Continuous Optimization (Chapter 7: MML Book)

Latex macros

Recall geometry of a derivative

Definition (Directional derivative)

Directional dervivative of a function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ with respect to a given vector $\mathbf{u} \in \mathbb{R}$ is defined as

$$D_{\mathbf{u}} f(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon}$$

Ref Khan Academy

Ref Libretexts/12%3A_Functions_of_Several_Variables/12.06%3A_Directional_Derivatives

Vector calculus chain rule (a theorem)

Given a function composition $\mathbf{f}(\mathbf{x}) = \mathbf{g}(\mathbf{h}(\mathbf{x})) = (\mathbf{g} \circ \mathbf{h})(\mathbf{x})$ where $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{g} : \mathbb{R}^m \to \mathbb{R}^p$ and $\mathbf{h} : \mathbb{R}^m \to \mathbb{R}^p$

$$\frac{\p \mathbf{f}}{\p \mathbf{x}} = \frac{\p \mathbf{f}}{\p \mathbf{h}} \frac{\p \mathbf{h}}{\p \mathbf{x}}$$

or denoting the derivatives as Jacobian matrices we have,

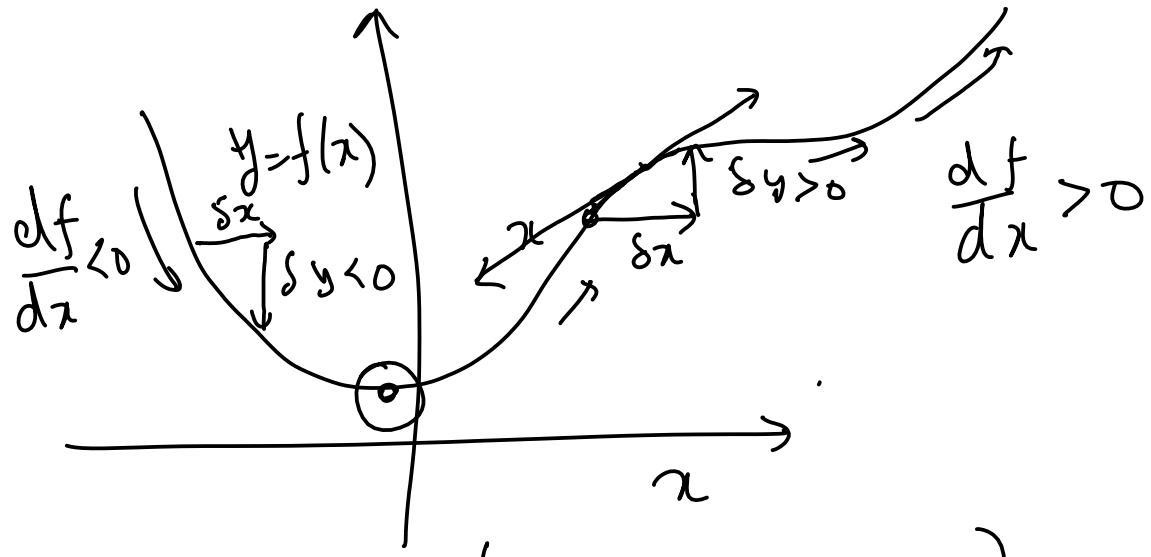
$$\mathcal{J}_{\mathbf{x}} \mathbf{f} = \mathcal{J}_{\mathbf{h}}[\mathbf{f}] \mathcal{J}_{\mathbf{x}}[\mathbf{h}]$$

Theorem (Directional derivative is gradient dot product with the direction)

Express the trajectory in the direction $\mathbf{u}$ as a function of time t as

$$\mathbf{g}(t) = \mathbf{x} + t\mathbf{u}$$

Note that the Jacobian of $\mathbf{g}(t)$ wrt t is simply $\mathbf{u}$,



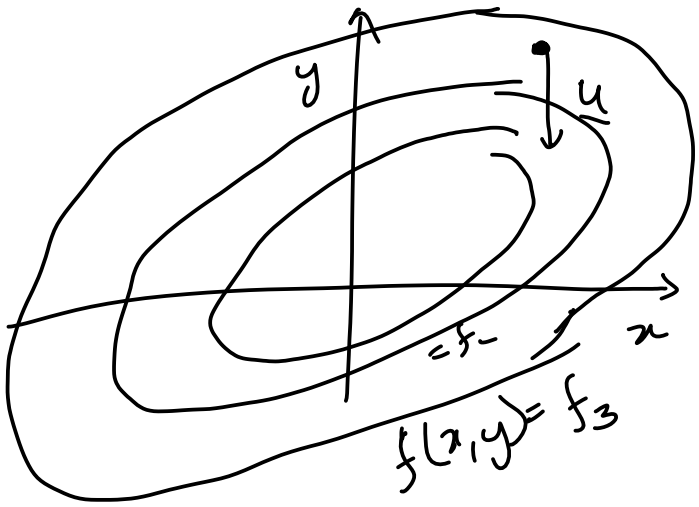
$$\frac{df(x)}{dx} = \lim_{\delta x \to 0} \frac{[f(x+\delta x) - f(x)]}{\delta x} = \lim_{\delta x \to 0} \frac{\delta y}{\delta x}$$

