Lagrange multipliers, introduction

Google Classroom

The "Lagrange multipliers" technique is a way to solve constrained optimization problems. Super useful!

Background

- Contour maps
- Gradient
- Local maxima and minima

What we're building to:

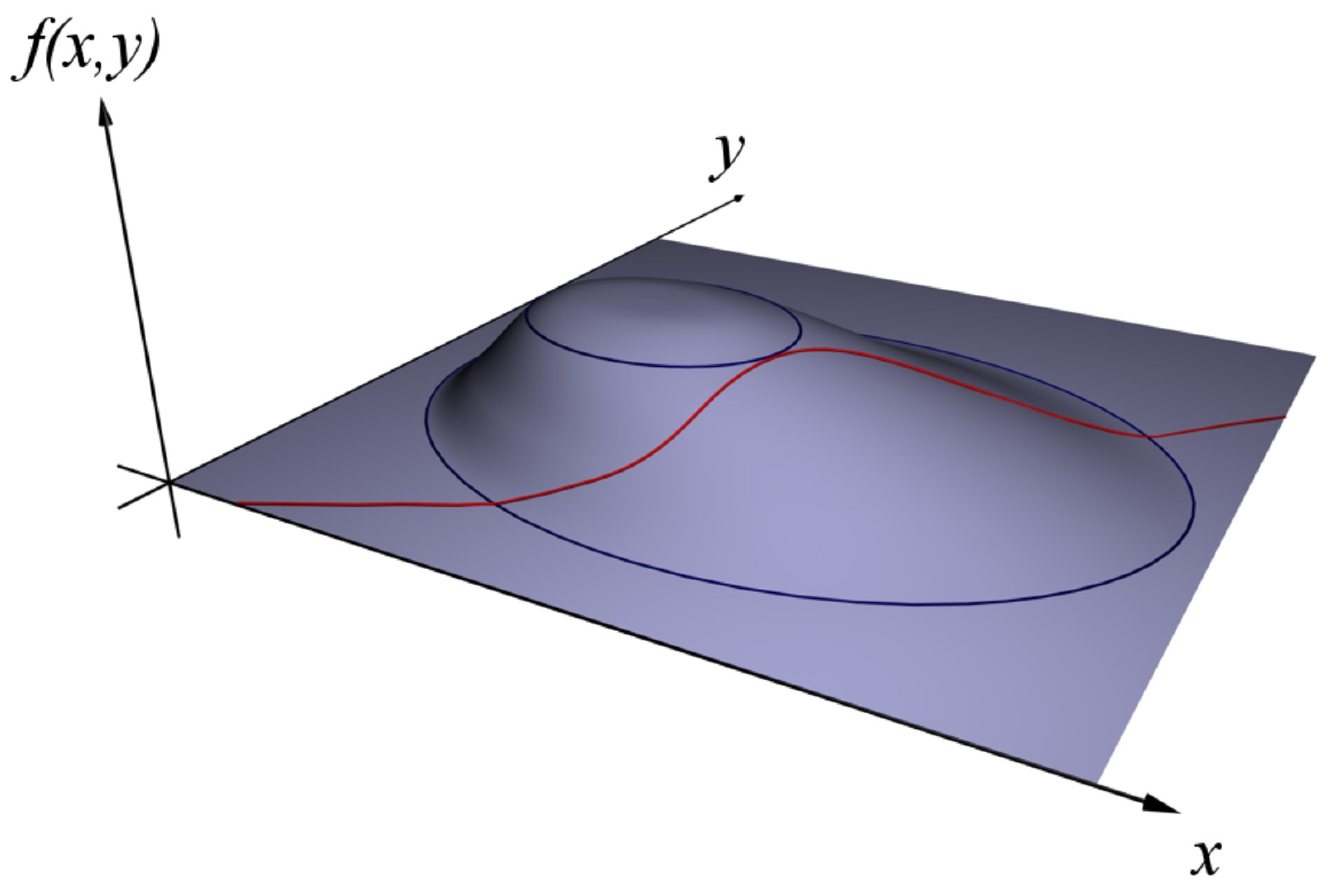
- The Lagrange multiplier technique lets you find the maximum or minimum of a multivariable function $f(x, y, \dots)$ when there is some constraint on the input values you are allowed to use.

- This technique only applies to constraints that look something like this:

$$g(x, y, \dots) = c$$

Here, g is another multivariable function with the same input space as f, and c is some constant. *[Hide picture]*

> For example, if the input space is two-dimensional, the graph of f with the line representing $g(x, y) = c$ projected onto it might look something like this:



> The goal is to find the highest point on that red line.

- The core idea is to look for points where the contour lines of f and g are tangent to each other.

- This is the same as finding points where the gradient vectors of f and g are parallel to each other.

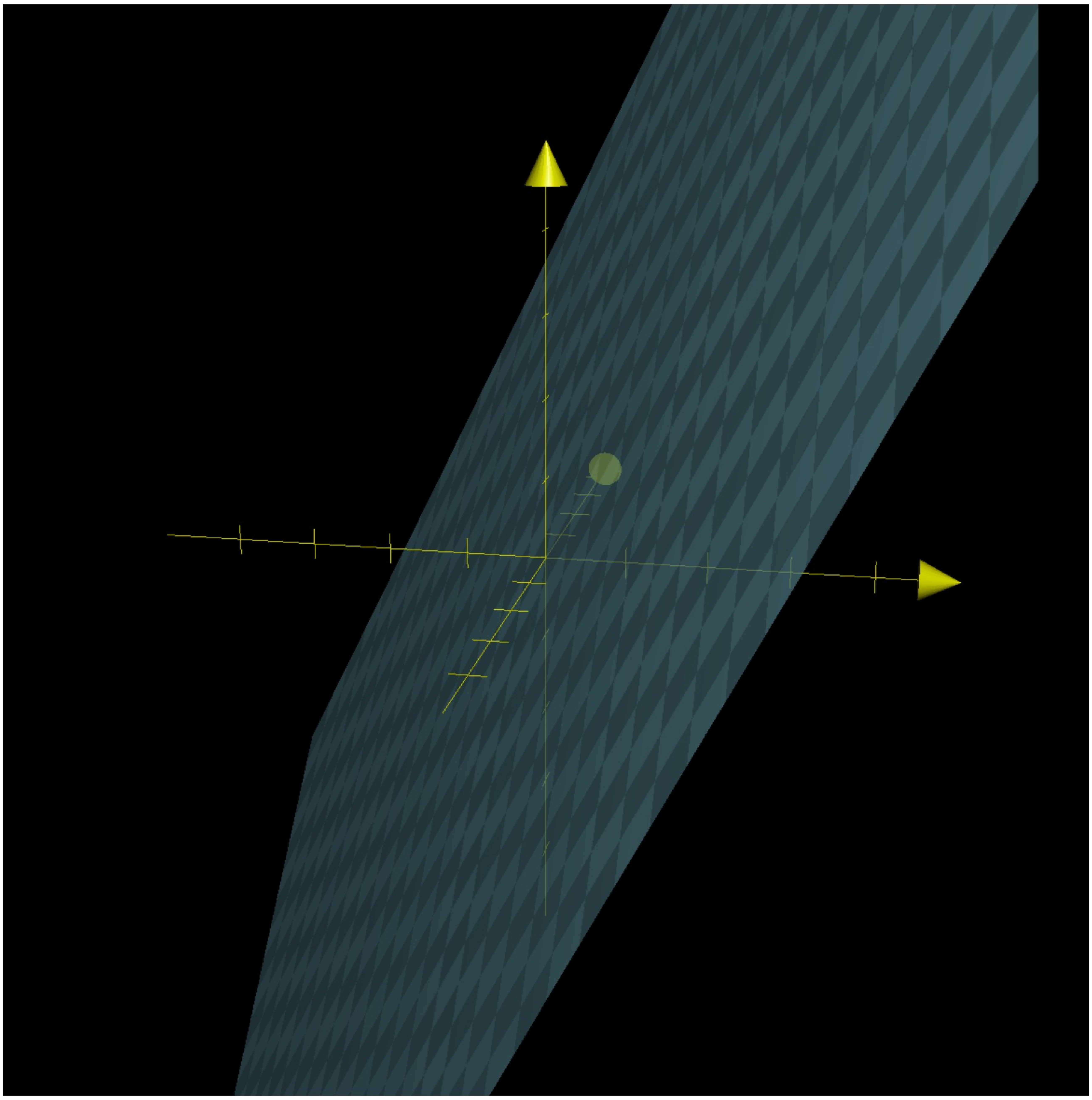
- The entire process can be boiled down into setting the gradient of a certain function, called the **Lagrangian**, equal to the zero vector.
  [Hide explanation]

> - **Step 1**: Introduce a new variable λ, and define a new function $\mathcal{L}$ as follows:
>
> $$\mathcal{L}(x, y, \dots, \lambda) = f(x, y, \dots) - \lambda(g(x, y, \dots) - c)$$
>
> This function $\mathcal{L}$ is called the "Lagrangian", and the new variable λ is referred to as a "Lagrange multiplier"
>
> - **Step 2**: Set the gradient of $\mathcal{L}$ equal to the zero vector.
>
> $$\nabla \mathcal{L}(x, y, \dots, \lambda) = \mathbf{0} \quad \leftarrow \text{Zero vector}$$
>
> In other words, find the **critical points** of $\mathcal{L}$.
>
> - **Step 3**: Consider each solution, which will look something like $(x_0, y_0, \dots, \lambda_0)$. Plug each one into f. Or rather, first remove the λ_0 component, then plug it into f, since f does not have λ as an input. Whichever one gives the greatest (or smallest) value is the maximum (or minimum) point your are seeking.

