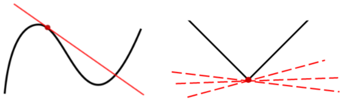
Optimization in Machine Learning

Mathematical Concepts: Differentiation and Derivatives



Learning goals

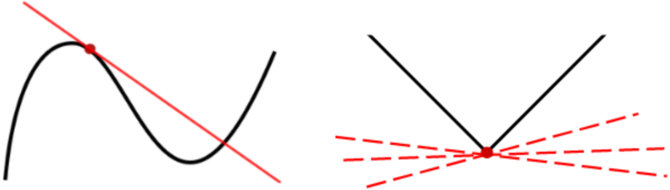
- Definition of smoothness
- Uni- & multivariate differentiation
- Gradient, partial derivatives
- Jacobi-Matrix
- Hessian Matrix

UNIVARIATE DIFFERENTIABILITY

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

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

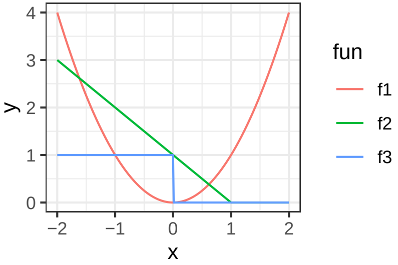
Intuitively: f can be approxed locally by a lin. fun. with slope $m = f'(x)$.



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

SMOOTH VS. NON-SMOOTH

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

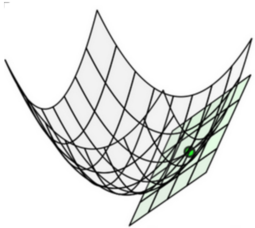


f_1 is smooth, f_2 is continuous but not differentiable, and f_3 is non-continuous.

MULTIVARIATE DIFFERENTIABILITY

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

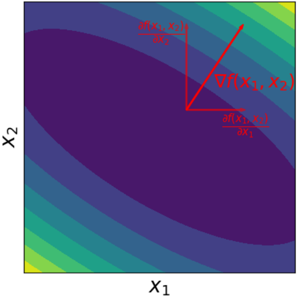
GRADIENT

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**.

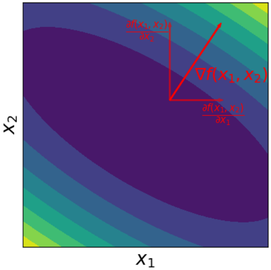
To compute a partial derivative in x_j, we treat the function as univariate in x_j (and everything else as constant), then compute the derivative in x_j.

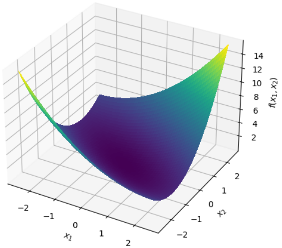


GRADIENT

Consider $f(\mathbf{x}) = 0.5x_1^2 + x_2^2 + x_1 x_2$. 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 $\boldsymbol{v}$:

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

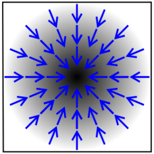
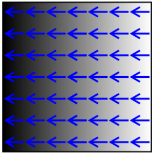
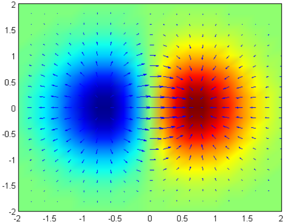
Example: The directional derivative for $\boldsymbol{v} = (1, 1)$ is:

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

NB: Some people require that $||\boldsymbol{v}|| = 1$. Then, we can identify $D_v f(\mathbf{x})$ with the instantaneous rate of change in direction $\boldsymbol{v}$ – and in our example we would have to divide by $\sqrt{2}$.