$$x_{t+1} = x_t - \alpha \frac{df(x)}{dx}$$ | want to generalize this to higher dim

---

Directional derivative

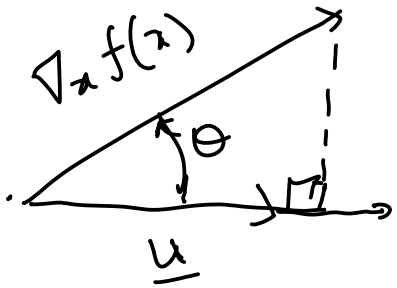
$$D_{\underline{u}} f(\underline{x}) = \lim_{h \to 0} \frac{f(\underline{x} + h\underline{u}) - f(\underline{x})}{h}$$



Theorem

$$D_{\underline{u}} f(\underline{x}) = \left[\nabla_{\underline{x}} f(\underline{x})\right]^T \underline{u} = \left(\nabla_{\underline{x}} f(x)\right) \cdot \underline{u}$$

$$= \left\| \nabla_x f(x) \right\| \|u\| \cos\theta$$



Suppose $\|u\| = 1$

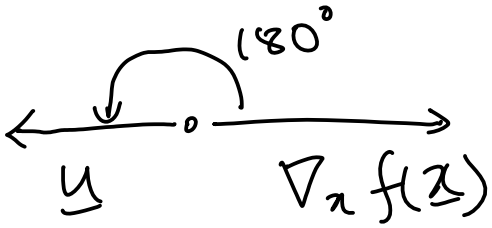
$$D_u f(x) = \|\nabla_x f(x)\| \cos(\theta)$$

find the direction in which $f(x)$ decreases fastest?

$$\min_u D_u f(x)$$

$$\theta = 180^\circ = \arg\min_\theta D_u f(x) = \|\nabla_x f(x)\| \cos(\theta)$$

steepest descent



$$u \propto -\nabla_x f(x)$$

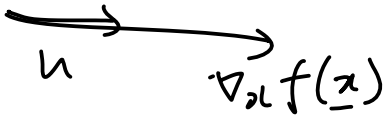
Proportionality

$$u = -\frac{\nabla_x f(x)}{\|\nabla_x f(x)\|}$$

$$0^\circ = \arg\max_\theta D_u f(x) = \|\nabla_x f(x)\| \cos(\theta)$$

steepest ascent

$$u \propto \nabla_x f(x)$$



$$\min_{\underline{x}} f(\underline{x})$$

$$\frac{\partial f(\underline{x})}{\partial \underline{x}} = 0$$

not always has analytical solution

Algorithm : Gradient descent

$\underline{x}_{t=0}$ ≃ choose randomly

while $\|\underline{x}_{t+1} - \underline{x}_t\| > 10^{-6}$:

$$x_{t+1} = x_t - \underbrace{\alpha_t}_{\text{step size / learning rate}} \left[\nabla_{\underline{x}} f(\underline{x}) \right]_{x_t}$$

x_{t+1} = minimum of the function $f(\underline{x})$

$$\mathcal{J}_t \mathbf{g}(t) = \mathbf{u}$$

Recall the definition of directional derivative,

$$D_{\mathbf{u}} f(\mathbf{x}) = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon}.$$

Compare it with the derivative of $f(\mathbf{g}(t))$ with respect to t at $t = 0$

$$\frac{\mathrm{p} f(\mathbf{g}(t))}{\mathrm{p} t} = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + (t+\epsilon)\mathbf{u}) - f(\mathbf{x} + t\mathbf{u})}{\epsilon} \Big|_{t=0}.$$

$$\frac{\mathrm{p} f(\mathbf{g}(t))}{\mathrm{p} t} = \lim_{\epsilon \to 0} \frac{f(\mathbf{x} + \epsilon \mathbf{u}) - f(\mathbf{x})}{\epsilon} = D_{\mathbf{u}} f(\mathbf{x}).$$

