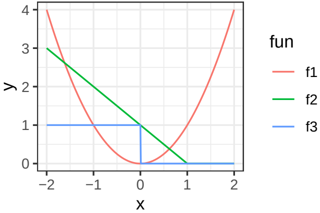
Optimization

Smoothness & Gradients

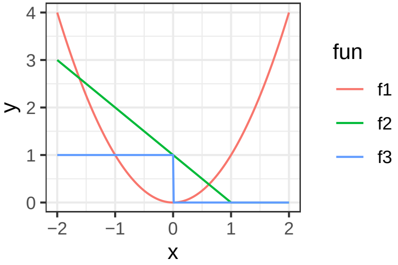


Learning goals

- Definition of smoothness
- Uni- & multivariate differentiation

SMOOTH VS. NON-SMOOTH

- **Smoothness** of a function $f : \mathcal{S} \to \mathbb{R}$ is measured by the number of its continuous derivatives
- If k-th derivative of f exists and is continuous on $\mathcal{S}$: f is k-times continuously differentiable; $f \in \mathcal{C}^k$ (class of k-times continuous differentiable functions)
- In this lecture, we call f “smooth”, if at least $f \in \mathcal{C}^1$



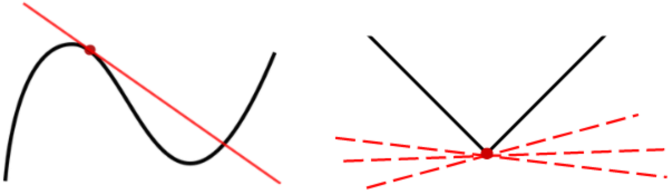
f_1 is smooth, f_2 is continuous but not differentiable, and f_3 is non-continuous.

UNIVARIATE DIFFERENTIABILITY

Definition: A function $f : \mathcal{S} \subseteq \mathbb{R} \to \mathbb{R}$ is said to be differentiable in $x \in \mathcal{S}$ if the following limit exists:

$$f'(x) := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Intuitively: f can be approximated locally by a linear function with slope $m = f'(x)$.



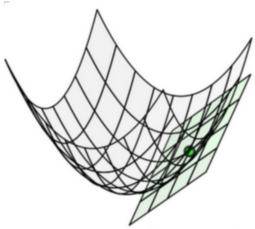
Left: Function is differentiable everywhere. Right: Not differentiable at the red point.

MULTIVARIATE DIFFERENTIABILITY

A similar definition of differentiability holds for multivariate functions.

Definition: A function $f : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ is differentiable in $\mathbf{x} \in \mathcal{S}$ if there exists a (continuous) linear map $\nabla f(\mathbf{x}) : \mathcal{S} \subseteq \mathbb{R}^d \to \mathbb{R}$ with

$$\lim_{\boldsymbol{h} \to 0} \frac{f(\mathbf{x} + \boldsymbol{h}) - f(\mathbf{x}) - \nabla f(\mathbf{x}) \cdot \boldsymbol{h}}{||\boldsymbol{h}||} = 0$$



Geometrically: The function can be locally approximated by a tangent hyperplane.

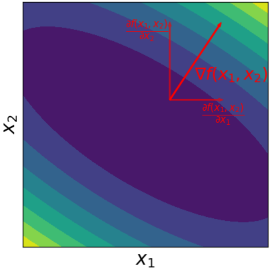
Source: `https://github.com/jermwatt/machine_learning_refined`.

GRADIENT

This linear approximation is given by the **gradient**:

$$\nabla f = \frac{\partial f}{\partial x_1}\boldsymbol{e}_1 + \cdots + \frac{\partial f}{\partial x_n}\boldsymbol{e}_n = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right)^\top.$$

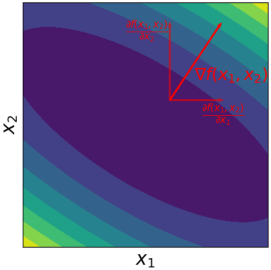
The elements of the gradient are called **partial derivatives**.