Motivating example

Suppose you want to maximize this function:

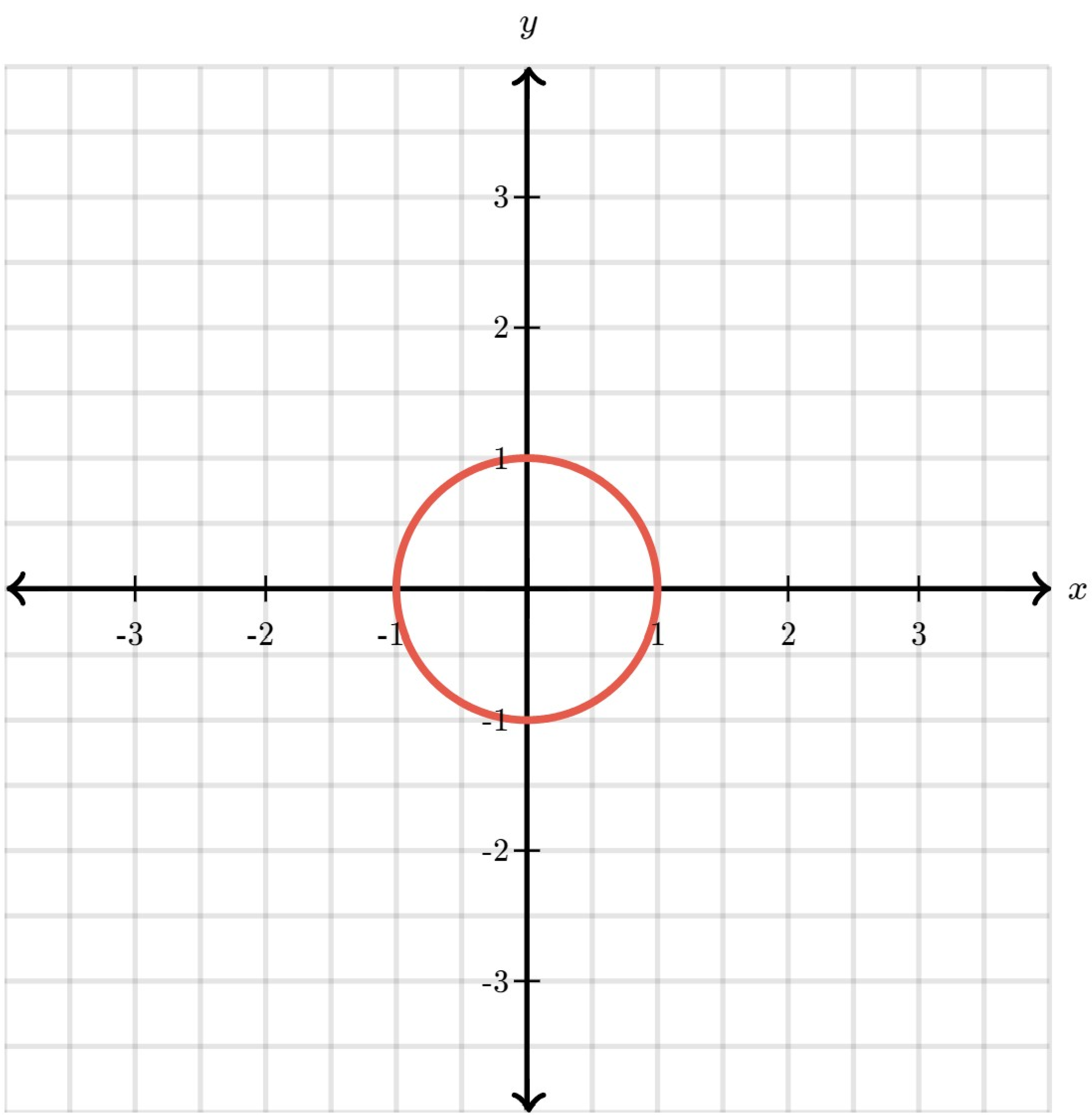
$$f(x, y) = 2x + y$$



Plot of the function $f(x, y) = 2x + y$

But let's also say you limited yourself to inputs (x, y) which satisfy the following equation:

$$x^2 + y^2 = 1$$

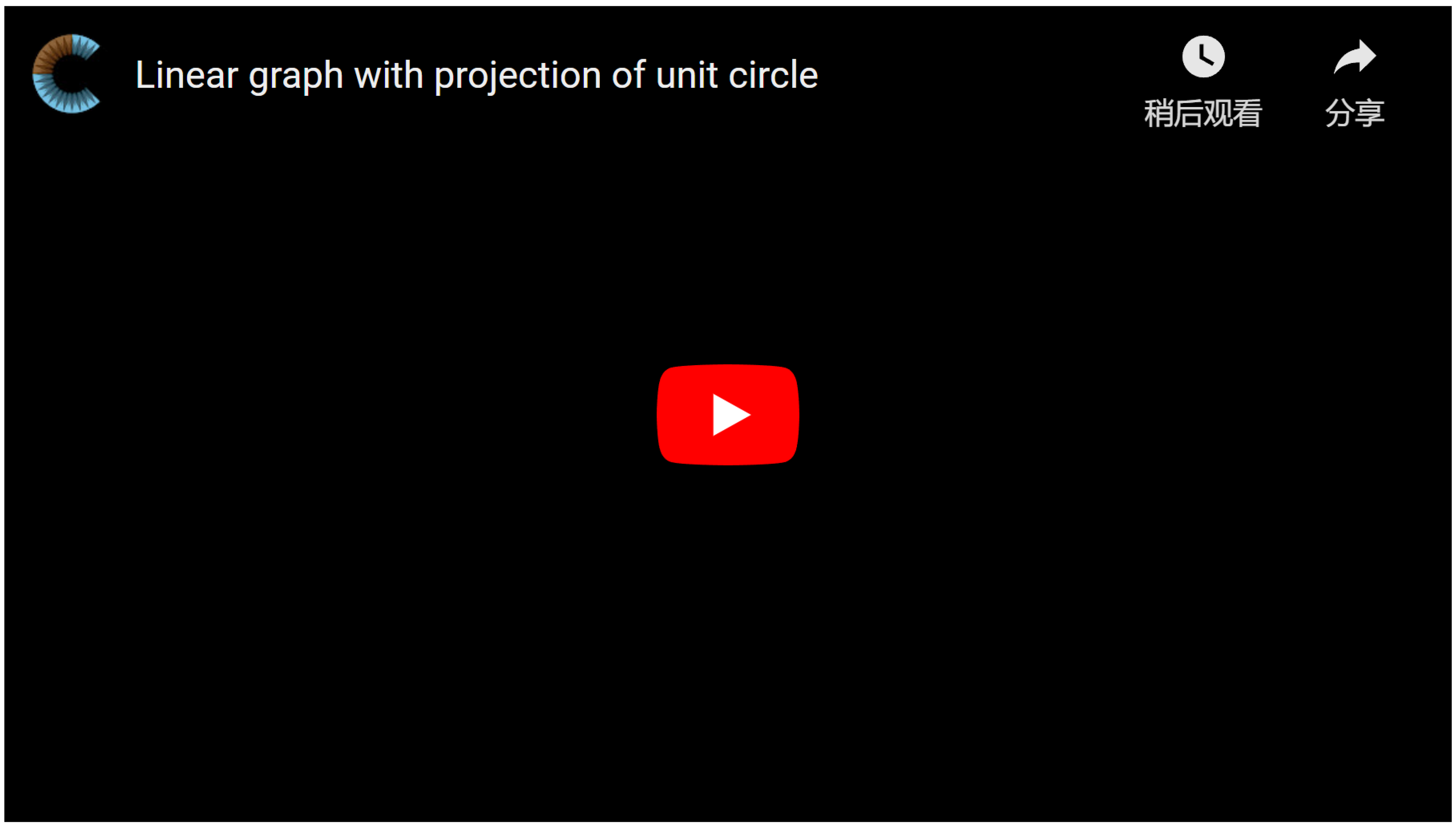


All points (x, y) satisfying $x^2 + y^2 = 1$, also known as the unit circle.

In other words, for which point (x, y) on the unit circle is the value $2x + y$ biggest?

This is what's known as a **constrained optimization problem**. The restriction to points where $x^2 + y^2 = 1$ is called a "constraint", and $f(x, y) = 2x + y$ is the function that needs to be optimized.

Here's one way to visualize this: First draw the graph of $f(x, y)$, which looks like a slanted plane since f is linear. Next, project the circle $x^2 + y^2 = 1$ from the xy-plane vertically onto the graph of f. The maximum we are seeking corresponds with the highest point of this projected circle on the graph.