We can compute $\frac{\mathrm{p} f(\mathbf{g}(t))}{\mathrm{p} t}$ by chain rule,

$$D_{\mathbf{u}} f(\mathbf{x}) = \mathcal{J}_t f(\mathbf{g}(t)) = \mathcal{J}_{\mathbf{x}} f(\mathbf{x}) \mathcal{J}_t \mathbf{g} = \nabla_{\mathbf{x}}^\top f(\mathbf{x}) \mathbf{u}$$

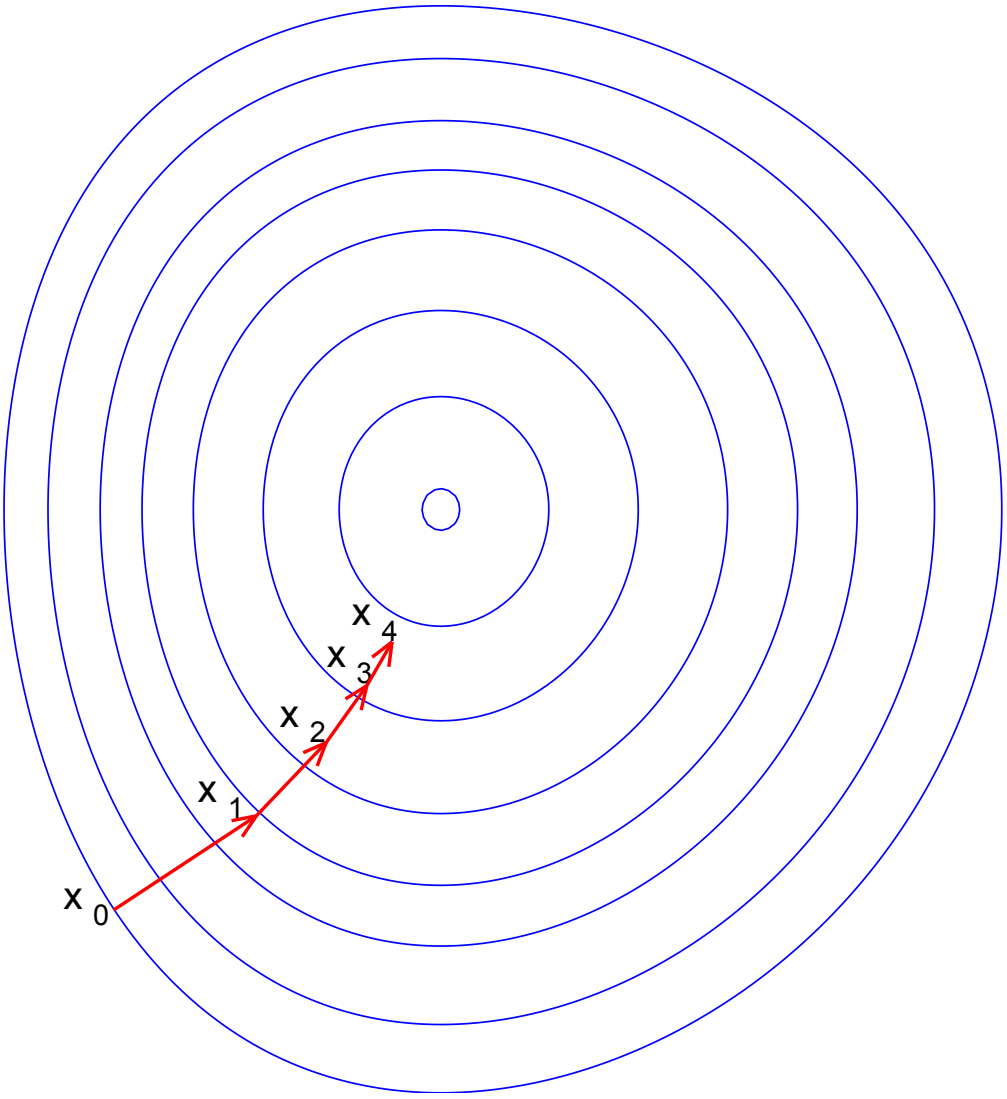
Theorem : The direction of stepest ascent and descent

