Multivariable Chain Rule, directional derivative, and gradient

What we're building to

- Given a multivariable function $f(x, y)$, and two single variable functions $x(t)$ and $y(t)$, here's what the multivariable chain rule says:

$$\underbrace{\frac{d}{dt} f(x(t), y(t))}_{\text{Derivative of composition function}} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

- Written with vector notation, where $\vec{\mathbf{v}}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$, this rule has a very elegant form in terms of the gradient of f and the vector-derivative of $\vec{\mathbf{v}}(t)$.

$$\underbrace{\frac{d}{dt} f(\vec{\mathbf{v}}(t))}_{\text{Derivative of composition function}} = \overbrace{\nabla f \cdot \vec{\mathbf{v}}'(t)}^{\text{Dot product of vectors}}$$

A more general chain rule

As you can probably imagine, the multivariable chain rule generalizes the chain rule from single variable calculus. The single variable chain rule tells you how to take the derivative of the composition of two functions:

$$\frac{d}{dt} f(g(t)) = \frac{df}{dg}\frac{dg}{dt} = f'(g(t))g'(t)$$

What if instead of taking in a one-dimensional input, t, the function f took in a two-dimensional input, (x, y)?

$$f(x, y) = \dots \text{some expression of } x \text{ and } y \dots$$

Well, in that case, it wouldn't make sense to compose it with a scalar-valued function $g(t)$. Instead, let's say there are two separate scalar-valued functions $x(t)$ and $y(t)$, and we plug these in as the coordinates of f. The overall composition will be a single variable function, with a single-number input t, and a single-number output $f(x(t), y(t))$, as shown in this diagram:

$$\begin{array}{rccc} \text{One final output} & & f(x(t), y(t)) & \\ & \nearrow & & \nwarrow \\ \text{Two intermediate outputs} & x(t) & & y(t) \\ & \nwarrow & & \nearrow \\ \text{One input} & & t & \end{array}$$

$$\underbrace{\frac{d}{dt}}_{\substack{\text{This is an ordinary derivative,} \\ \text{not a partial derivative } \frac{\partial}{\partial t} \\ \text{because the total composition} \\ \text{has one input and one output.}}} f(x(t), y(t)) = \underbrace{\overbrace{\frac{\partial f}{\partial x}}^{\substack{\text{How } f \text{ changes} \\ \text{due to a tiny} \\ \text{change in } x}} \overbrace{\frac{dx}{dt}}^{\substack{\text{How } x \text{ changes} \\ \text{due to a tiny} \\ \text{change in } t}}}_{\substack{\text{Total change in } f \\ \text{due to the influence} \\ t \text{ has on } x}} +$$

$$\underbrace{\frac{\partial f}{\partial y} \frac{dy}{dt}}_{\substack{\text{Total change in } f \\ \text{due to the influence} \\ t \text{ has on } y}}$$

Keep in mind, an expression like $\frac{\partial f}{\partial x} \frac{dx}{dt}$ is shorthand for

$$\frac{\partial f}{\partial x}(x(t), y(t)) \frac{dx}{dt}(t)$$

That is, both are functions of t, but $\frac{\partial f}{\partial x}$ is evaluated via the intermediate functions $x(t)$ and $y(t)$.

Summary

- Given a multivariable function $f(x, y)$, and two single variable functions $x(t)$ and $y(t)$, here's what the multivariable chain rule says:

$$\underbrace{\frac{d}{dt} f(x(t), y(t))}_{\text{Derivative of composition function}} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

- Written with vector notation, where $\vec{\mathbf{v}}(t) = \begin{bmatrix} x(t) \\ y(t) \end{bmatrix}$, this rule has a very elegant form in terms of the gradient of f and the vector-derivative of $\vec{\mathbf{v}}(t)$.

$$\underbrace{\frac{d}{dt} f(\vec{\mathbf{v}}(t))}_{\text{Derivative of composition function}} = \overbrace{\nabla f \cdot \vec{\mathbf{v}}'(t)}^{\text{Dot product of vectors}}$$

Application of multivariable chain rule:

1. find the derivative of f(x) =$(x-1)^2 * (x+5)^2$
  Let u= (x-1), v= (x+5), respectively;

f(x) = f(u,v)= u^2*v^2

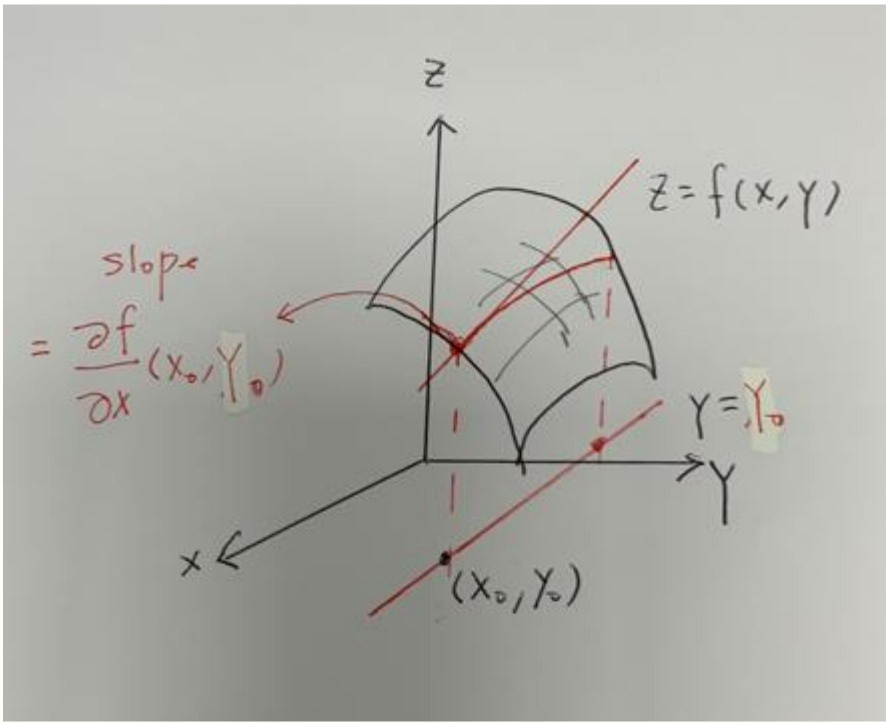
f'(x) = $2uv^2 + 2u^2 v=2(x-1)*(x+5)^2 + 2(x-1)^2 * (x+5)$

formula $(f/g)' = (f'g – fg')/g^2$

Directional derivative and gradient:

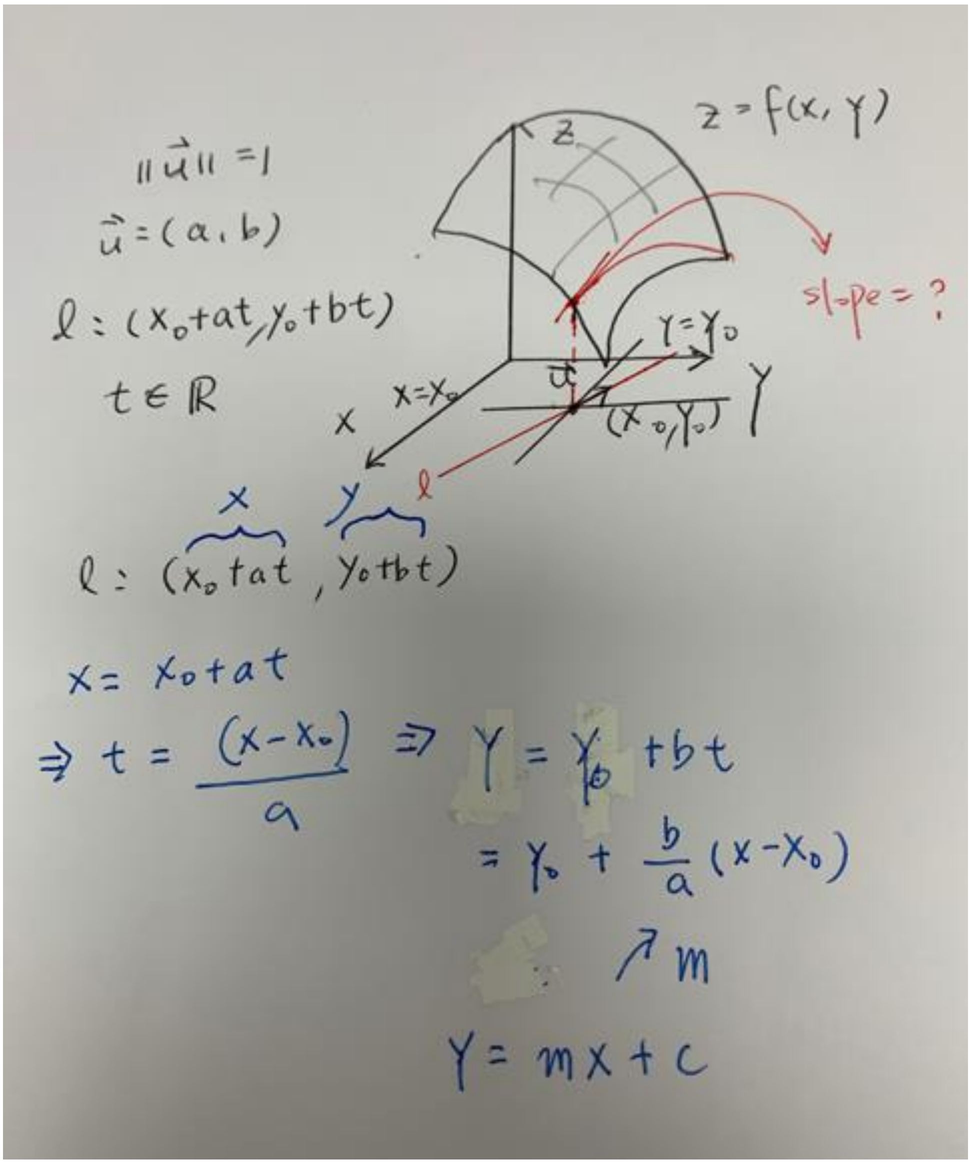
Gradient 2

z is a function of x, y. The derivative of z at point (x0, y0) along line y=y0 is the slope of the tangent line on its surface. And it is the rate of change of function z at (x0, y0) along line y=y0.



Similarly, the slope of the tangent line on z surface along line x=x0 at point (x0, y0) is the rate of change at (x0, y0) along line x=x0.

How about the slope (or rate of change) of z at (x0, y0) along a general direction of u= (a, b)



It is called the directional derivative of z along direction **u**=(a, b) at point (x0, y0)

We use the parametric line equation and the chain rule to find the directional derivative.

Chain rule:

using chain rule :

$$f(x_0+at,\ y_0+bt)$$

$$x = x_0+at \qquad y = y_0+bt$$

$$t$$

directional derivative at (x_0, y_0)

$$D_{\vec{u}} f(x_0,y_0) = \left.\frac{df}{dt}\right|_{t=0}$$

$$\frac{df}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}$$

since $x = x_0 + at$, $y = y_0 + bt$

we have : $$\left.\frac{df}{dt}\right|_{t=0} = \frac{\partial f}{\partial x} a + \frac{\partial f}{\partial y} b$$

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

inner product

$$= \nabla f(x_0, y_0) \cdot \vec{u}$$

A gradient is a vector!

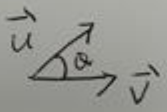
Geometric interpretation: if **u** coincides with (or lies on) the gradient vector $\nabla f(x0, y0)$, we shall have the greatest slope (rate of change), and the greatest rate of change is the length of the gradient vector. Therefore, at any point (x0, y0) on the X-Y plane, the direction to have the greatest rate of change (i.e., amount of increment on z due to a small unit of change of t, i.e., Δt) is the direction of the gradient vector.

$$\nabla f(x_0, y_0) = \left(\frac{\partial f(x_0, y_0)}{\partial x}, \frac{\partial f(x_0, y_0)}{\partial y}\right)$$