See video transcript

More general form

In general, constrained optimization problems involve maximizing/minimizing a multivariable function whose input has any number of dimensions:

$$f(x, y, z, \dots)$$

Its output will always be one-dimensional, though, since there's not a clear notion of "maximum" with vector-valued outputs.

The type of constraints that the Lagrange multiplier technique applies to must take the form of some other multivariable function $g(x, y, z, \dots)$ being set equal to a constant c.

$$g(x, y, z, \dots) = c$$

Since this is meant to be a constraint on the input of f, the number of dimensions in the input of g is the same as that of f. For example, the example outlined above fits this general form as follows:

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

$$g(x, y) = x^2 + y^2$$

$$c = 1$$

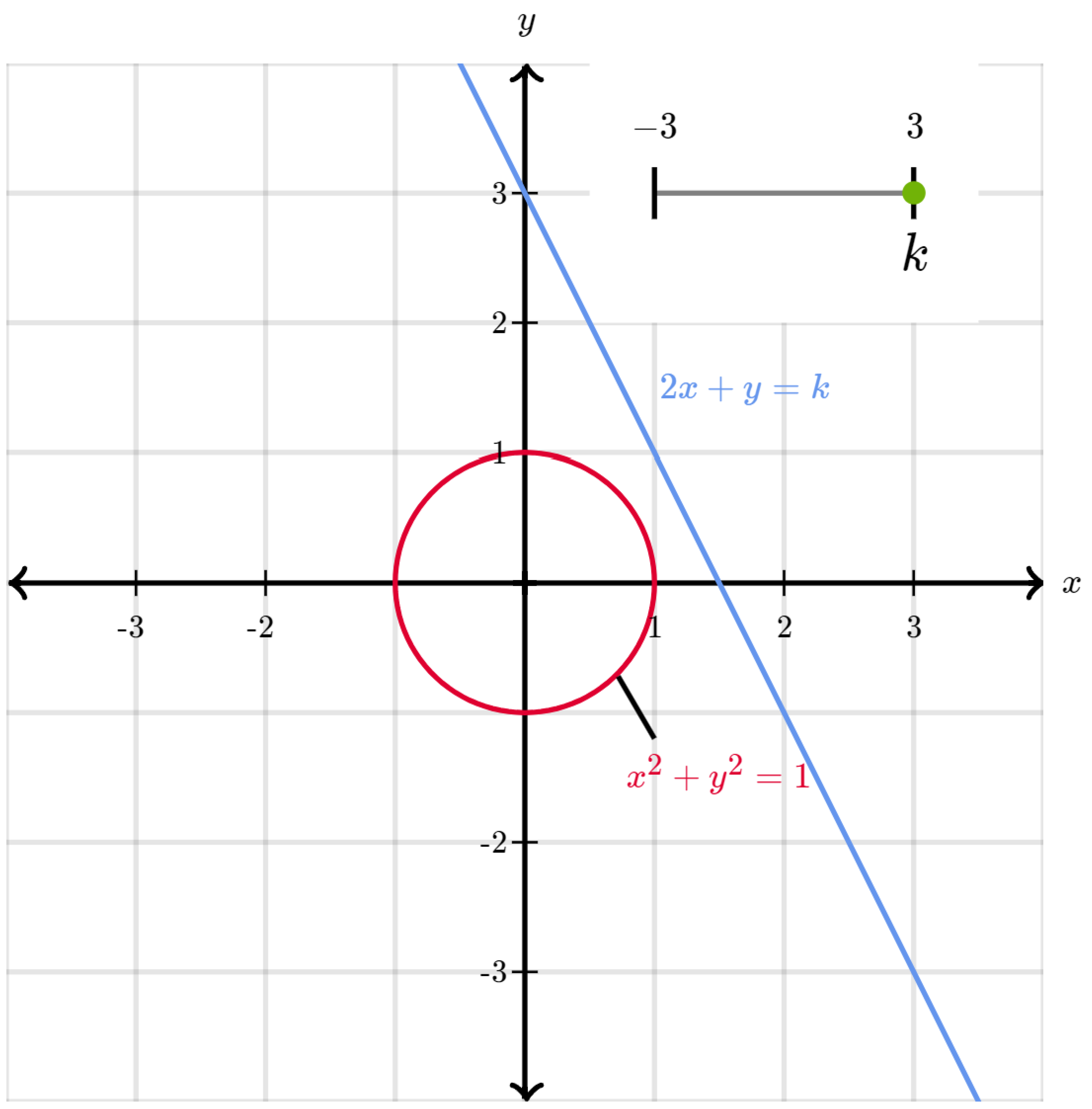
[Hide explanation]

> There is a generalization of the Lagrange multiplier technique that applies to situations with multiple constants on the input.
>
> $$\begin{aligned} g_1(x, y, z, \dots) &= c_1 \\ g_2(x, y, z, \dots) &= c_2 \\ g_3(x, y, z, \dots) &= c_3 \\ &\vdots \end{aligned}$$
>
> However, in this article, I will only talk about the single-constraint case.

Using contour maps

Reasoning about this problem becomes easier if we visualize f not with a graph, but with its contour lines.

As a reminder, a contour line of $f(x, y)$ is the set of all points where $f(x, y) = k$ for some constant k. The following interactive tool shows how this line (drawn in blue) changes as the constant k changes. The circle $g(x, y) = 1$ is also shown (in red). Try to make k as big/small as possible while still allowing contour line of f to intersect the circle.



Concept check: What does it mean if for a particular value of k, the blue line representing $f(x, y) = k$ does **not** intersect the red circle representing $g(x, y) = 1$?

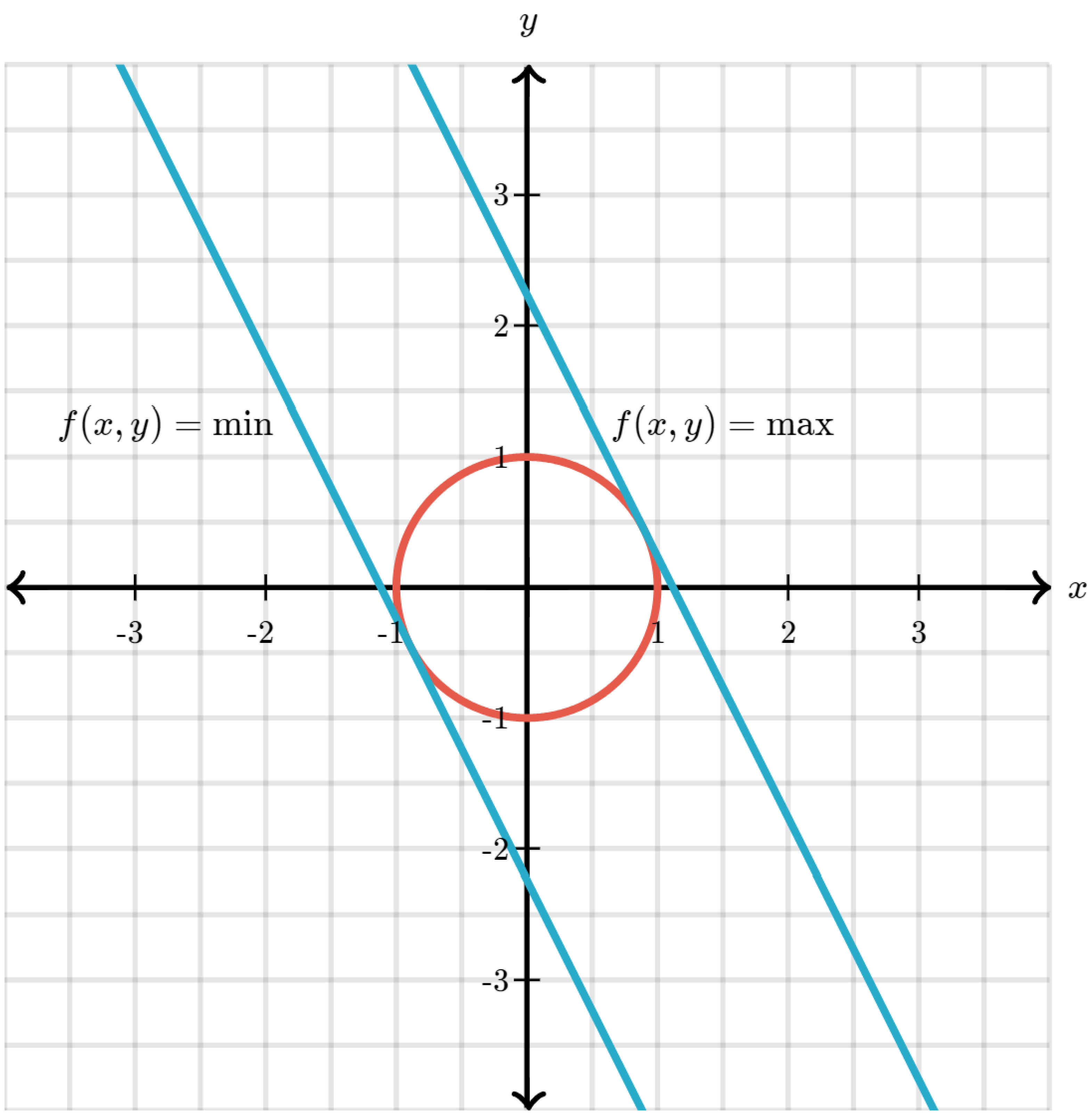
Choose 1 answer:

- (A) There are no values of x and y satisfying both $2x + y = k$ and $x^2 + y^2 = 1$
- (B) The given optimization problem has no solutions.

