Directional derivatives (going deeper)

Google Classroom

A more thorough look at the formula for directional derivatives, along with an explanation for why the gradient gives the slope of steepest ascent.

Background:

- Introduction to directional derivatives
- Partial derivatives
- Gradient

This article is targetted for those who want a deeper understanding of the directional derivative and its formula.

Formal definition of the directional derivative

There are a couple reasons you might care about a formal definition. For one thing, really understanding the formal definition of a new concept can make clear what it is really going on. But more importantly than that, I think the main benefit is that it gives you the confidence to recognize when such a concept can and cannot be applied.

As a warm up, let's review the formal definition of the partial derivative, say with respect to x:

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{h \to 0} \frac{f(x_0 + h, y_0) - f(x_0, y_0)}{h}$$

The connection between the informal way to read $\frac{\partial f}{\partial x}$ and the formal way to read the right-hand side is as follows:

Symbol	Informal understanding	Formal understanding
∂x	A tiny nudge in the x direction.	A limiting variable h which goes to 0, and will be added to the first component of the function's input.
∂f	The resulting change in the output of f after the nudge.	The difference between $f(x_0 + h, y_0)$ and $f(x_0, y_0)$, taken in the same limit as $h \to 0$.

We could instead write this in vector notation, viewing the input point (x_0, y_0) as a two-dimensional vector

$$\mathbf{x}_0 = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix}$$

Here $\mathbf{x}_0$ is written in bold to emphasize its vectoriness. It's a bit confusing to use a bold $\mathbf{x}$ for the entire input rather than some other letter, since the letter x is already used in an un-bolded form to denote the first component of the input. But hey, that's convention, so we go with it.

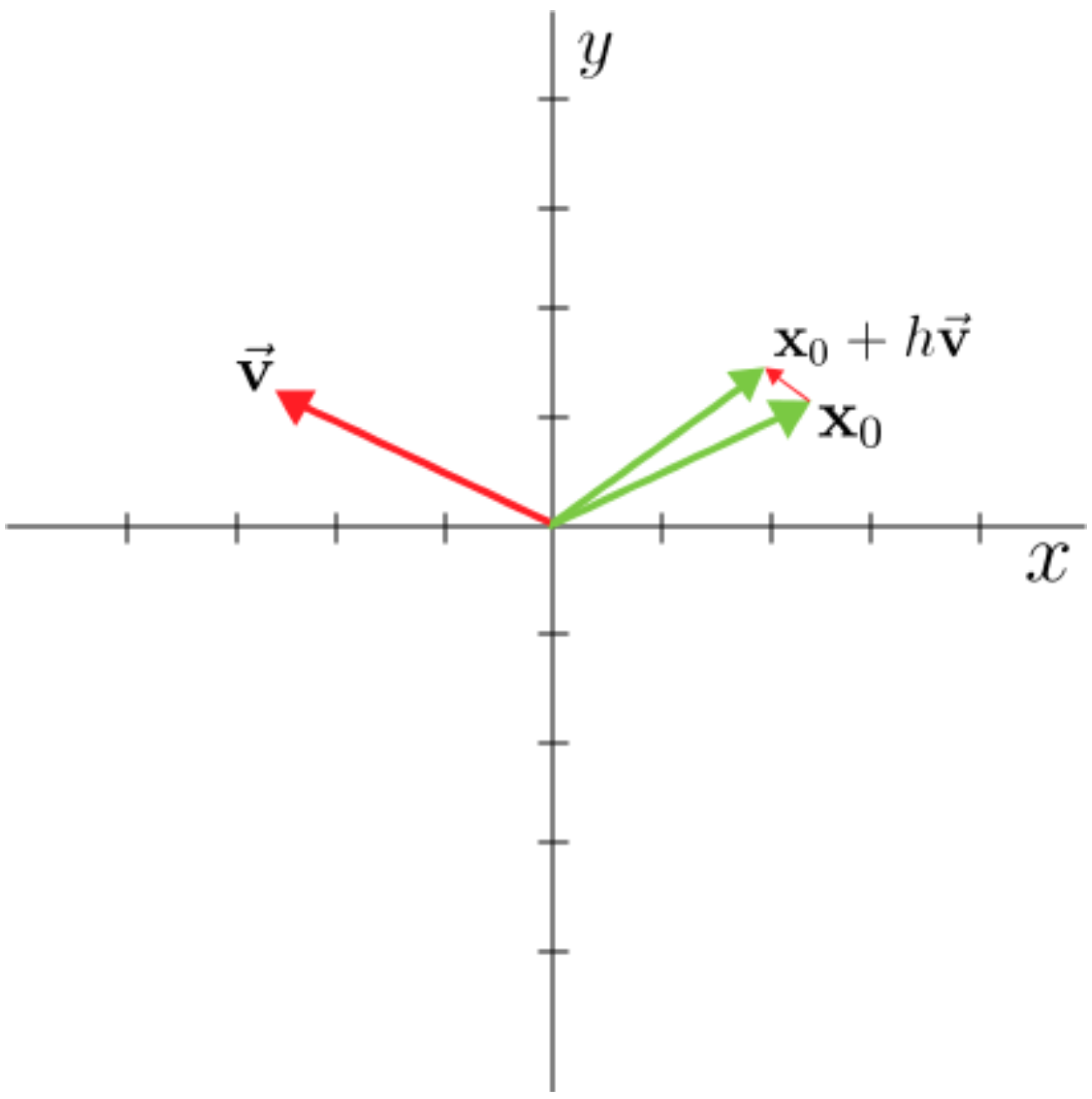
Instead of writing the "nudged" input as $(x_0 + h, y_0)$, we write it as $\mathbf{x}_0 + h\hat{\mathbf{i}}$, where $\hat{\mathbf{i}}$ is the unit vector in the x-direction:

$$\frac{\partial f}{\partial x}(\mathbf{x}_0) = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\hat{\mathbf{i}}) - f(\mathbf{x}_0)}{h}$$

In this notation, it's much easier to see how to generalize the partial derivative with respect to x to the directional derivative along any vector $\vec{\mathbf{v}}$:

$$\nabla_{\vec{\mathbf{v}}} f(\mathbf{x}_0) = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\vec{\mathbf{v}}) - f(\mathbf{x}_0)}{h}$$

In this case, adding $h\vec{\mathbf{v}}$ to the input for a limiting variable $h \to 0$ formalizes the idea of a tiny nudge in the direction of $\vec{\mathbf{v}}$.



Seeking connection between the definition and computation

Computing the directional derivative involves a dot product between the gradient ∇f and the vector $\vec{\mathbf{v}}$. For example, in two dimensions, here's what this would look like:

$$\nabla_{\vec{\mathbf{v}}} f(x, y) = \nabla f \cdot \vec{\mathbf{v}}$$

$$= \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \cdot \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$= v_1 \frac{\partial f}{\partial x}(x, y) + v_2 \frac{\partial f}{\partial y}(x, y)$$

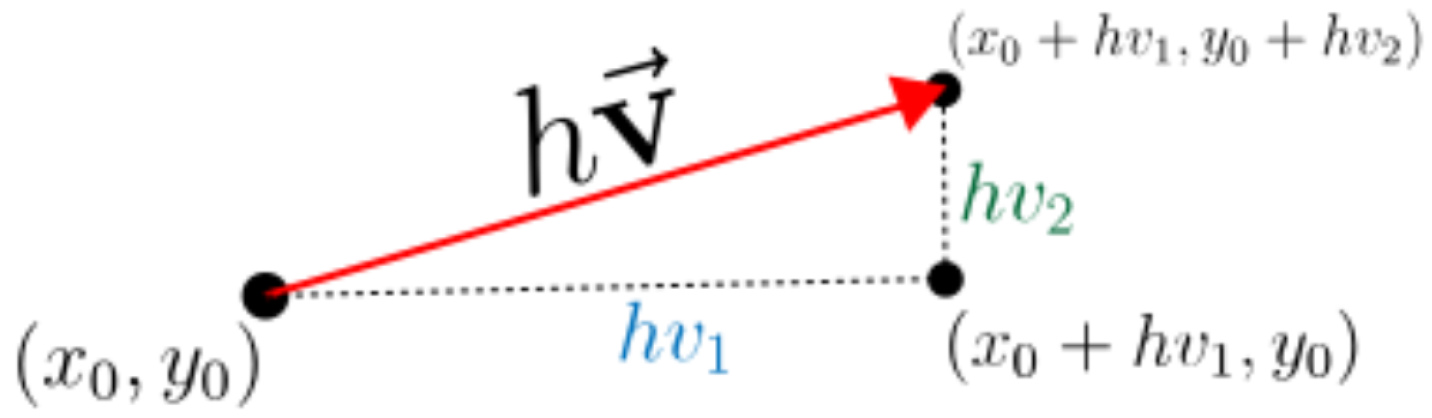
Here, v_1 and v_2 are the components of $\vec{\mathbf{v}}$.

$$\vec{\mathbf{v}} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

The central question is, what does this formula have to do with the definition given above?

