Week 4: Derivatives and Gradients

Product Rule for Derivatives (M&T p. 125)

Suppose $f : \mathbb{R}^n \to \mathbb{R}^k$ and $c : \mathbb{R}^n \to \mathbb{R}$ are differentiable at $\mathbf{x}_0$. Let $h(\mathbf{x}) = \underbrace{c(\mathbf{x})}_{\text{Scalar}}\underbrace{f(\mathbf{x})}_{\text{Vector}} : \mathbb{R}^n \to \mathbb{R}^k$.

$$\mathbf{D}h(\mathbf{x}_0) = c(\mathbf{x}_0)\mathbf{D}f(\mathbf{x}_0) + f(\mathbf{x}_0)\mathbf{D}c(\mathbf{x}_0)$$

Notice the dimensions of the matrices involved:

They match!

- $\mathbf{D}h(\mathbf{x}_0)$ and $c(\mathbf{x}_0)\mathbf{D}f(\mathbf{x}_0)$ are $k \times n$.
- $f(\mathbf{x}_0)$ is $k \times 1$.
- $\mathbf{D}c(\mathbf{x}_0)$ is $1 \times n$.

$$f\,\mathbf{D}c = (k\times 1)(1\times n) = k \times n$$

Yay, the dimensions all match! :)

Activity: Think-Pair-Board (3 min)

Think about these following task alone for one minute, then chat with your neighbour about it for two minutes, then share your ideas with the class by writing them on the board.

Generalize the addition rule and scaling rule for derivatives. (for D)

Example: Compute $\mathbf{D}\left(xy\begin{bmatrix}x\\y\end{bmatrix}\right)$ in Two Ways

The function $h(x,y) = xy\begin{bmatrix}x\\y\end{bmatrix}$ can be written in the form $c(\mathbf{x})f(\mathbf{x})$ where $c(\mathbf{x}) = xy$ and $f(\mathbf{x}) = \begin{bmatrix}x\\y\end{bmatrix}$.

1. Apply the product rule to calculate $\mathbf{D}h$.
2. Re-write $h(x,y)$ as $\begin{bmatrix}x^2y\\xy^2\end{bmatrix}$ and compute $\mathbf{D}h$ directly.

(1) Apply product rule to Dh

$$Dh = cDf + fDc$$

∂ \ partial $\quad$ 2 two

for $f(x,y) = \begin{bmatrix}x\\y\end{bmatrix}$ and $c(x,y) = xy$

$$Dh = xy\begin{bmatrix}\frac{\partial x}{\partial x} & \frac{\partial x}{\partial y}\\ \frac{\partial y}{\partial x} & \frac{\partial y}{\partial y}\end{bmatrix} + \begin{bmatrix}x\\y\end{bmatrix}\left[\frac{\partial(xy)}{\partial x}, \frac{\partial(xy)}{\partial y}\right]$$

$$= xy\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix} + \begin{bmatrix}x\\y\end{bmatrix}[y, x] = \begin{bmatrix}xy & 0\\0 & xy\end{bmatrix} + \begin{bmatrix}xy & x^2\\y^2 & xy\end{bmatrix}$$

$$= \begin{bmatrix}2xy & x^2\\y^2 & 2xy\end{bmatrix}$$

(2) Calculate Dh directly from $h = \begin{bmatrix}x^2y\\xy^2\end{bmatrix}$

$$Dh = \begin{bmatrix}\frac{\partial(x^2y)}{\partial x} & \frac{\partial(x^2y)}{\partial y}\\ \frac{\partial(xy^2)}{\partial x} & \frac{\partial(xy^2)}{\partial y}\end{bmatrix} = \begin{bmatrix}2xy & x^2\\y^2 & 2xy\end{bmatrix}$$

Yay! They match!

Invertibility of Df is important.

$$\underset{\mathbb{R}^n}{U} \xrightarrow{g} \underset{\mathbb{R}^k}{V} \xrightarrow{f} \mathbb{R}^d$$

Chain Rule for Total Derivatives (M&T p. 126)

Let $U \subset \mathbb{R}^n$ and $V \subset \mathbb{R}^k$ be open sets. Suppose that $g : U \subset \mathbb{R}^n \to \mathbb{R}^k$ and $f : V \subset \mathbb{R}^k \to \mathbb{R}^d$. Furthermore, assume $g(U) \subset V$ so that $f \circ g : U \subset \mathbb{R}^k \to \mathbb{R}^d$ is defined on all of U. If g is differentiable at $\mathbf{x}_0$ and f is differentiable at $\mathbf{y}_0 = g(\mathbf{x}_0)$ then:

$$\mathbf{D}(f \circ g)(\mathbf{x}_0) = \mathbf{D}f(\mathbf{y}_0)\mathbf{D}g(\mathbf{x}_0)$$

$f'(g(x))\,g'(x)$

$= f'(y)\,g'(x)$

TL;DR: The chain-rule says multiply the total derivatives evaluated at the appropriate points.