Check

Notice, the circle where $g(x, y) = 1$ can be thought of as a particular contour line of the function g. So with that, here's the clever way to think about constrained optimization problems:

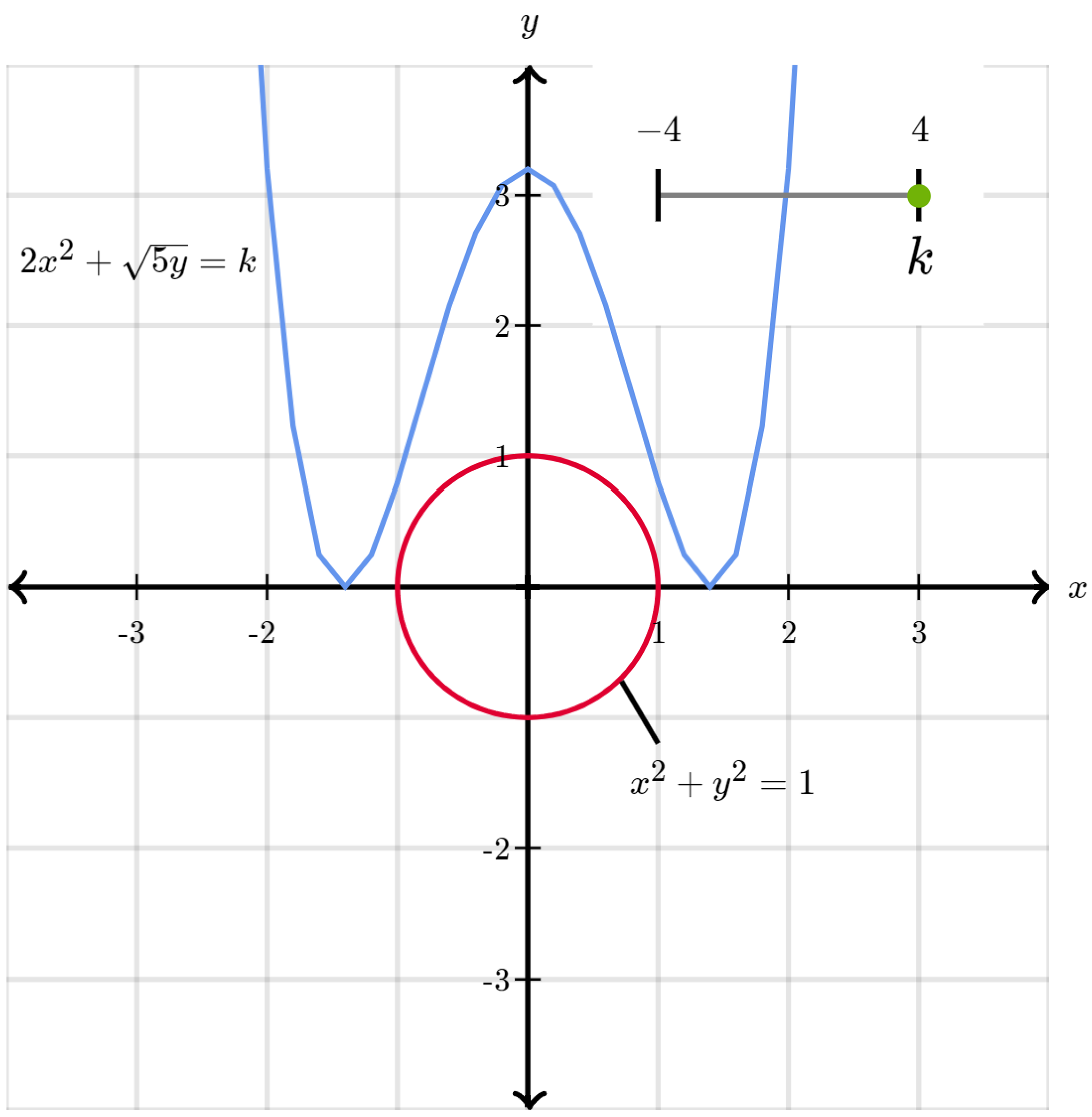
Key observation: The maximum and minimum values of f, subject to the constraint $g(x, y) = 1$, correspond with contour lines of f that are **tangent** to the contour representing $g(x, y) = 1$.



If f were a different function, its contours might not always be straight lines. This is unique to our example since f is linear. For example, take a look at this function:

$$f(x,y) = 2x^2 + \sqrt{5y},$$

Its contour lines look like this:

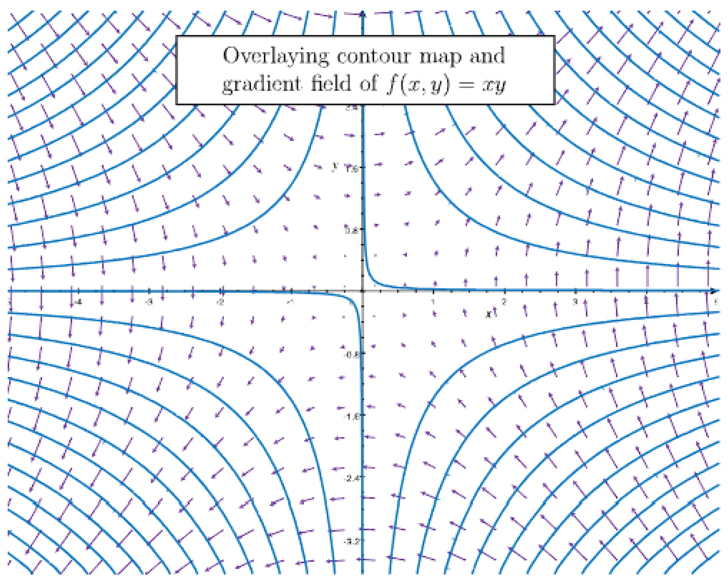


That said, the key observation still holds, and is worth repeating: When k is a maximum or minimum of f subject to the constraint, the contour line for $f(x,y) = k$ will be tangent to contour representing $g(x,y) = 1$.

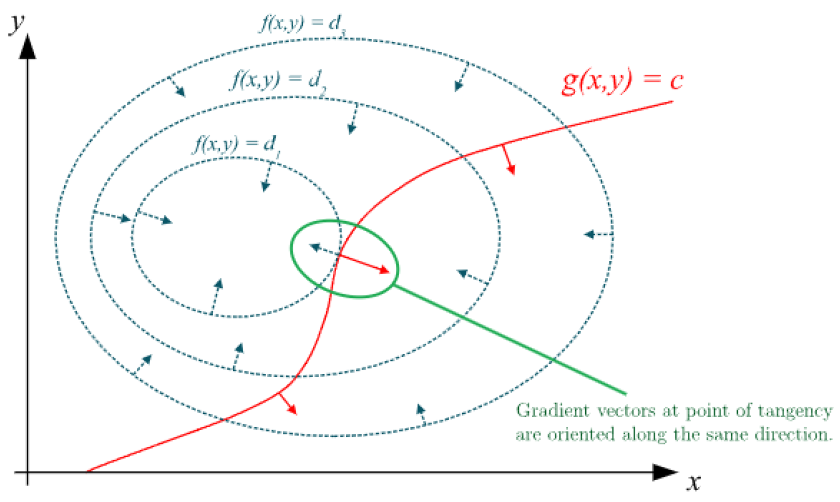
Where the gradient comes into play

How do you put the idea of two contour lines being tangent into a formula you can solve?

To answer this, we turn to our loyal friend the gradient. There are many ways to interpret ∇f: The direction of steepest ascent, a tool for computing directional derivatives, etc. But for our purposes here, the property we care about is that **the gradient of f evaluated at a point (x_0, y_0) always gives a vector perpendicular to the contour line passing through that point.**



This means when the contour lines of two functions f and g are tangent, their gradient vectors are parallel. Here's what that might look like for arbitrary functions f and g:



The fact that contour lines are tangent tells us nothing about the magnitude of each of these gradient vectors, but that's okay. When two vectors point in the same direction, it means we can multiply one by some constant to get the other. Specifically, let (x_0, y_0) represent a particular point where the contour lines of f and g are tangent (writing x_0 and y_0 with a 0 subscripts just indicates that we are considering constant values, and hence a specific point). Since this tangency means their gradient vectors align, here's what you might write down:

$$\nabla f(x_0, y_0) = \lambda_0 \nabla g(x_0, y_0)$$

Here, λ_0 represents some constant. Some authors use a negative constant, $-\lambda_0$, but I personally prefer a positive constant, as it gives a cleaner interpretation of λ_0 down the road.