gradient or nabla

$$D_{\vec{u}} f(x_0, y_0) = \nabla f(x_0, y_0) \cdot \vec{u}$$

Note that $\vec{u} \cdot \vec{v} = \|\vec{u}\| \|\vec{v}\| \cos Q$

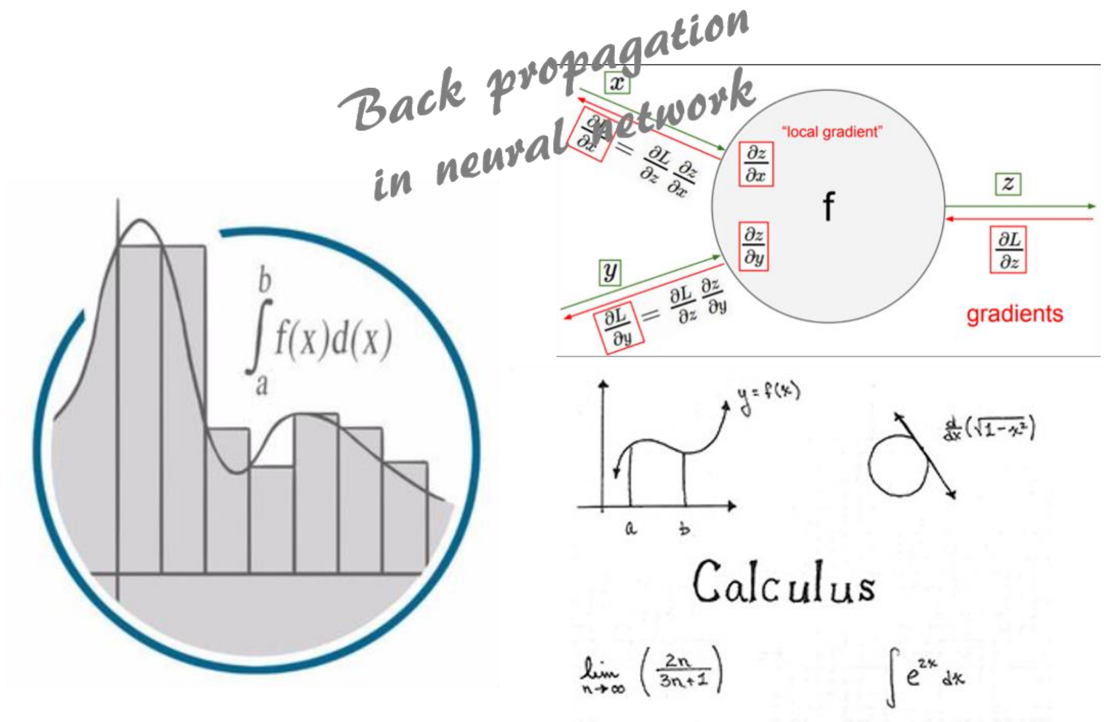


therefore, the maximum directional dervative happens when $Q = 0$, that is, when $\vec{u}$ coincides with the gradient vector

Summary

If we want to move on the x-y plane to (quickly) maximize the value of function z=f(x,y) at (x0, y0), we should move along the direction of the gradient vector at (x0, y0). In contrast, if we want to (quickly) minimize the value of f, we should move along the opposite direction of the gradient vector.
The idea of gradient descent is "to move along the opposite direction of the gradient vector to minimize the cost (or Loss) function."

More examples: A **gradient is always perpendicular to the level curve.**
[Gradient 1]



Gradient (Ascent or descent?)

- Let's start with a simple one!

  f(x, y)=x + y

- Given x = a, y = b, how to update x and y to make f(x, y) larger?
- Follow gradient directions!

  $$f(x, y) = x + y \Longrightarrow \frac{\partial f}{\partial x} = 1 \qquad \frac{\partial f}{\partial y} = 1$$

  Step size

  x= a + 0.01 * 1,
  y= b + 0.01 * 1 (a, b) + $\underline{0.01}$ * (1, 1)

  f(x, y): a+b $\Longrightarrow$ a+b+0.02

- A more complex one!

  f(x, y)=x * y

- Given x = a, y = b, how to update x and y to make f(x, y) larger?
- Follow gradient directions!

  f(x, y)= xy ⟹ $\frac{\partial f}{\partial x} = y \quad \frac{\partial f}{\partial y} = x$

  x= a + 0.01 * b,

  y= b + 0.01 * a

  f(x, y): a*b ⟹ (a+0.01*b)*(b+0.01*a)

  f(x,y):4*(-3) ⟹ 3.97*(-2.96) =-11.7512

set of training examples.

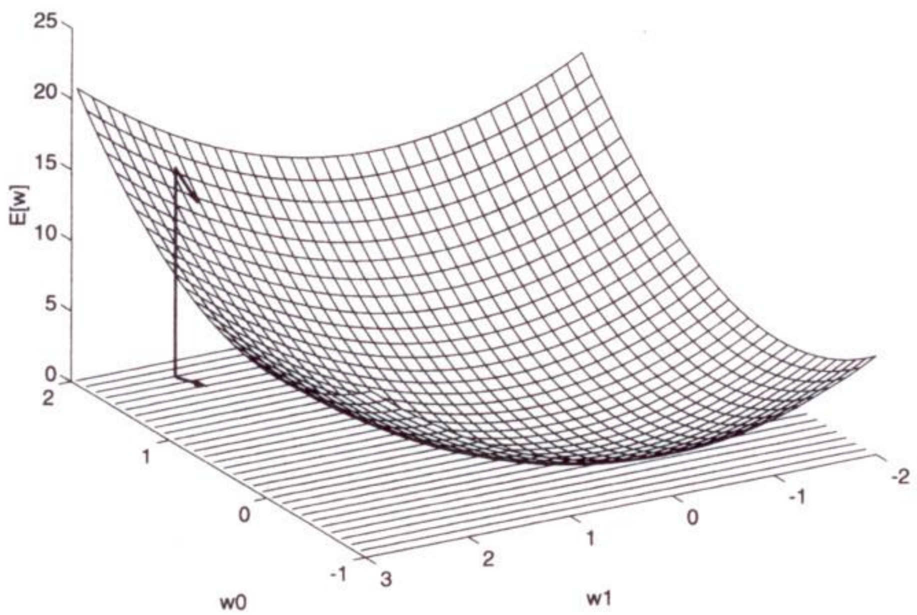


FIGURE 4.4
Error of different hypotheses. For a linear unit with two weights, the hypothesis space H is the w_0, w_1 plane. The vertical axis indicates the error of the corresponding weight vector hypothesis, relative to a fixed set of training examples. The arrow shows the negated gradient at one particular point, indicating the direction in the w_0, w_1 plane producing steepest descent along the error surface.