Breaking down the nudge

The computation for $\nabla_{\mathbf{v}} f$ can be seen as a way to break down a tiny step in the direction of $\mathbf{v}$ into its x and y components.



Specifically, you can imagine the following procedure:

1. Start at some point (x_0, y_0).

2. Choose a tiny value h.

3. Add hv_1 to x_0, which means stepping to the point $(x_0 + hv_1, y_0)$. From what we know of partial derivatives, this will change the output of the function by about

$$hv_1 \left(\frac{\partial f}{\partial x}(x_0, y_0) \right)$$

- Now add hv_2 to y_0 to bring us up/down to the point $(x_0 + hv_1, y_0 + hv_2)$. The resulting change to f is now about

$$hv_2 \left(\frac{\partial f}{\partial y}(x_0 + hv_1, y_0) \right)$$

Adding the results of steps 3 and 4, the total change to the function upon moving from the input (x_0, y_0) to the input $(x_0 + hv_1, y_0 + hv_2)$ has been about

$$hv_1 \left(\frac{\partial f}{\partial x}(x_0, y_0) \right) + hv_2 \left(\frac{\partial f}{\partial y}(x_0 + hv_1, y_0) \right)$$

This is very close to the expression for the directional derivative, which says the change in f due to this step $h\vec{\mathbf{v}}$ should be about

$$h\nabla_{\vec{\mathbf{v}}} f(x_0, y_0)$$

$$= h\vec{\mathbf{v}} \cdot \nabla f(x_0, y_0)$$

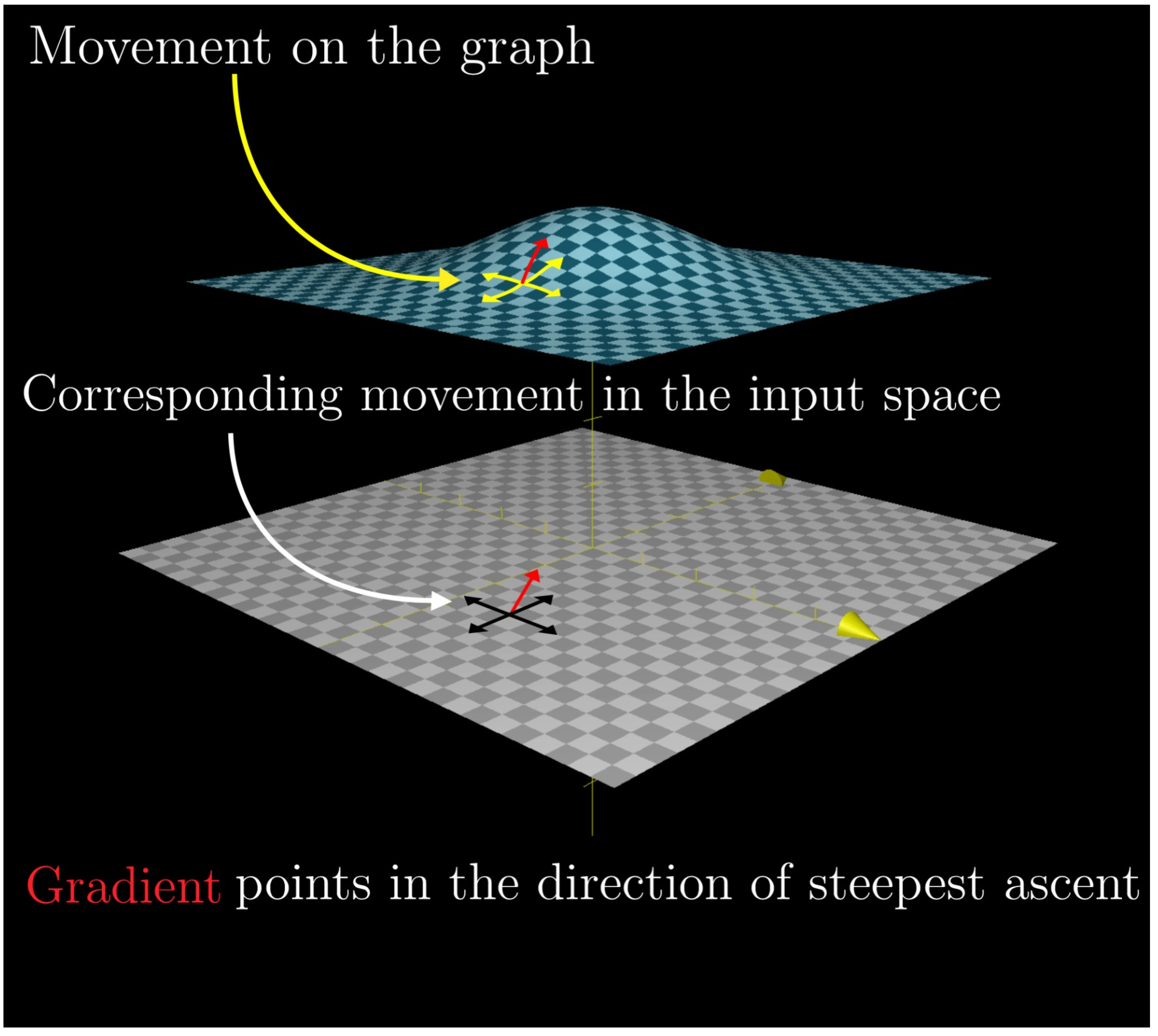
$$= hv_1 \frac{\partial f}{\partial x}(x_0, y_0) + hv_2 \frac{\partial f}{\partial y}(x_0, y_0)$$