Let's see what this looks like in our example where $f(x, y) = 2x + y$ and $g(x, y) = x^2 + y^2$. The gradient of f is

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial}{\partial x}(2x + y) \\ \\ \frac{\partial}{\partial y}(2x + y) \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and the gradient of g is

$$\nabla g(x, y) = \begin{bmatrix} \frac{\partial}{\partial x}(x^2 + y^2 - 1) \\ \\ \frac{\partial}{\partial y}(x^2 + y^2 - 1) \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

Therefore, the tangency condition ends up looking like this:

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \lambda_0 \begin{bmatrix} 2x_0 \\ 2y_0 \end{bmatrix}$$

Solving the problem in the specific case

To sum up where we are so far, we are looking for input points (x_0, y_0) with the following properties:

- $g(x_0, y_0) = 1$, which for our example means

  $$x_0^2 + y_0^2 = 1$$

- $\nabla f(x_0, y_0) = \lambda_0 \nabla g(x_0, y_0)$ for some constant λ_0, which for our example means

  $$\begin{aligned} 2 &= 2\lambda_0 x_0 \\ 1 &= 2\lambda_0 y_0 \end{aligned}$$

There are 3 equations and 3 unknowns, so this is a perfectly solvable situation.

[Hide explanation]

The approach will be to first solve for λ_0, then use the solution to find x_0 and y_0.

Using the last two equations above, write x_0 and y_0 in terms of λ_0.

$$2 = 2\lambda_0 x_0 \Rightarrow x_0 = \frac{1}{\lambda_0}$$

$$1 = 2\lambda_0 y_0 \Rightarrow y_0 = \frac{1}{2\lambda_0}$$

Now to bring in the third equation, plug these results into the equation $x_0^2 + y_0^2 = 1$.

$$\begin{aligned} {x_0}^2 + {y_0}^2 &= 1 \\ \left(\frac{1}{\lambda_0}\right)^2 + \left(\frac{1}{2\lambda_0}\right)^2 &= 1 \\ \frac{1}{\lambda_0^2} + \frac{1}{4\lambda_0^2} &= 1 \end{aligned}$$

To get λ_0 out of the denominators, we multiply everything by $4\lambda_0^2$ and simplify.

$$\begin{aligned} 4 + 1 &= 4\lambda_0^2 \\ \frac{5}{4} &= \lambda_0^2 \end{aligned}$$

$$\pm\sqrt{\frac{5}{4}} = \lambda_0$$

$$\frac{\pm\sqrt{5}}{2} = \lambda_0$$

Using the expressions for x_0 and y_0 in terms of λ_0 that we found above, these two solutions correspond with the pairs

$$\begin{aligned} (x_0, y_0) &= \left(\frac{1}{\lambda_0}, \frac{1}{2\lambda_0}\right) \\ &= \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right) \quad \text{or} \quad \left(\frac{-2}{\sqrt{5}}, \frac{-1}{\sqrt{5}}\right) \end{aligned}$$

We can see which of these is a maximum point and which is a minimum point by plugging these solutions into $f(x, y)$ and seeing which is bigger.

$$f\left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}\right) = 2\frac{2}{\sqrt{5}} + \frac{1}{\sqrt{5}}$$

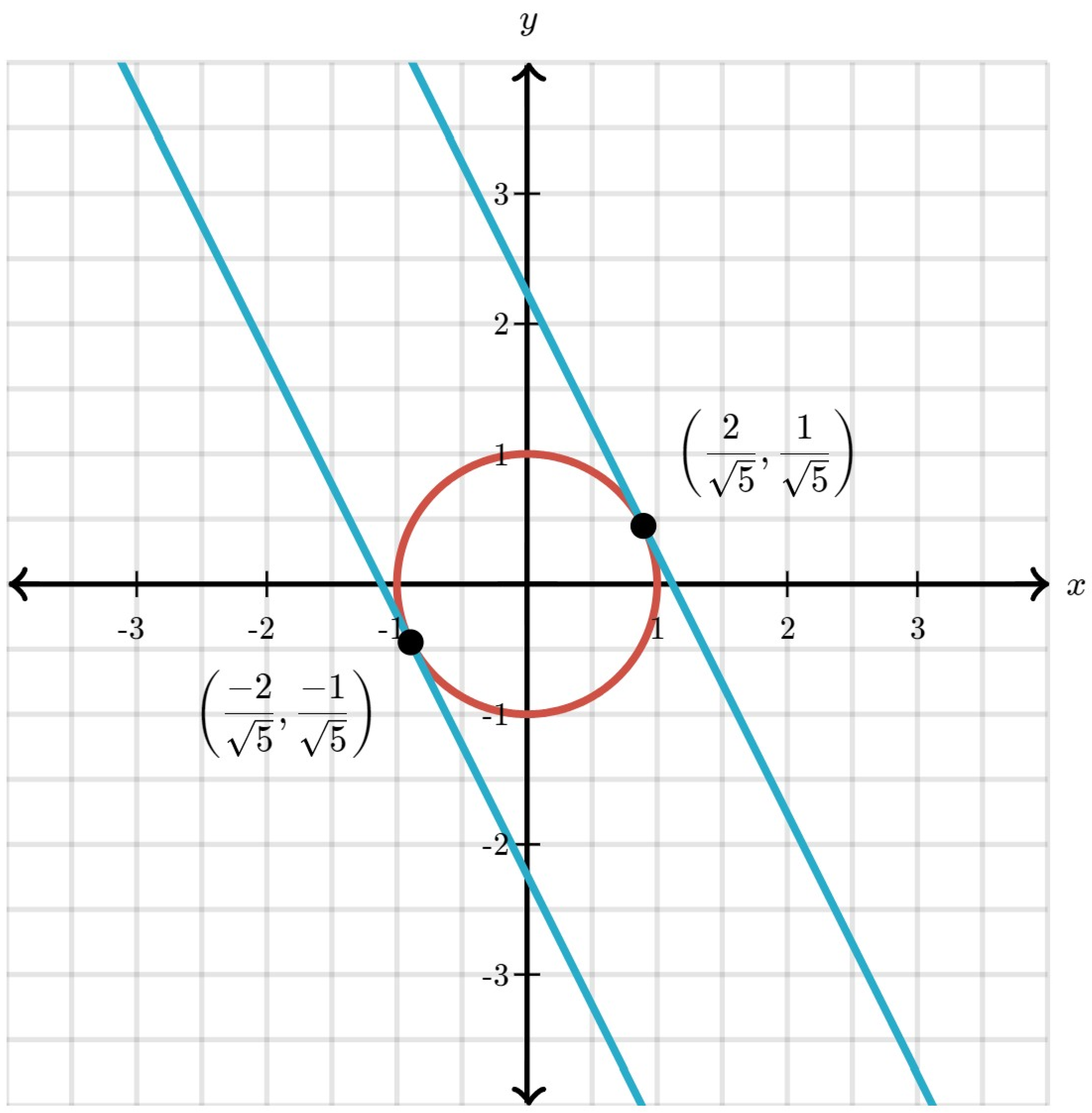
$$= \frac{5}{\sqrt{5}}$$

$$= \sqrt{5} \quad \leftarrow \text{Maximum}$$

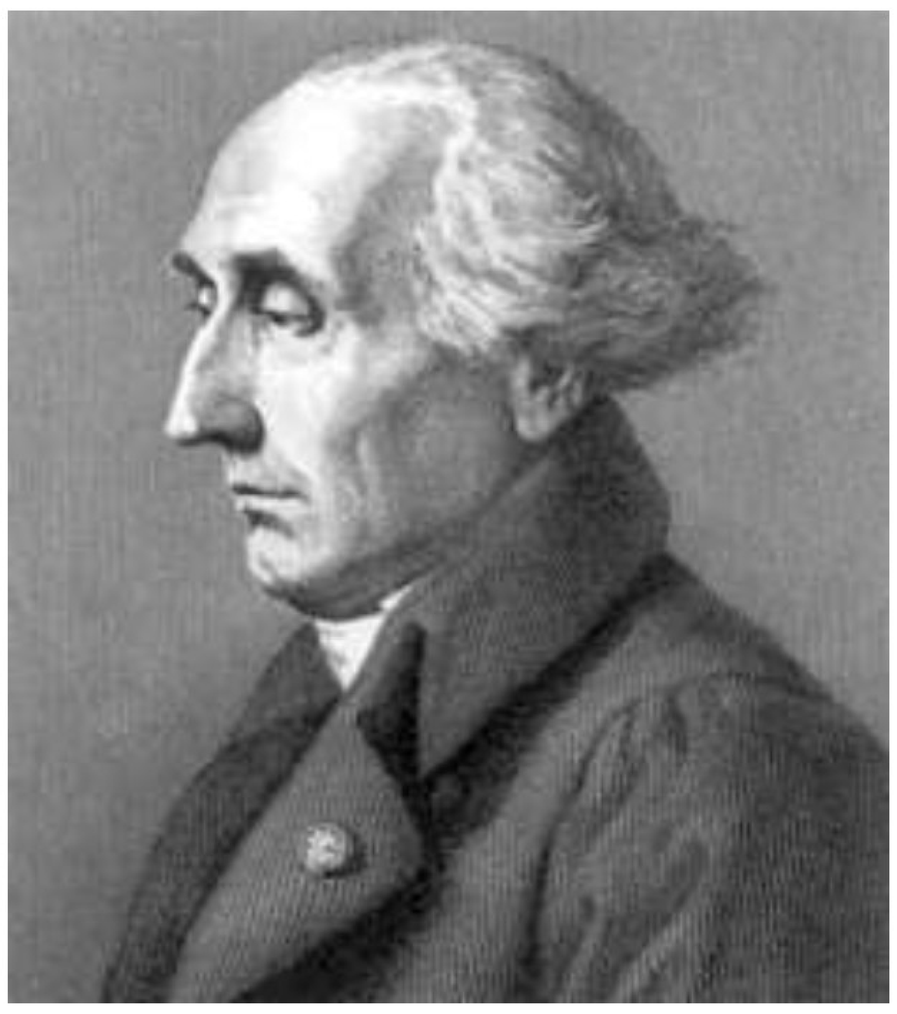
$$f\left(-\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}}\right) = 2\frac{-2}{\sqrt{5}} + \frac{-1}{\sqrt{5}}$$

