed.

Example 1 (Finding minimizers and maximizers) Let $f(x, y) = x^4 + y^4 + 4xy$, which we have used in previous examples. Let D be the closed disk of radius 4 centered on the origin. We will now find the maximizers and minimizers of f in D .

We can write the equation for the boundary in the form $g(x, y) = 0$ by putting

$$g(x, y) = x^2 + y^2 - 16 .$$

First, we look for the critical points. In Example 7 from Section 3, we found that f has exactly 3 critical points in all of \mathbb{R}^2 , and all of them happen to be in the interior of D . They are

$$(0, 0) \quad (1, -1) \quad \text{and} \quad (-1, 1) . \tag{2.5}$$

Next, since

$$\nabla f(x, y) = 4 \begin{bmatrix} x^3 + y \\ y^3 + x \end{bmatrix} \quad \text{and} \quad \nabla g(x, y) = 2 \begin{bmatrix} x \\ y \end{bmatrix} .$$

We can ignore the factors of 4 and 2 in these gradients, and they have no effect on (2.4), and so we compute the determinant of

$$\begin{bmatrix} x^3 + y & x \\ y^3 + x & y \end{bmatrix} .$$

Setting it equal to zero gives us the equation $x^3y + y^2 - y^3x - x^2 = 0$. Combining this with $g(x, y) = 0$ we have the system of equations

$$\begin{aligned} x^3y + y^2 - y^3x - x^2 &= 0 \\ g(x, y) = x^2 + y^2 - 16 &= 0 . \end{aligned}$$

The first equation can be written as

$$(x^2 - y^2)(xy - 1) = 0 .$$

Either $x^2 - y^2 = 0$, or else $xy - 1 = 0$, or both.

Suppose that $x^2 - y^2 = 0$. Then we can eliminate y from the second equation, obtaining $2x^2 = 16$, or $x = \pm 2\sqrt{2}$. If $y^2 = x^2$, then $y = \pm x$, so we get 4 solutions of the system this way:

$$(2\sqrt{2}, 2\sqrt{2}) \quad (2\sqrt{2}, -2\sqrt{2}) \quad (-2\sqrt{2}, -2\sqrt{2}) \quad \text{and} \quad (-2\sqrt{2}, 2\sqrt{2}) . \tag{2.6}$$

On the other hand, if $xy - 1 = 0$, $y = 1/x$, and eliminating y from the second equation gives us

$$x^2 + x^{-2} - 16 = 0 \tag{2.7}$$

Multiplying through by x^2 , and writing $u = x^2$, we get

$$u^2 - 16u = -1$$

so $u = 8 \pm \sqrt{63}$. Since $u = x^2$, there are four values of x that solve (2.7), namely $\pm\sqrt{8 \pm \sqrt{63}}$. The corresponding y values are given by $y = 1/x$. We obtain the final 4 solutions of (2.4):

$$(\pm\sqrt{8 \pm \sqrt{63}}, \pm 1/(\sqrt{8 \pm \sqrt{63}})) . \tag{2.8}$$

Now for the interrogation phase: First the critical points:

$$f(0, 0) = 0$$

while

$$f(1, -1) = f(-1, 1) = -2 .$$

At each of the 4 points in (2.6), $f = 96$. At two of the four points in (2.8), f takes on the value

$$(8 + \sqrt{63})^2 + (8 + \sqrt{63})^{-2} - 4(8 + \sqrt{63}) ,$$

while at the other two, it takes on the value

$$(8 - \sqrt{63})^2 + (8 - \sqrt{63})^{-2} - 4(8 - \sqrt{63}) ,$$

Since

$$\sqrt{63} = \sqrt{64(1 - 1/64)} = 8\sqrt{(1 - 1/64)} \approx 8 \left(1 - \frac{1}{128}\right) = 8 - \frac{1}{16} .$$

Hence

$$8 + \sqrt{63} \approx 16 \quad \text{and} \quad 8 - \sqrt{63} \approx \frac{1}{16} .$$