Example: A Weird Composition

Consider the following functions:

$$f(u, v, w) = (u - w, \cos(u + v), \sin(u + v + w)) \qquad g(x, y) = (e^x, \cos(y - x), e^{-y})$$

Compute $\mathbf{D}(f \circ g)$ by using the chain rule.

compute $Dg(x_0)$

$$Dg = \begin{bmatrix} \frac{\partial(e^x)}{\partial x} & \frac{\partial(\cancel{\cos(y-x)})}{\partial y} \ e^x & \cancel{\frac{\partial(e^x)}{\partial z}} \\ \frac{\partial(\cos(y-x))}{\partial x} & \frac{\partial(\cos(y-x))}{\partial y} & \cancel{\frac{\partial(\cos(y-x))}{\partial z}} \\ \frac{\partial(e^{-y})}{\partial x} & \frac{\partial(e^{-y})}{\partial y} & \cancel{\frac{\partial(e^{-y})}{\partial z}} \end{bmatrix}$$

$$= \begin{bmatrix} e^x & 0 \\ \sin(y-x) & -\sin(y-x) \\ 0 & -e^{-y} \end{bmatrix}$$

Compute $Df(y_0)$

Recall $f(u,v,w) = \left(u-w,\ \cos(u+v),\ \sin(u+v+w)\right)$

$$Df = \begin{bmatrix} \dfrac{\partial(u-w)}{\partial u} & \dfrac{\partial(u-w)}{\partial v} & \dfrac{\partial(u-w)}{\partial w} \\[2ex] \dfrac{\partial(\cos(u+v))}{\partial u} & \dfrac{\partial(\cos(u+v))}{\partial v} & \dfrac{\partial(\cos(u+v))}{\partial w} \\[2ex] \dfrac{\partial(\sin(u+v+w))}{\partial u} & \dfrac{\partial(\sin(u+v+w))}{\partial v} & \dfrac{\partial(\sin(u+v+w))}{\partial w} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & -1 \\ -\sin(u+v) & -\sin(u+v) & 0 \\ \cos(u+v+w) & \cos(u+v+w) & \cos(u+v+w) \end{bmatrix}$$

We now have Df and Dg.

Input g in to Df

Multipy $Df(g(x_0))$ and $Dg(x_0)$

Exercise: Calculate this matrix.

The result will be a $\boxed{3\times 2}$ matrix which represents $D(f\circ g)$

We see $f\circ g : \mathbb{R}^2 \to \mathbb{R}^3$

Alternatively, Df is 3×3 and Dg is 3×2.

Chain Rule and Gradients (M&T p. 127)

The following special case of the chain rule is very important.

Yay! single variable calc

If $\mathbf{c} : \mathbb{R} \to \mathbb{R}^k$ is a path and $f : \mathbb{R}^k \to \mathbb{R}$ is a function then $f \circ \mathbf{c} : \mathbb{R} \to \mathbb{R}$. In this case,

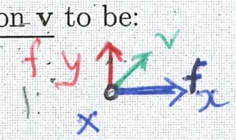
$$\frac{d(f \circ \mathbf{c})}{dt} = \mathbf{D}f(\mathbf{c}(t))\mathbf{D}\mathbf{c}(t) = \underbrace{\nabla f(\mathbf{c}(t))}_{\text{gradient}} \cdot \underbrace{\mathbf{c}'(t)}_{\text{vector}}$$

Directional Derivatives (M&T p. 136)

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$. We can define the <u>derivative of f at $\mathbf{x}$ in the direction $\mathbf{v}$</u> to be:

The notation from our book

$$\mathbf{D}f(\mathbf{x})\mathbf{v} = \frac{d}{dt} f(\mathbf{x} + t\mathbf{v})|_{t=0} = \lim_{t \to 0} \frac{f(\mathbf{x} + t\mathbf{v}) - f(\mathbf{x})}{\|t\mathbf{v}\|}$$



By convention, we *always* pick $\mathbf{v}$ to be a unit vector when calculating directional derivatives.