$$= \frac{-5}{\sqrt{5}}$$

$$= -\sqrt{5} \quad \leftarrow \text{Minimum}$$



The Lagrangian function



Joseph Louis Lagrange, looking peaceful, content, and sleepy, all at the same time. [Wikimedia Commons]

In the 1700's, our buddy Joseph Louis Lagrange studied constrained optimization problems of this kind, and he found a clever way to express all of our conditions into a single equation.

You can write these conditions generally by saying we are looking for constants x_0, y_0 and λ_0 that satisfy the following conditions:

- **The constraint**:

$$g(x_0, y_0) = c$$

- **The tangency condition**:

$$\nabla f(x_0, y_0) = \lambda_0 \nabla g(x_0, y_0).$$

This can be broken into its components as follows:

- $f_x(x_0, y_0) = \lambda_0 g_x(x_0, y_0)$

- $f_y(x_0, y_0) = \lambda_0 g_y(x_0, y_0)$

Lagrange wrote down a special new function which takes in all the same input variables as f and g, along with the new kid in town λ, thought of now as a variable rather than a constant.

$$\mathcal{L}(x, y, \lambda) = f(x, y) - \lambda(g(x, y) - c)$$

For example, consider our example above.

$$\begin{aligned} f(x, y) &= 2x + y \\ g(x, y) &= x^2 + y^2 \\ c &= 1 \end{aligned}$$

Here's how this new function would look:

$$\mathcal{L}(x, y, \lambda) = 2x + y - \lambda(x^2 + y^2 - 1).$$

Notice, the partial derivative of $\mathcal{L}$ with respect to λ is $-(g(x, y) - c)$:

$$\begin{aligned} \mathcal{L}_\lambda(x, y, \lambda) &= \frac{\partial}{\partial \lambda}(f(x, y) - \lambda(g(x, y) - c) \\ &= 0 - (g(x, y) - c) \end{aligned}$$

So we can translate the condition $g(x, y) = c$ as

$$\mathcal{L}_\lambda(x, y, \lambda) = -g(x, y) + c = 0$$

What's more, look at what we get when we set one of the other partial derivatives equal to 0:

$$\mathcal{L}_x(x, y, \lambda) = 0$$

$$\frac{\partial}{\partial x}(f(x, y) - \lambda(g(x, y) - c)) = 0$$

$$f_x(x, y) - \lambda g_x(x, y) = 0$$

$$f_x(x, y) = \lambda g_x(x, y)$$

That just so happens to be another one of our conditions! Almost identically, the condition $\mathcal{L}_y(x, y, \lambda) = 0$ unravels to become

$$f_y(x, y) = \lambda g_y(x, y)$$

Together, these conditions are the same as saying.

$$\nabla f(x, y) = \lambda \nabla g(x, y)$$

Therefore, the three conditions we need to solve to find x, y and λ come down to the various partial derivatives of $\mathcal{L}$ being equal to 0. This can be written extremely compactly by setting the gradient of $\mathcal{L}$ equal to the zero vector:

$$\nabla \mathcal{L} = \mathbf{0}$$

For example, using our specific functions from above, we see how this encodes the system of equations we need to solve:

$$\nabla \mathcal{L} = \begin{bmatrix} \frac{\partial}{\partial x}(2x + y - \lambda(x^2 + y^2 - 1)) \\ \\ \frac{\partial}{\partial y}(2x + y - \lambda(x^2 + y^2 - 1)) \\ \\ \frac{\partial}{\partial \lambda}(2x + y - \lambda(x^2 + y^2 - 1)) \end{bmatrix} = \begin{bmatrix} 2 - 2\lambda x \\ 1 - 2\lambda y \\ -x^2 - y^2 + 1 \end{bmatrix} =$$

As a tribute to ol' Joey Lou, we call this function $\mathcal{L}$ the "**Lagrangian**", and the new variable λ that we introduce is called a "**Lagrange multiplier**". Imagine if someone added "-ian" the end of your last name and made it the name of a function everybody uses. Pretty sweet!

Warning: Some authors use a convention where the sign of λ is reversed:

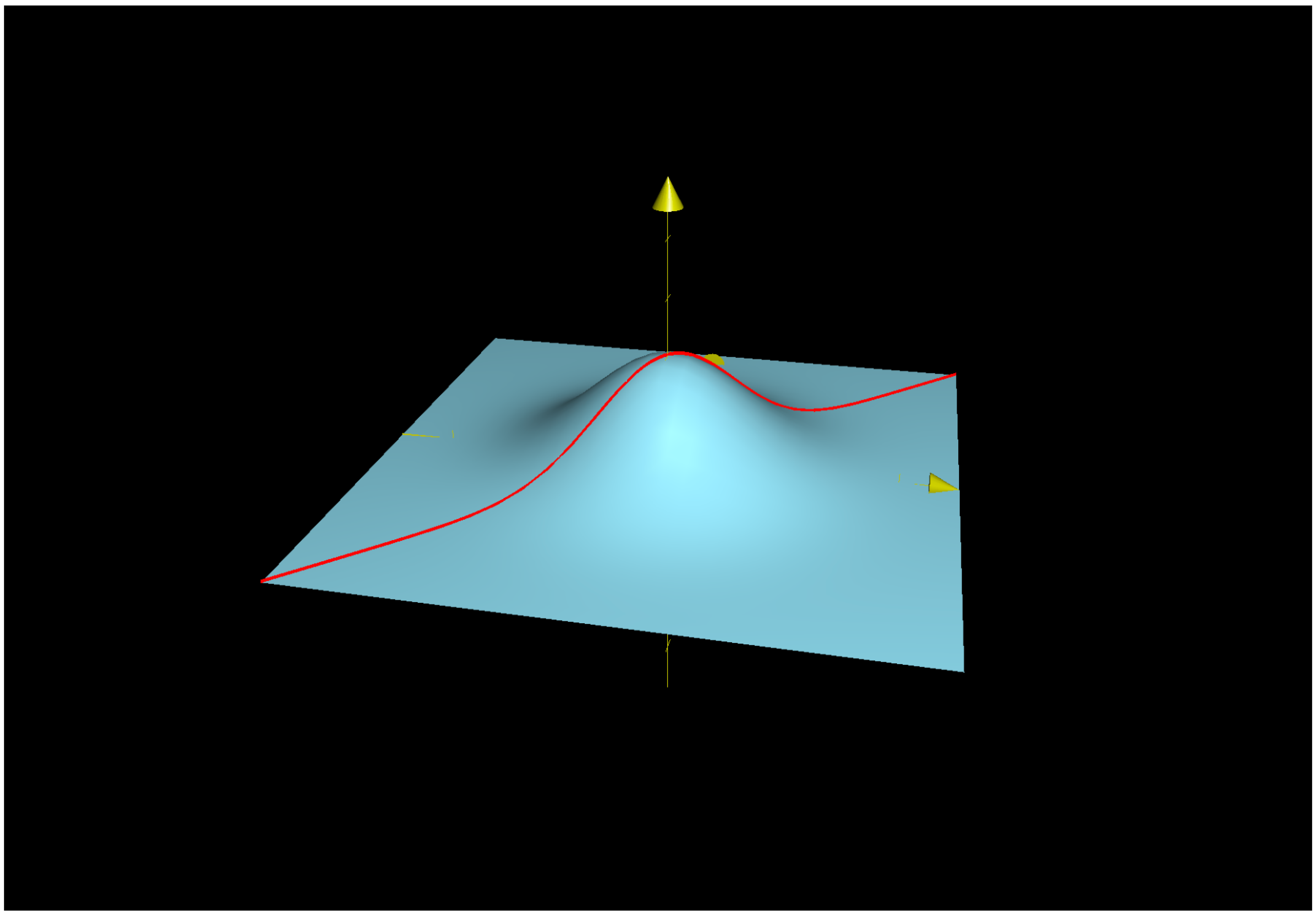
$$\mathcal{L}(x, y, \lambda) = f(x, y) + \lambda(g(x, y) - c)$$

This doesn't make any difference when it comes to solving the problem, but you should keep it in mind in case the course you are taking or the text you are reading follows this convention.