From this is clear that the largest value on our finite list is $(8 - \sqrt{63})^2 + (8 - \sqrt{63})^{-2} - 4(8 - \sqrt{63})$. This is the maximum value, and the corresponding maximizers are

$$(\sqrt{8 - \sqrt{63}}, 1/(\sqrt{8 - \sqrt{63}})) \quad \text{and} \quad (-\sqrt{8 - \sqrt{63}}, -1/(\sqrt{8 - \sqrt{63}})) .$$

The smallest value on our list is -2 . This is the minimum values, and the corresponding minimizers are

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

Proof of Theorem 1: let $\mathbf{x}(t)$ be a parameterization of a portion of the level curve of g through \mathbf{x}_0 . As long as g has continuous partial derivatives in a neighborhood of \mathbf{x}_0 , and $\nabla g(\mathbf{x}_0) \neq 0$, the Implicit Function Theorem guarantees the existence of such a parameterized curve with $|\mathbf{v}| = 1$ where.

$$\mathbf{v} = \frac{d}{dt} \mathbf{x}(0) .$$

That is, the parameterization can be adjusted so that the curve passes through \mathbf{x}_0 at unit speed. In terms of the informal discussion we gave above, this parameterized path represents your motion across the landscape as you hike along the path.

Since $g(\mathbf{x}(t))$ is constant, by the definition of a level curve,

$$\mathbf{v} \cdot \nabla g(\mathbf{x}_0) = 0 . \tag{2.9}$$

What is your instantaneous rate of change of altitude as you pass through \mathbf{x}_0 ? This is the directional derivative

$$\mathbf{v} \cdot \nabla f(\mathbf{x}_0) .$$

This quantity *must* be zero if \mathbf{x}_0 is to either minimize or maximize f along B . The reason is that if $\mathbf{v} \cdot \nabla f(\mathbf{x}_0) > 0$, you are going strictly uphill as you pass through \mathbf{x}_0 , and so there is higher ground just ahead on your path, and lower ground just behind. Likewise,

if $\mathbf{v} \cdot \nabla f(\mathbf{x}_0) < 0$, you are going strictly downhill as you pass through \mathbf{x}_0 , and so there is lower ground just ahead on your path, and higher ground just behind.

Hence the only way that \mathbf{x}_0 can be a minimizer if

$$\mathbf{v} \cdot \nabla f(\mathbf{x}_0) = 0 . \quad (2.10)$$

To make this equation more useful, let's express it directly in terms of f and g . This is easy: Since $\mathbf{v} \neq 0$, and since we have from (2.9) and (2.10) that both $\nabla g(\mathbf{x}_0)$ and $\nabla f(\mathbf{x}_0)$ are orthogonal to \mathbf{v} , they must be parallel to each other. Hence the columns of $[\nabla f(\mathbf{x}_0), \nabla g(\mathbf{x}_0)]$ are proportional, and its rank is one, and so its determinant is zero.

On the other hand, if

$$\det([\nabla f(\mathbf{x}_0), \nabla g(\mathbf{x}_0)]) = 0 , \quad (2.11)$$

then the rank of the matrix cannot be two. Since $\nabla g(\mathbf{x}_0) \neq 0$, it is at least one, and hence it is exactly one. So this implies that $\nabla g(\mathbf{x}_0)$ and $\nabla f(\mathbf{x}_0)$ are parallel, and so (2.10) follows from (2.9) whenever (2.11) holds. ■

2.3 More variables

If there are more than two variables, we can have more than one constrain equation. Consider the problem of maximizing $f(x, y, z) = xyz$ subject to the constraints $g(x, y, z) = 0$ and $h(x, y, z) = 0$ where

$$g(x, y, z) = x^2 + y^2 + z^2 - 1$$

and

$$h(x, y, z) = x + y + z - 1 .$$

Then $g(x, y, z) = 0$ is the implicit description of the unit sphere, and $h(x, y, z) = 0$ is the implicit description of a plane slicing through the sphere. The admissible points – the ones satisfying both constraints – are exactly the points on the intersection of the unit sphere and this plane. The intersection is a circle consisting of infinitely many points. which one maximizes f ?