However, this differs slightly from the result of our step-by-step argument, in which the partial derivative with respect to y is taken at the point $(x_0 + hv_1, y_0)$, not at the point (x_0, y_0).

Luckily we are considering very, very small values of h. In fact, more technically, we should be talking about the limit as $h \to 0$. Therefore evaluating $\frac{\partial f}{\partial y}$ at $(x_0 + hv_1, y_0)$ will be almost the same as evaluating it at (x_0, y_0). Moreover, as h approaches 0, so does the difference between these two, but we have to assume that f is continuous.

Why does the gradient point in the direction of steepest ascent?

Having learned about the directional derivatives, we can now understand *why* the direction of the gradient is the direction of steepest ascent.



Specifically, here's the question at hand.

Setup:

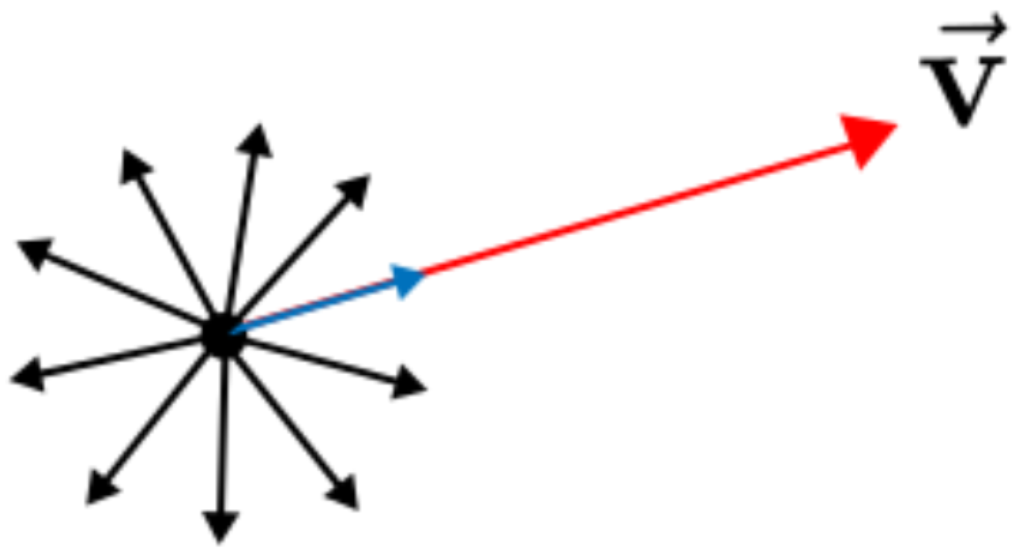
- Let f be some scalar-valued multivariable function, such as $f(x, y) = x^2 + y^2$.
- Let (x_0, y_0) be a particular input point
- Consider all possible *directions*, i.e. all unit vectors $\hat{\mathbf{u}}$ in the input space of f.