[Hide explanation]

Side note: What if the constraint isn't so constraining?

There's a slight twist to this story, best illustrated with an example.



Suppose we needed to maximize the function

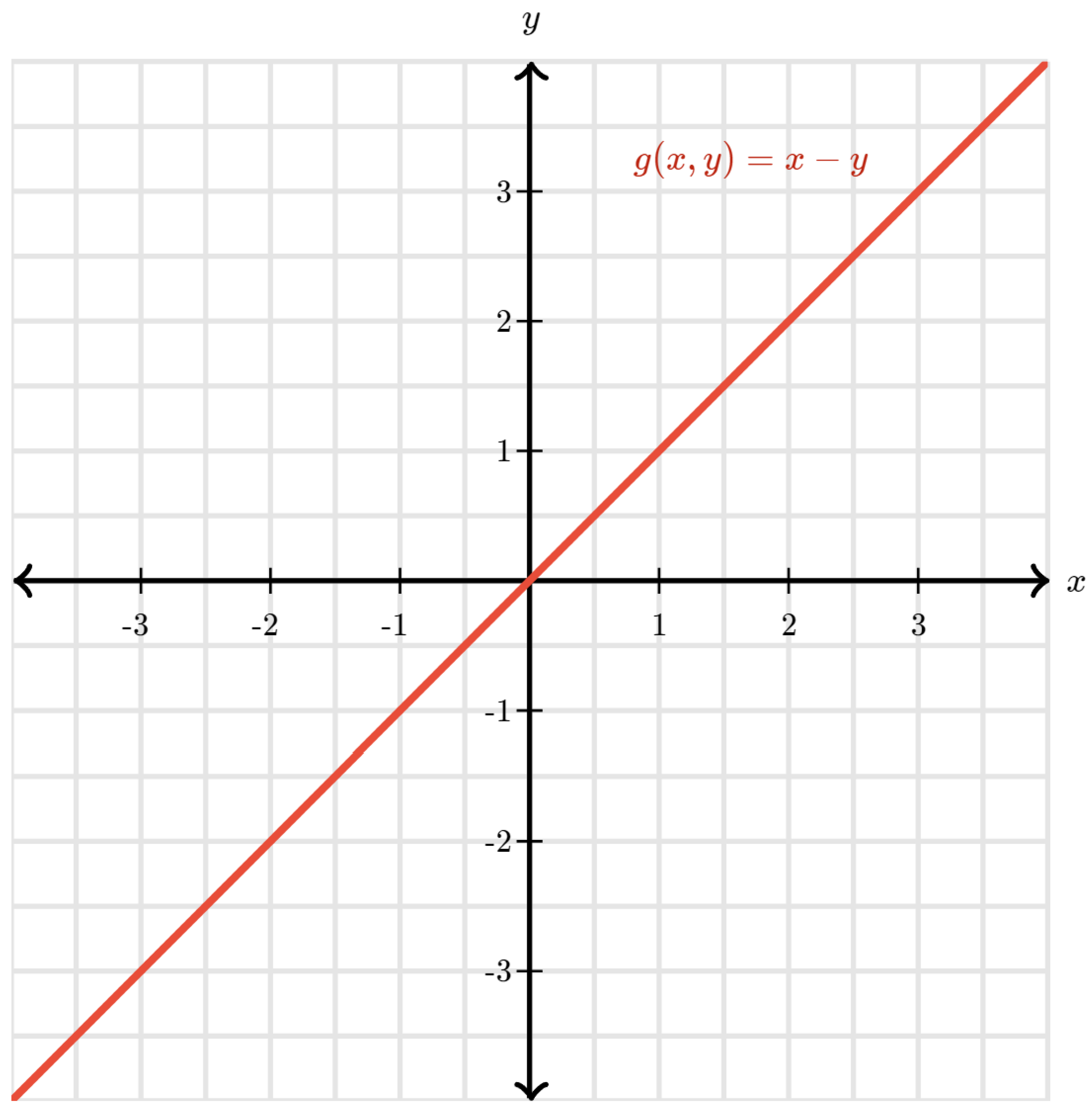
$$f(x, y) = e^{-(x^2+y^2)}$$

Subject to the constraint

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

The graph of $f(x, y) = e^{-(x^2+y^2)}$ is a "bell curve", with a round bump above the point $(x_0, y_0) = (0, 0)$, as pictured above.

This constraint can be pictured as a diagonal line in the xy-plane (shown in red).



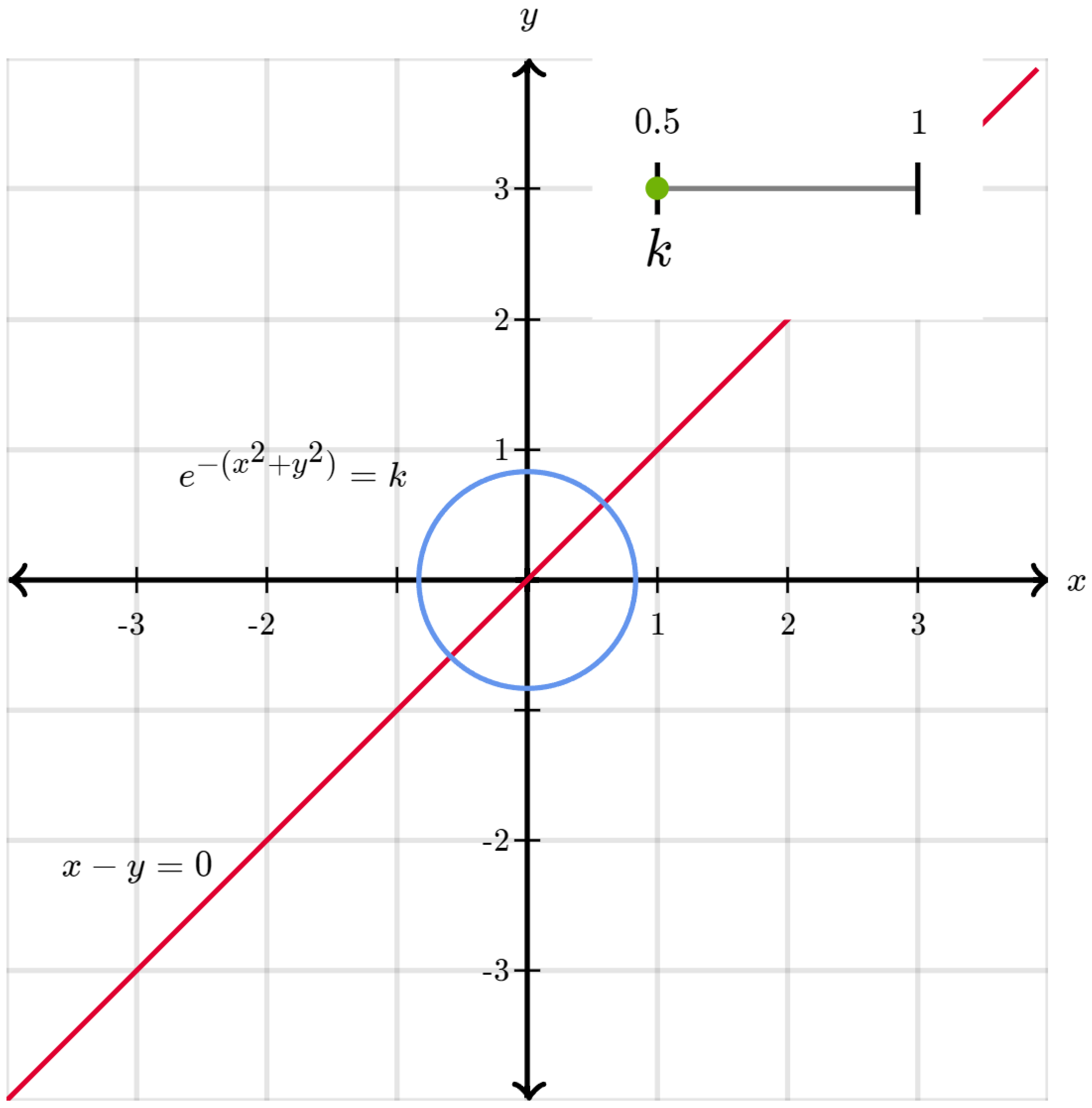
What makes this problem a bit silly is that the (unconstrained) maximum point $(0, 0)$ of f already satisfies the constraint $g(x, y) = 0$, since

$$g(0, 0) = 0 + 0 = 0$$

It might feel like this makes things easier on us. After all, not having to worry about the constraint should be more straightforward than worrying about it, right? However, if you (or more realistically a computer) were solving a given constrained optimization problem, it's not like you would first find the unconstrained maximum, check if it fits the constraint, then turn to the Lagrange multiplier technique. You would just start with the Lagrange multiplier approach, since it's almost never the case that the unconstrained maximum is also the constrained maximum.

It turns out that **the Lagrange multiplier technique still works when the constrained maximum is also an unconstrained maximum**. The reason is a bit subtle since our contour-tangency argument above doesn't quite apply.