Let \mathbf{x}_0 be any point on the circle produced by the intersection of the plane and the unit sphere. Consider the tangent planes to the surfaces $g(\mathbf{x}) = 0$ and $h(\mathbf{x}) = 0$ at \mathbf{x}_0 . The normal vectors to these planes are $\nabla g(\mathbf{x}_0)$ and $\nabla h(\mathbf{x}_0)$ respectively. any vector that is orthogonal to both of these vectors is tangent to the circle. Let \mathbf{v} be any such vector. Then $\mathbf{v} \cdot \nabla f(\mathbf{x}_0)$ is the rate of change of f as one passes through \mathbf{x}_0 in the direction \mathbf{v} , which is tangent to the circle. Thus, if \mathbf{x}_0 is either a minimum or a maximum, this rate of change must be zero. That is:

- If \mathbf{x}_0 is a maximizer of $f(\mathbf{x})$ subject to $g(\mathbf{x}) = 0$ and $h(\mathbf{x}) = 0$, then $\mathbf{v} \cdot \nabla f(\mathbf{x}_0) = 0$ for all vectors \mathbf{v} satisfying $\mathbf{v} \cdot \nabla g(\mathbf{x}_0) = 0$ and $\mathbf{v} \cdot \nabla h(\mathbf{x}_0) = 0$

Now, $\mathbf{v} \cdot \nabla f(\mathbf{x}_0) = 0$ for all vectors \mathbf{v} satisfying $\mathbf{v} \cdot \nabla g(\mathbf{x}_0) = 0$ and $\mathbf{v} \cdot \nabla h(\mathbf{x}_0) = 0$ if and only if $\nabla f(\mathbf{x}_0)$ is a linear combination of $\nabla g(\mathbf{x}_0)$ and $\nabla h(\mathbf{x}_0)$, and this is the case if and only if

$$\det \left(\begin{bmatrix} \nabla f(\mathbf{x}_0) \\ \nabla g(\mathbf{x}_0) \\ \nabla h(\mathbf{x}_0) \end{bmatrix} \right) = 0 . \quad (2.12)$$

This gives us an equation to check. Along with $g(x, y, z) = 0$ and $h(x, y, z) = 0$, we now have three equations in three variables, and can solve to find a list of candidates for the maximizer.

Nothing in the reasoning leading to this system of equations relied on the specific forms of f , g and h , so this is generally valid. Also, the reasoning applies equally well to minima and maxima. Since we do have explicit forms for f , g and h we can go on and apply it.

Example 2 (Finding minima and maxima in three variables with two constraints) As above, consider the problem of maximizing $f(x, y, z) = xyz$ subject to the constraints

$$g(x, y, z) = 0 \quad \text{and} \quad h(x, y, z) = 0$$

where

$$g(x, y, z) = x^2 + y^2 + z^2 - 1$$

and

$$h(x, y, z) = x + y + z - 1 .$$

Then

$$\begin{bmatrix} \nabla f(\mathbf{x}) \\ \nabla g(\mathbf{x}) \\ \nabla h(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} yz & xz & xy \\ 2x & 2y & 2z \\ 1 & 1 & 1 \end{bmatrix}$$

so that (2.12) reduces to

$$x^2(y - z) + y^2(z - x) + z^2(x - y) = 0 .$$

If you go through the algebra carefully – this takes some doing – you will find 6 solutions:

$$(1, 0, 0) \quad (0, 1, 0) \quad (0, 0, 1)$$

and

$$(2/3, 2/3, -1/3) \quad (2/3, -1/3, 2/3) \quad (-1/3, 2/3, 2/3) .$$

At each of the first 3 points $f = 0$, at each of the remaining 3 points $f = -4/9$. Hence the maximum value of f subject to these constraints is 0, and the minimum value is $-4/9$.