Example: Directional Derivatives and Gradients

Suppose that $f : \mathbb{R}^k \to \mathbb{R}$ and $\mathbf{v} \in \mathbb{R}^k$ (unit). The function $h(t) = f(\mathbf{x}+t\mathbf{v})$ can be written as a composition: $h(t) = f(\mathbf{c}(t))$ for the path $\mathbf{c}(t) = \mathbf{x} + t\mathbf{v}$. Use the chain rule to relate directional derivatives and gradients.

This composition has a function f and a path $c(t)$.

$$Df(x)v = \frac{d}{dt} f(x+tv)\Big|_{t=0}$$

$$= \nabla f(c(t)) \cdot c'(t)$$

$$c'(t) = \frac{d}{dt}\left[x+tv\right] = v$$

$$= \nabla f \cdot v$$

∂ = "partial"
d = "single"

We get: the directional derivative $Df(x)v$ is a gradient ~~###~~ ∇f dot with v

$$\boxed{Df(x)\,v = \nabla f \cdot v}$$

Theorem: ∇f is the Direction of Fastest Increase

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$ and $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$. The unit vector $\mathbf{n}$ that maximizes $\mathbf{D}f(\mathbf{x}_0)\mathbf{n}$ is $\mathbf{n} = \frac{\nabla f(\mathbf{x}_0)}{\|\nabla f(\mathbf{x}_0)\|}$.

Machine learning: gradient ascent

The Big Idea: $\mathbf{D}f(\mathbf{x}_0)\mathbf{n} = \nabla f(\mathbf{x}_0) \cdot \mathbf{n} = \|\nabla f(\mathbf{x}_0)\| \, \|\mathbf{n}\| \cos\theta = \|\nabla f(\mathbf{x}_0)\| \cos\theta$.
And we know that $\cos\theta$ is maximized when $\theta = 0$.

$$\begin{aligned}
Df(x_0)\,n &= \nabla f(x_0) \cdot n && \text{\# apply definition} \\
&= \|\nabla f(x_0)\| \, \|n\| \cos\theta && \text{\# geometric dot product} \\
&= \|\nabla f(x_0)\| \cdot 1 \cdot \cos\theta && \text{\# n is a unit vector} \\
&= \underbrace{\|\nabla f(x_0)\|}_{\text{fixed non-zero vector}} \cos\theta && \leftarrow \text{This is maximized when } \theta = 0 \text{ and } \cos\theta = 1.
\end{aligned}$$

The biggest value of $Df(x_0)\,n$ we can possibly get is when n is parallel ($\theta = 0$) to $\nabla f(x_0)$

Thus, the max occurs at $n = \dfrac{\nabla f(x_0)}{\|\nabla f(x_0)\|}$

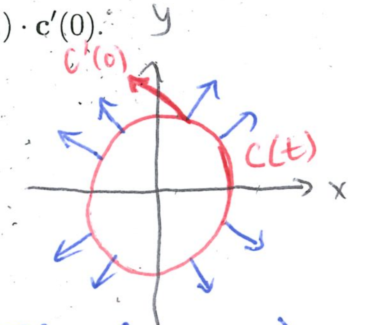
Theorem: ∇f is Perpendicular to Level Surfaces

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$ and $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$.
Let $c = f(\mathbf{x}_0)$ and consider the level surface $L_c = \{\mathbf{x} : c = f(\mathbf{x})\} \subset \mathbb{R}^n$ passing through $\mathbf{x}_0$.
The gradient vector $\nabla f(\mathbf{x}_0)$ is perpendicular to the tangent plane of L_c at $\mathbf{x}_0$.

The Big Idea: Pick any curve $\mathbf{c}(t)$ in the level surface L_c passing through $\mathbf{x}_0$ at $t = 0$.
We know that $\dfrac{d(f \circ \mathbf{c})}{dt} = 0$ because $f \circ \mathbf{c}$ is constant. However, $\dfrac{d(f \circ \mathbf{c})}{dt} = \nabla f(\mathbf{x}_0) \cdot \mathbf{c}'(0)$.

We pick $c(t)$ to be contained in a level surface L_c. It follows:

$$f(c(t)) = c = \text{const}.$$



$\nabla f = (2x, 2y)$
$f(x,y) = x^2 + y^2$
L_c is a circle of radius $\sqrt{c}$ about $(0,0)$.

We then get:

$$\frac{d}{dt}\left[f(c(t))\right] = \frac{d}{dt}\left[\text{const}\right] = 0$$

We also have:

$$\frac{d}{dt}\left[f(c(t))\right] = \boxed{\nabla f(x_0)} \cdot \boxed{c'(0)} = 0$$

Thus, $c'(0)$ is orthogonal to $\nabla f(x_0)$.