PROPERTIES OF THE GRADIENT

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

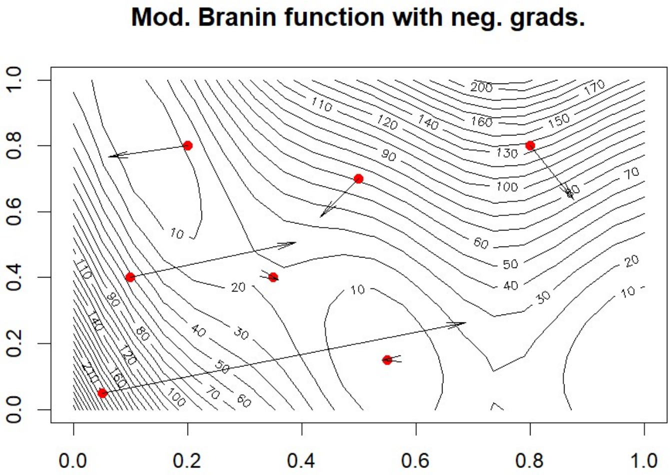
Proof: Let $\boldsymbol{v}$ be a vector of length 1. Let θ the angle between $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$.

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

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

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

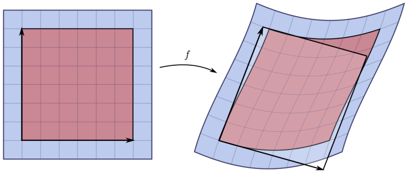


(Length of arrow is scaled norm of gradient)

JACOBI MATRIX

Let $f : \mathcal{S} \to \mathbb{R}^m$ be vector-valued with components $f_1, f_2, ..., f_m$.
Jacobian generalizes the gradient by placing the ∇f_j in its rows:

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



$f : \mathbb{R}^2 \to \mathbb{R}^2$ sends a small square (left, red) originating from a given input point to a distorted parallelogram (right, red) . Jacobian gives best linear approximation of distorted parallelogram near that point. Source: Wikipedia.

DEFINITION HESSIAN MATRIX

The 2nd derivative of a multivariate function $f \in \mathcal{C}^2(\mathcal{S}, \mathbb{R})$, $\mathcal{S} \subseteq \mathbb{R}^d$ (if it exists) is defined by the **Hessian** matrix

$$H(\mathbf{x}) = \nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}\right)_{i,j=1...d}$$

Example: Let $f(x_1, x_2) = sin(x_1) \cdot cos(2x_2)$. Then:

$$H(\mathbf{x}) = \begin{pmatrix} \text{-cos}(2x_2) \cdot \sin(x_1) & \text{-2cos}(x_1) \cdot \sin(2x_2) \\ \text{-2cos}(x_1) \cdot \sin(2x_2) & \text{-4cos}(2x_2) \cdot \sin(x_1) \end{pmatrix}$$

- If all 2nd partial derivatives are continuous, H will be symmetric
- Many local properties, w.r.t. geometry, convexity, critical points, are encoded by the Hessian and its Eigenspectrum ($\rightarrow$ later).

HESSIAN DESCRIBES LOCAL CURVATURE

Let w.l.o.g. $A(\mathbf{x}) = \{\lambda_{1,\mathbf{x}}, ..., \lambda_{d,\mathbf{x}}\}$ be Eigenspectrum with $\lambda_{1,\mathbf{x}} \leq \lambda_{2,\mathbf{x}} \leq ... \leq \lambda_{d,\mathbf{x}}$ of $H(\mathbf{x})$; let $\boldsymbol{v}_{i,\mathbf{x}}$ define the respective Eigenvectors. We can read off it:

- $\boldsymbol{v}_d/\boldsymbol{v}_1$ points in the direction of largest/smallest curvature

Example (continued): $H(\mathbf{x}) = \begin{pmatrix} -\cos(2x_2)\cdot\sin(x_1) & -2\cos(x_1)\cdot\sin(2x_2) \\ -2\cos(x_1)\cdot\sin(2x_2) & -4\cos(2x_2)\cdot\sin(x_1) \end{pmatrix}$.

- $H(a),\ a = (\frac{-\pi}{2}, 0)$: $\lambda_{1,a} = 1, \lambda_{2,a} = 4$; $v_{1,a} = (1,0)^\top$, $v_{2,a} = (0,1)^\top$