For example, let's play around with the contours representing $e^{-(x^2+y^2)} = k$ as k varies between 0.5 and 1:



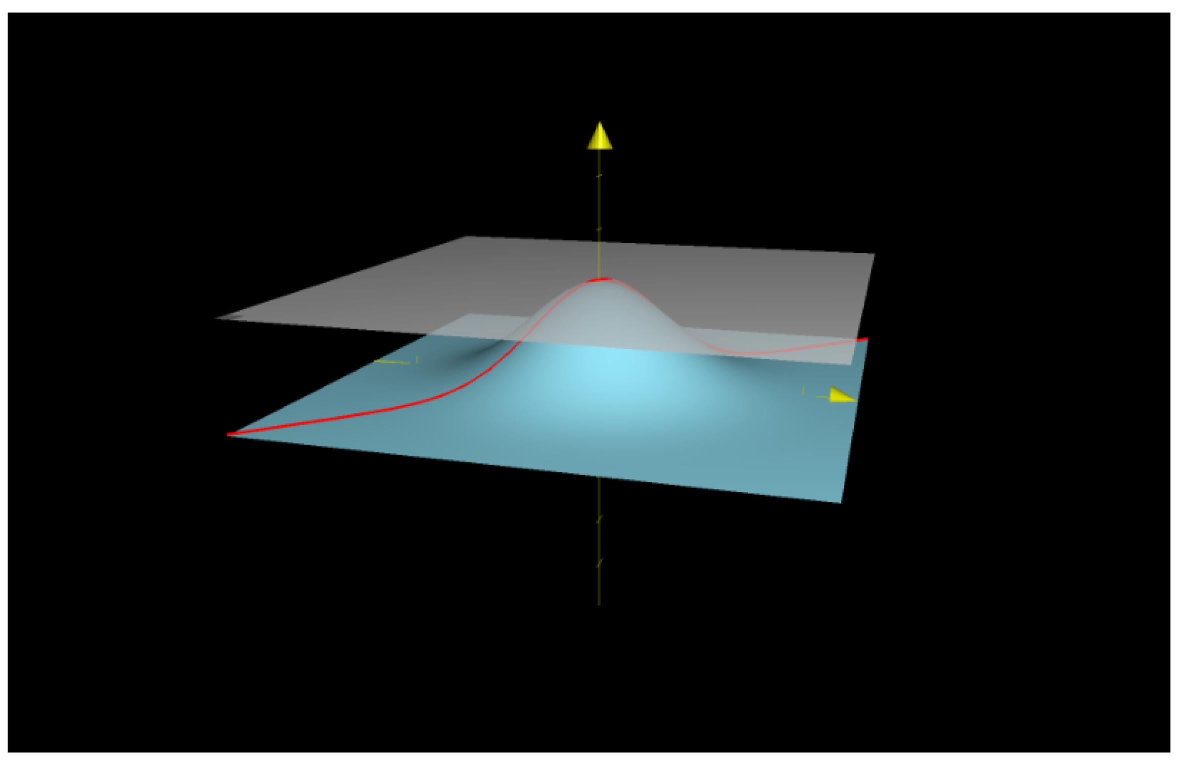
The contour representing $e^{-(x^2+y^2)} = k$ always crosses the line $x - y = 0$, and this contour shrinks to a point when k is as big as it can be. Previously, we said the contour lines would be tangent when f achieves its constrained maximum, but it feels weird to say a point is tangent to a line.

Why isn't this a problem then?

If f achieves an unconstrained local maximum at some point (x_0, y_0), its gradient at that point will be 0.

$$\nabla f(x_0, y_0) = \mathbf{0}$$

You can think of this as saying the tangent plane at a local maximum is flat.



Flat tangent plane of f at its maximum.

In this case, the property

$$\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0)$$

still holds, because we can set $\lambda = 0$.

A funny way to think about this is to say a point is "tangent" to all lines passing through it in the same sense that the zero vector is "proportional" to all other vectors.

> So not only does the Lagrange multiplier technique cleverly consolidate many conditions into the one equation
>
> $$\nabla \mathcal{L} = \mathbf{0},$$
>
> it also automatically takes into account this special edge case!

Summary

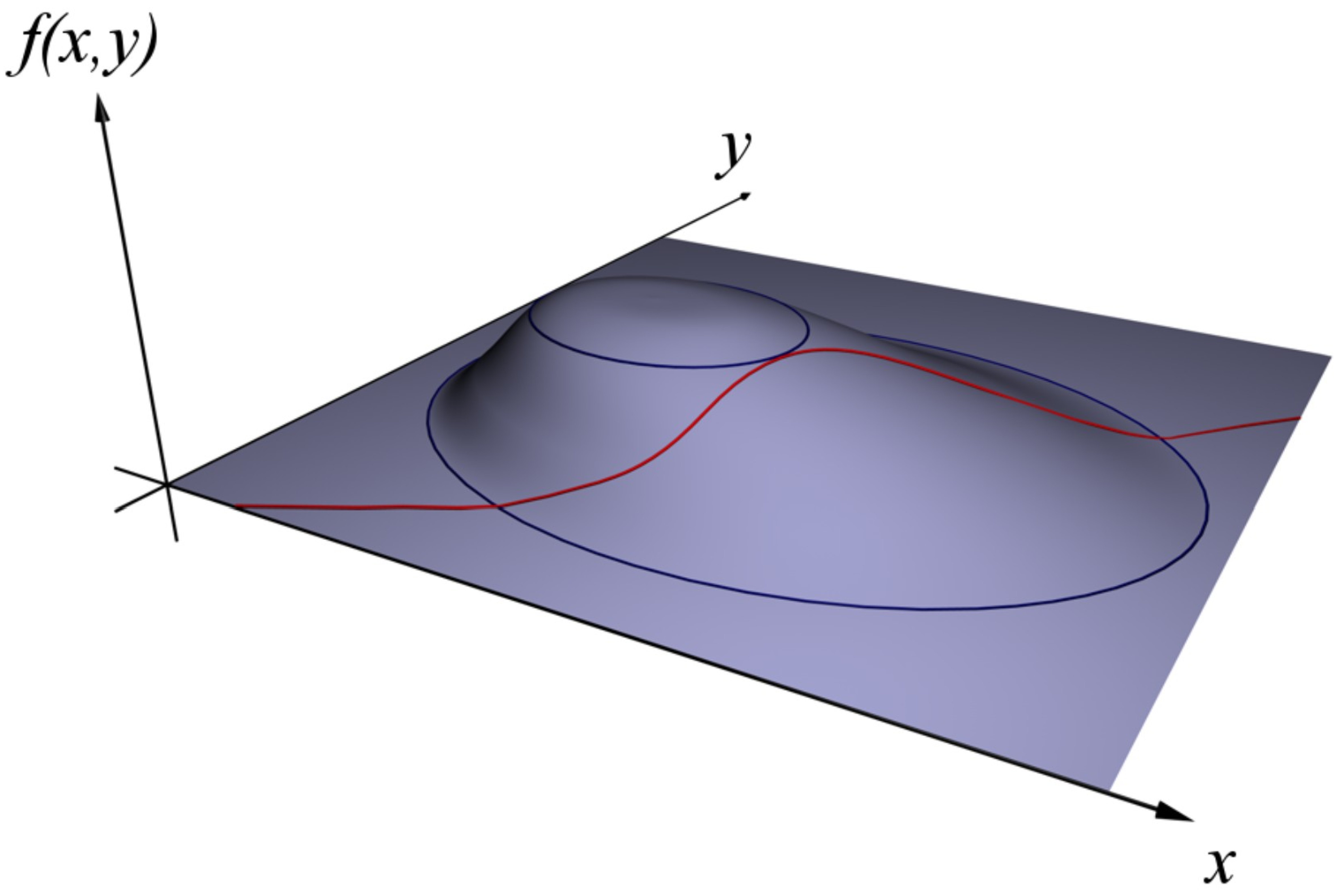


Image credit: By Nexcis (Own work) [Public domain], via Wikimedia Commons

When you want to maximize (or minimize) a multivariable function $f(x, y, \dots)$ subject to the constraint that another multivariable function equals a constant, $g(x, y, \dots) = c$, follow these steps:

- **Step 1**: Introduce a new variable λ, and define a new function $\mathcal{L}$ as follows:

  $$\mathcal{L}(x, y, \dots, \lambda) = f(x, y, \dots) - \lambda(g(x, y, \dots) - c)$$

  This function $\mathcal{L}$ is called the "Lagrangian", and the new variable λ is referred to as a "Lagrange multiplier"

- **Step 2**: Set the gradient of $\mathcal{L}$ equal to the zero vector.

  $$\nabla \mathcal{L}(x, y, \dots, \lambda) = \mathbf{0} \quad \leftarrow \text{Zero vector}$$

In other words, find the **critical points** of $\mathcal{L}$.

- **Step 3**: Consider each solution, which will look something like $(x_0, y_0, \ldots, \lambda_0)$. Plug each one into f. Or rather, first remove the λ_0 component, then plug it into f, since f does not have λ as an input. Whichever one gives the greatest (or smallest) value is the maximum (or minimum) point you are seeking.