Example: The Two Unit Normals to a Curve

Find all unit vectors normal to the curve $x^3 + xy + y^3 = 11$ at $(x, y) = (1, 2)$.

The curve $x^3+xy+y^3=11$ is a level curve of the function $f(x,y) = x^3+xy+y^3$.

No idea what the curve looks like but we can do geometry on it!

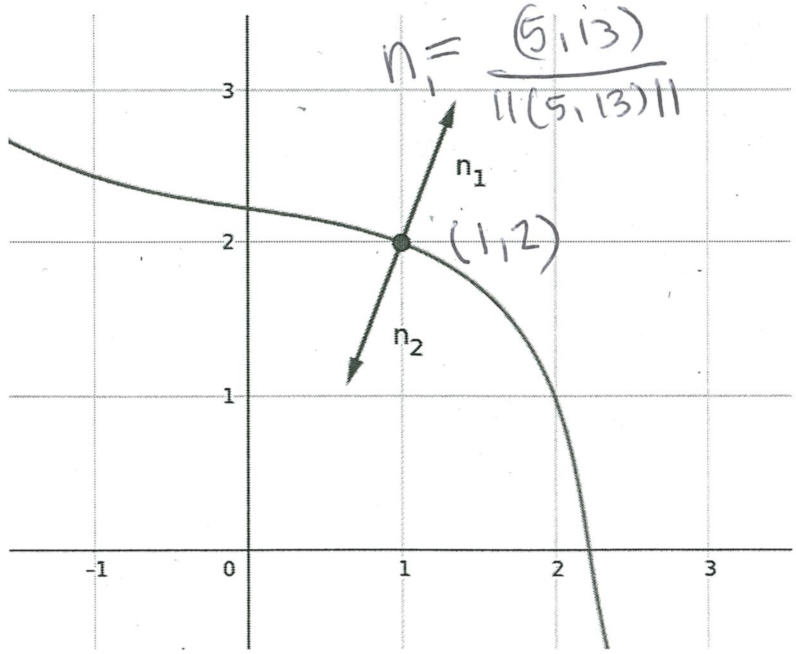
It is L_{11} of $f(x,y)$.

We calculate $\nabla f = \left(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right) = (3x^2+y, x+3y^2)$

This gives $\nabla f(1,2) = (5, 13)$ ← This is perpendicular to the curve L_{11}.

We make the two normal vectors as ~~the~~ unit vectors:

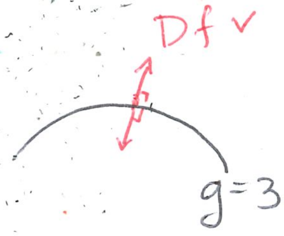
$$n_1 = \frac{(5,13)}{\|(5,13)\|} \qquad n_2 = \frac{(-5,-13)}{\|(5,13)\|}$$



Example: Rate of Change Normal to a Surface

Find the rate of change of $f(x, y, z) = xyz$ in the direction normal to the surface $yx^2 + xy^2 + yz^2 = 3$ at $(x, y, z) = (1, 1, 1)$.

Find directions normal to $yx^2 + xy^2 + yz^2 = 3$



Define $g(x,y,z) = yx^2 + xy^2 + yz^2$

We calculate $\nabla g = \left(\frac{\partial g}{\partial x}, \frac{\partial g}{\partial y}, \frac{\partial g}{\partial z}\right)$

$$= \left(2xy + y^2,\ x^2 + 2xy + z^2,\ 2yz\right)$$

At $(1,1,1)$ we get $\nabla g = (3,4,2)$

This gives $n_1 = \frac{(3,4,2)}{\|(3,4,2)\|} = \frac{1}{\sqrt{29}}(3,4,2)$ and $n_2 = \frac{1}{\sqrt{29}}(-3,-4,-2)$

Evaluate the directional derivatives of f along n_1 and n_2

$Df(1,1,1)\,v = \nabla f(1,1,1) \cdot v$

Calculate Df

$\nabla f = (yz, xz, xy)$

$\nabla f(1,1,1) = (1,1,1)$

$= (1,1,1) \cdot v$

This gives:

$$Df(1,1,1)\,n_1 = (1,1,1) \cdot n_1 = \frac{1 \cdot 3 + 1 \cdot 4 + 1 \cdot 2}{\sqrt{29}} = \frac{9}{\sqrt{29}}$$

$$Df(1,1,1)\,n_2 = \frac{-9}{\sqrt{29}}$$