Question (informal): If we start at (x_0, y_0), which direction should we walk so that the output of f increases most quickly?

Question (formal): Which unit vector $\hat{\mathbf{u}}$ maximizes the directional derivative along $\hat{\mathbf{u}}$?

$$\nabla_{\hat{\mathbf{u}}} f(x_0, y_0) = \underbrace{\hat{\mathbf{u}} \cdot \nabla f(x_0, y_0)}_{\text{Maximize this quantity}}$$

The famous triangle inequality tells us that this will be maximized by the unit vector in the direction $\nabla f(x_0, y_0)$.



The unit vector $\hat{\mathbf{u}}$ maximizing $\hat{\mathbf{u}} \cdot \vec{\mathbf{v}}$ is the one in the same direction as $\vec{\mathbf{v}}$

Notice, the fact that the gradient points in the direction of steepest ascent is a consequence of the more fundamental fact that all directional derivatives require taking the dot product with ∇f.

[Hide explanation]

> Okay, this is kind of a long non-sequitur, but I think it helps shed some light on the fact that a dot product is maximized with both vectors point in the same direction. It's also a fun brain teaser in its own right.
>
> Suppose I present to you a stack of $\$100$ bills, a stack of $\$20$ bills, and a stack of $\$2$ bills. I explain that you should choose three different numbers A, B and C that satisfy the equation
>
> $$A^2 + B^2 + C^2 = 10{,}404$$
>
> And I will let you take A bills from the stack of $\$100$'s, B bills from the stack of $\$20$'s and C bills from the stack of $\$2$'s.
>
> So the natural question is, what choice of A, B and C maximizes the amount of money you get, and what is that maximum value?

[Hide explanation]

It might feel like you should take as many $\$100$ bills as possible. That is, let $B = C = 0$, and let

$$A = \sqrt{10{,}404} = 102$$

The problem here is that each additional $\$100$ bill increases the sum $A^2 + B^2 + C^2$ more and more as A gets larger, so you get less and less "bang for your buck", so to speak. For example, if you have already grabbed one hundred $\$100$ bills, adding another one increases this sum by 201:

$$\begin{aligned} 101^2 - 100^2 &= 10{,}201 - 10{,}000 \\ &= 201 \end{aligned}$$

So you would have been better off grabbing 5 twenty dollar bills, increasing your sum by only 25:

$$5^2 - 0^2 = 25$$

In fact, if you grab twenty $\$20$ bills, the sum of squares is still less than the constraint,

$$\begin{aligned} A^2 + B^2 + C^2 &= 100^2 + 20^2 + 0^2 \\ &= 10{,}400 \\ &< 10{,}404 \end{aligned}$$

and you have $\$10{,}400$, which is better than the $102 \times \$100 = \$10{,}200$ you got by greedily grabbing as many $\$100$ bills as you could.

In fact, the best possible choice is to take one hundred $\$100$ bills, twenty $\$20$ bills, and two $\$2$ bills.

I encourage you to try proving this for yourself. As a hint, think of $A^2 + B^2 + C^2$ as the number of "points" you have, and think about maximizing the total "dollars per point" with each new choice.

This problem is analogous to finding the steepest ascent, since you are maximizing a dot product,

$$\begin{bmatrix} 100 \\ 20 \\ 2 \end{bmatrix} \cdot \begin{bmatrix} A \\ B \\ C \end{bmatrix} = 100A + 20B + 2C$$

subject to a constraint on the norm $A^2 + B^2 + C^2$. Likewise, in finding the steepest ascent, you are maximizing a dot product,

$$\nabla f \cdot \hat{\mathbf{u}} = u_1 \frac{\partial f}{\partial x} + u_2 \frac{\partial f}{\partial y} + u_3 \frac{\partial f}{\partial z}$$

subject to the constraint that $u_1^2 + u_2^2 + u_3^2 = 1$. In the money example, the best choice is one hundred $100's, twenty $20's, and two $2's. Similarly, to maximize $\nabla f \cdot \hat{\mathbf{u}}$,

- u_1 should be proportional to $\frac{\partial f}{\partial x}$,
- u_2 should be proportional to $\frac{\partial f}{\partial y}$,
- u_3 should be proportional to $\frac{\partial f}{\partial z}$.