The reasoning leading to (2.12) is more flexible and powerful than the formula itself. For instance, suppose we have just one constraint $g(x, y, z) = 0$ in three variables. Suppose for example that we want to maximize the function $f(x, y, z) = xyz$ over the unit sphere. We still have, for the exact same reasons, that is \mathbf{x}_0 is a maximizer, $\nabla f(\mathbf{x}_0)$ must be a linear combination of the gradients of the constraint functions. When there is just one constraint function g , this becomes the statement that $\nabla f(\mathbf{x}_0)$ must be a multiple of $\nabla g(\mathbf{x}_0)$. This is the case, if and only if

$$\nabla f(\mathbf{x}_0) \times \nabla g(\mathbf{x}_0) = 0 .$$

Example 3 (Finding minima and maxima in three variables with one constraints) As above, consider the problem of maximizing $f(x, y, z) = xyz$ subject to the constraint $g(x, y, z) = 0$ where

$$g(x, y, z) = x^2 + y^2 + z^2 - 1 .$$

Then

$$\nabla f(\mathbf{x}) \times \nabla g(\mathbf{x}) = \begin{bmatrix} x(z^2 - y^2) \\ y(x^2 - z^2) \\ z(y^2 - x^2) \end{bmatrix} .$$

Setting this equal to zero would at first seem to give us three equations, but only two of them are independent. Keeping the first two together with the constraint equation gives us the system

$$\begin{aligned}x(z^2 - y^2) &= 0 \\y(x^2 - z^2) &= 0 \\x^2 + y^2 + z^2 &= 1 .\end{aligned}$$

The first equation says $x = 0$ or $z^2 = y^2$. If $x = 0$, the second equation says $yz^2 = 0$, so then either $y = 0$ or $z = 0$. If $x = y = 0$, the third equation says $x = \pm 1$. Hence if $x = 0$, we have the solutions $(0, 0, 1)$ and $(0, 1, 0)$. Otherwise if $z^2 = y^2$. If $y = 0$, we get the solution $(1, 0, 0)$. Otherwise, if $y \neq 0$, we get from the second equation that $x^2 = z^2$ too, so

$$x^2 = y^2 = z^2 .$$

Now the third equation says that the common value is $1/3$. Hence we have the nine solutions

$$(1, 0, 0) \quad (0, 1, 0) \quad (0, 0, 1)$$

and

$$(\pm 1/\sqrt{3}, \pm 1/\sqrt{3}, \pm 1/\sqrt{3}) .$$

From this list we see that the minimum value of f on the unit sphere is $-3^{-3/2}$, and the maximum value is $3^{-3/2}$.

Problems

1 Let D be the ellipse bounded by the ellipse $x^2 + 4y^2 + 3y = 8$ so that D consists of all points (x, y) satisfying

$$x^2 + 4y^2 + 3y \leq 8 .$$

Let $f(x, y) = 2 - x - 2y$. Find the minimum and maximum values of f on D , and find all minimizers and maximizers.

2 Let D be the ellipse bounded by the ellipse $x^2 + 4y^2 + 3y = 8$ so that D consists of all points (x, y) satisfying

$$x^2 + 4y^2 + 3y \leq 8 .$$

Let $f(x, y) = (x + y)(2 - x - 2y)$. Find the minimum and maximum values of f on D , and find all minimizers and maximizers.

2 Find the point on the graph of $y = x^2$ that is closest to $(3, 1)$.

4 Find the maximum value of xy given that $x, y \geq 0$ and $x + y \leq 4$.

5 Let D be the region consisting of all points (x, y) satisfying

$$x^2 - 1 \leq y \leq 1 + x^2 .$$

$x^2 + y^2 \leq 1$. Let $f(x, y) = x + 2y + 3$. Find the minimum and maximum values of f on D , and find all minimizers and maximizers.

6 Let D be the region consisting of all points (x, y) satisfying

$$x^2 - 4 \leq y \leq 4 + x^2 .$$

$x^2 + y^2 \leq 1$. Let $f(x, y) = (x - 1)^2 + (y - 1)^2$. Find the minimum and maximum values of f on D , and find all minimizers and maximizers.

7 Compute the minimum and maximum values of $f(x, y, z) = xy + 2yz + 3xy$ on the unit sphere.