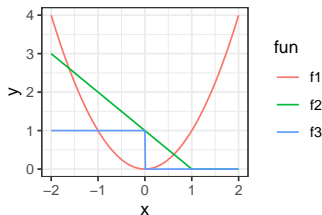


Optimization

Smoothness & Gradients



Learning goals

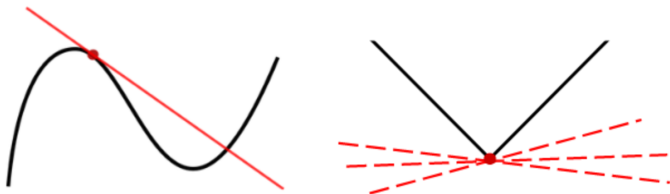
- Definition of smoothness
- Uni- & multivariate differentiation
- Gradient, partial derivatives, gradient
- Jacobi-Matrix

UNIVARIATE DIFFERENTIABILITY

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

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

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.

SMOOTH VS. NON-SMOOTH

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

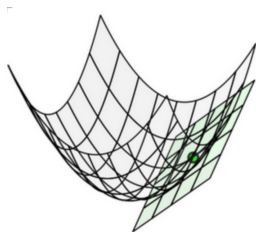


f_1 is smooth, f_2 is continuous but not differentiable, and f_3 is non-continuous.

MULTIVARIATE DIFFERENTIABILITY

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

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



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

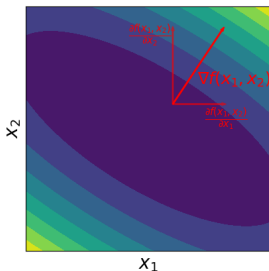
Source: https://github.com/jermwatt/machine_learning_refined.

GRADIENT

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

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

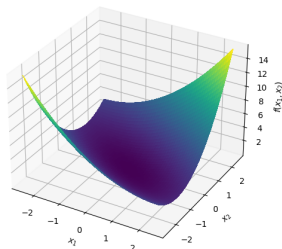
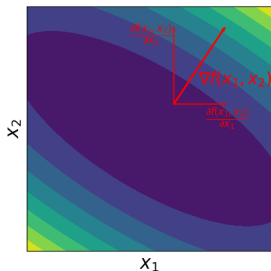
The elements of the gradient are called **partial derivatives**.



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 $f : \mathcal{S} \rightarrow \mathbb{R}$ is changing w.r.t. an arbitrary direction \mathbf{v} :

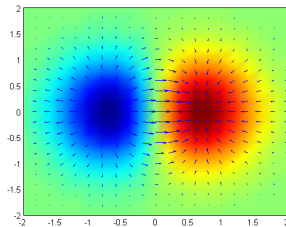
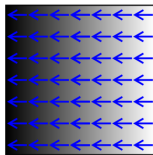
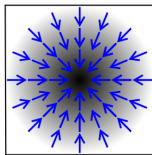
$$D_{\mathbf{v}}f(\mathbf{x}) := \lim_{h \rightarrow 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h} = \nabla f(\mathbf{x}) \cdot \mathbf{v}.$$

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

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

PROPERTIES OF THE GRADIENT

- Orthogonal to level curves / surfaces of a function
- Points in direction of greatest increase of f



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

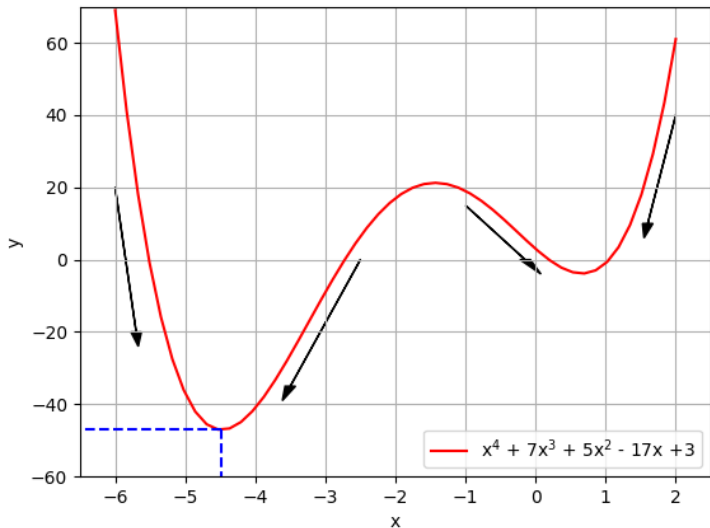
$$D_{\mathbf{v}}f(\mathbf{x}) = \nabla f(\mathbf{x})^{\top} \mathbf{v} = \|\nabla f(\mathbf{x})\| \|\mathbf{v}\| \cos(\theta) = \|\nabla f(\mathbf{x})\| \cos(\theta)$$

using the cosine formula for dot products and because $\|\mathbf{v}\| = 1$ by assumption. $\cos(\theta)$ is maximal if $\theta = 0$, which is if \mathbf{v} and $\nabla f(\mathbf{x})$ point in the same direction.

(Alternative proof: Apply Cauchy-Schwarz to $\nabla f(\mathbf{x})^{\top} \mathbf{v}$ and show for which \mathbf{v} the inequality holds with equality.)

PROPERTIES OF THE GRADIENT

- Negative gradient $-\nabla f(\mathbf{x})$ points in direction of greatest decrease

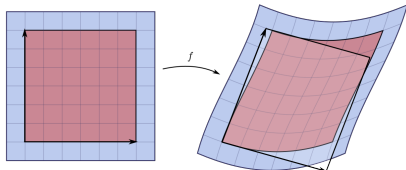


JACOBI MATRIX

Let $f : \mathcal{S} \rightarrow \mathbb{R}^m$ be vector-valued with components f_1, f_2, \dots, f_m .

Jacobian matrix as generalization of gradient:

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



$f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ sends a small square (left, red) to a distorted parallelogram (right, red).

Jacobian gives best linear approximation of distorted parallelogram near that point.

Source: Wikipedia.