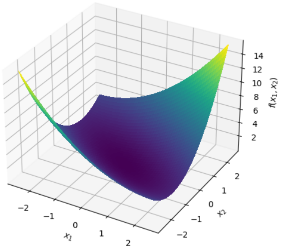


GRADIENT

Consider $f(\mathbf{x}) = 0.5x_1^2 + x_2^2 + x_1x_2$. The gradient is

$$\nabla f(\mathbf{x}) = (x_1 + x_2, 2x_2 + x_1)^\top.$$





DIRECTIONAL DERIVATIVE

The directional derivative tells how fast $f : \mathcal{S} \to \mathbb{R}$ is changing w.r.t. an arbitrary direction $\boldsymbol{v}$:

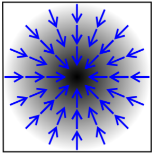
$$D_v f(\mathbf{x}) := \lim_{h \to 0} \frac{f(\mathbf{x} + h\boldsymbol{v}) - f(\mathbf{x})}{h} = \nabla f(\mathbf{x}) \cdot \boldsymbol{v}.$$

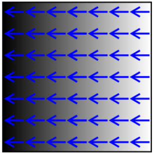
Example: The instantaneous rate of change in direction $\boldsymbol{v} = (1, 1)$ is:

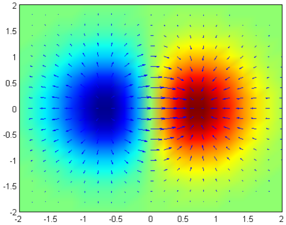
$$D_v f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{\partial f}{\partial x_1} + \frac{\partial f}{\partial x_2}$$

PROPERTIES OF THE GRADIENT

- Gradient is orthogonal to level curves / surfaces of a function
- The gradient points in the direction of greatest increase of f





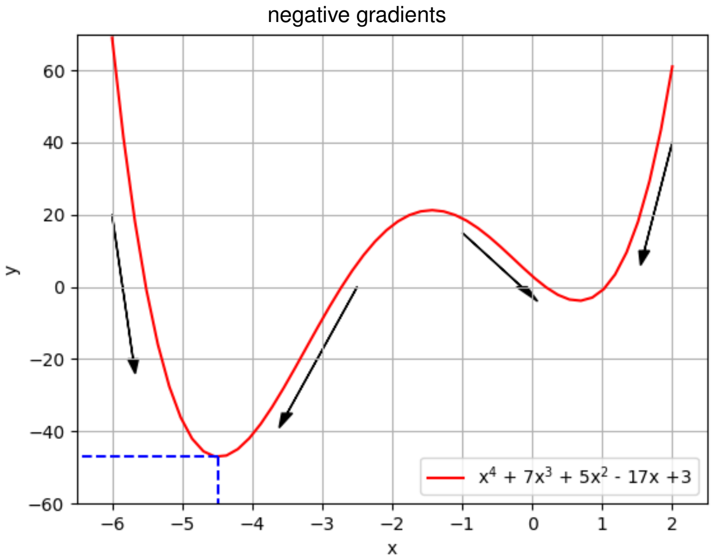


Proof: Let $\boldsymbol{v}$ be a vector of length 1. Let θ the angle between $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$.

$$D_{\boldsymbol{v}} f(\mathbf{x}) = \nabla f(\mathbf{x})^\top \boldsymbol{v} = \|\nabla f(\mathbf{x})\| \, \|\boldsymbol{v}\| \cos(\theta) = \|\nabla f(\mathbf{x})\| \cos(\theta)$$

using the cosine formula for dot products and because $\|v\| = 1$ by assumption.
$\cos(\theta)$ is maximal if $\theta = 0$, which is if $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$ point in the same direction.
(Alternative proof: Apply Cauchy-Schwarz to $\nabla f(\mathbf{x})^\top \boldsymbol{v}$ and show for which $\boldsymbol{v}$ the inequality holds with equality.)

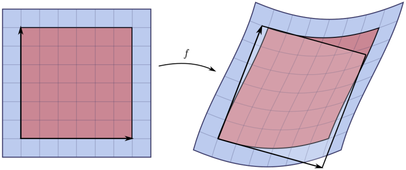
PROPERTIES OF THE GRADIENT



JACOBI MATRIX

Let $f : \mathcal{S} \to \mathbb{R}^m$ be a vector-valued function with components $f_1, f_2, ..., f_m$. **Jacobian** matrix as generalization of gradient:

$$J_f = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$



$f : \mathbb{R}^2 \to \mathbb{R}^2$ sends a small square (left, red) to a distored parallelogram (right, red). Jacobian gives best linear aproximation of di storted parallelogram near that point. Source: Wikipedia.