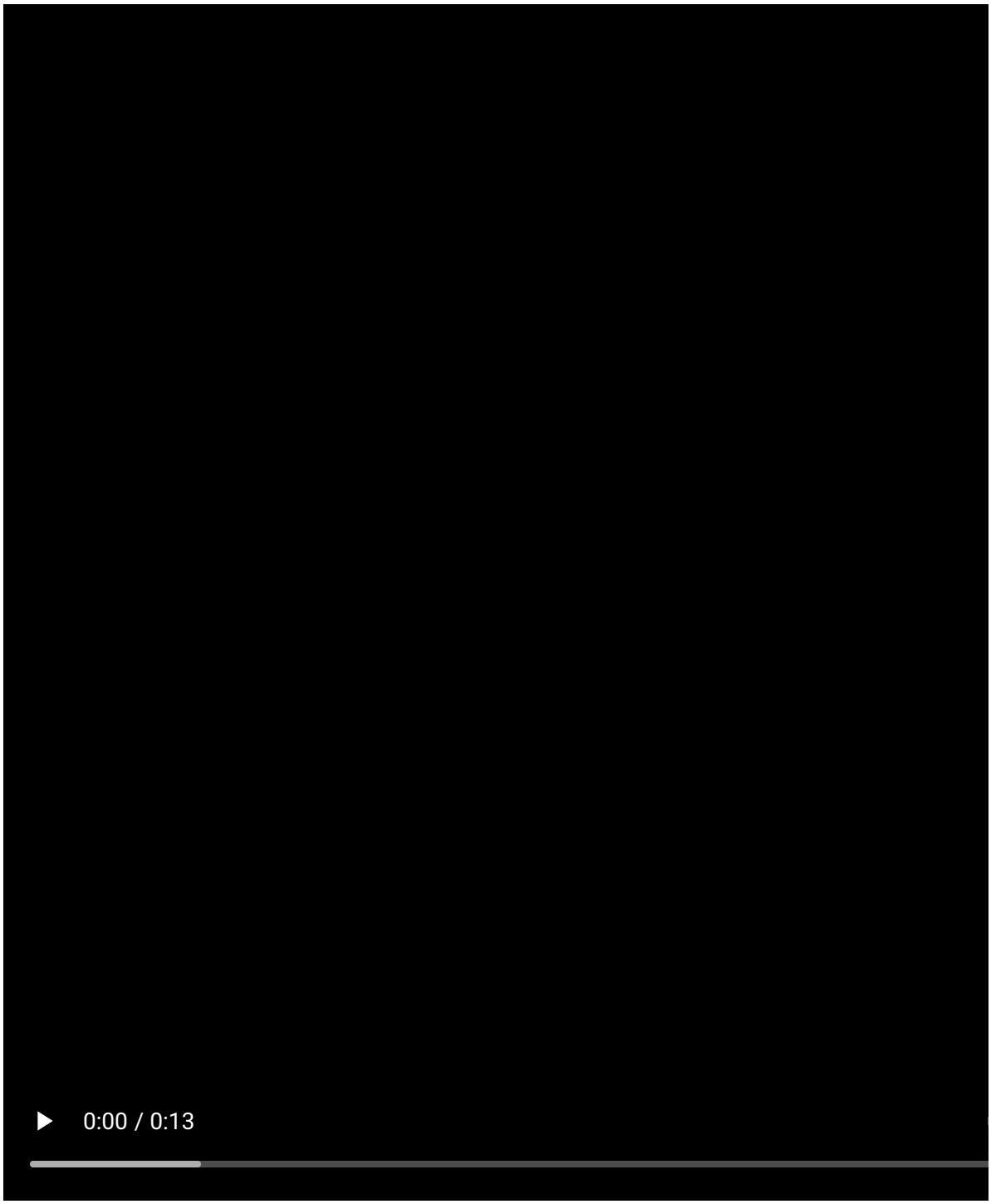
Let $\hat{\mathbf{u}}$ be of unit magnitude. The directional derivative represents how the function changes in the direction $\hat{\mathbf{u}}$.

$$D_{\hat{\mathbf{u}}} f(\mathbf{x}) = \nabla_{\mathbf{x}}^\top f(\mathbf{x}) \hat{\mathbf{u}} = \|\nabla_{\mathbf{x}} f(\mathbf{x})\| \cos(\theta),$$

where θ is the angle between $\nabla_{\mathbf{x}} f(\mathbf{x})$ and $\hat{\mathbf{u}}$. The change is maximum when $\theta = 0$ and $\cos(\theta) = 1$ and the change is minimum when $\theta = 180^\circ$ and $\cos(\theta) = -1$.

In other words the function f increases the most (stepest ascent) when $\hat{\mathbf{u}} \propto \nabla_{\mathbf{x}} f(\mathbf{x})$ and decreases the most (steepest descent) when $\hat{\mathbf{u}} \propto -\nabla_{\mathbf{x}} f(\mathbf{x})$.


x 0
x 1
x 2
x 3
x 4


0:00 / 0:13