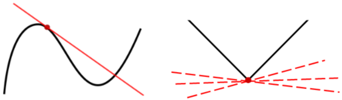
Optimization in Machine Learning

Mathematical Concepts
Differentiation and Derivatives



Learning goals

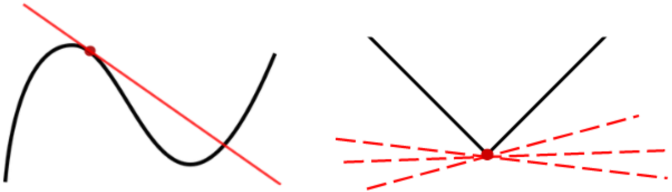
- Definition of smoothness
- Uni- & multivariate differentiation
- Gradient, partial derivatives
- Jacobian matrix
- Hessian matrix
- Lipschitz continuity

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:

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

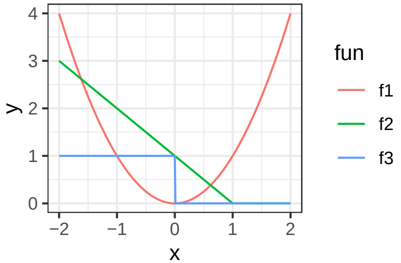
Intuitively: f can be approxed locally by a lin. fun. with slope $m = f'(x)$.



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
- $\mathcal{C}^k$ is class of k-times continuously differentiable functions ($f \in \mathcal{C}^k$ means $f^{(k)}$ exists and is continuous)
- In this lecture, we call f "smooth", if at least $f \in \mathcal{C}^1$

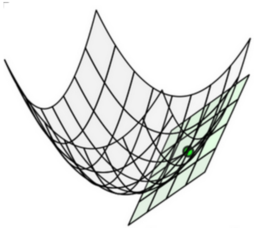


f_1 is smooth, f_2 is continuous but not differentiable, and f_3 is non-continuous.

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

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



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

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

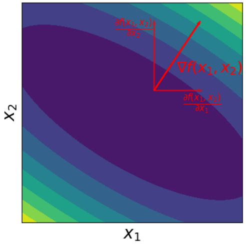
GRADIENT

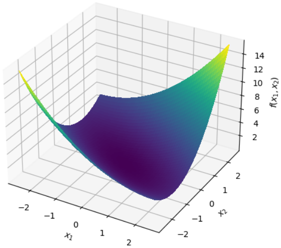
- Linear approximation is given by the **gradient**:

$$\nabla f = \frac{\partial f}{\partial x_1}\boldsymbol{e}_1 + \cdots + \frac{\partial f}{\partial x_d}\boldsymbol{e}_d = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d}\right)^T$$

- Elements of the gradient are called **partial derivatives.**
- To compute $\partial f / \partial x_j$, regard f as function of x_j only (others fixed)

Example: $f(\mathbf{x}) = x_1^2/2 + x_1 x_2 + x_2^2 \Rightarrow \nabla f(\mathbf{x}) = (x_1 + x_2, x_1 + 2x_2)^T$





DIRECTIONAL DERIVATIVE

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

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

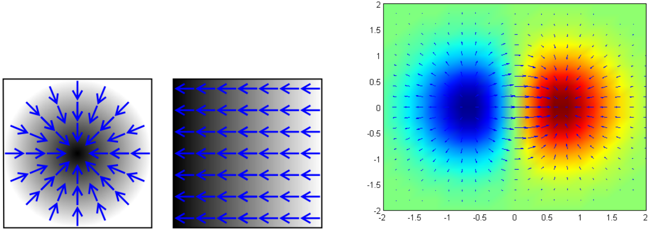
Example: The directional derivative for $\boldsymbol{v} = (1, 1)$ is:

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

NB: Some people require that $||\boldsymbol{v}|| = 1$. Then, we can identify $D_{\boldsymbol{v}} f(\mathbf{x})$ 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



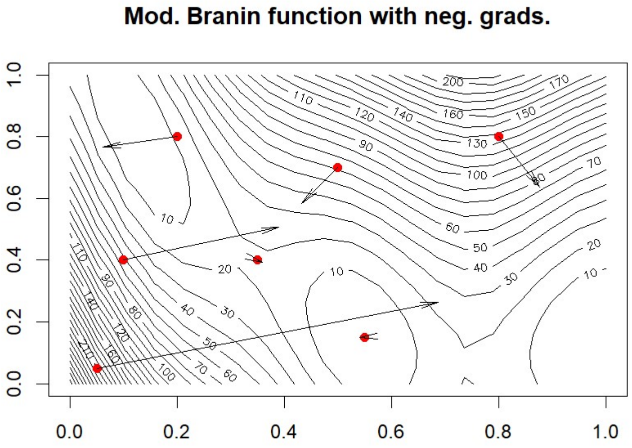
Proof: Let $\boldsymbol{v}$ be a vector with $\|\boldsymbol{v}\| = 1$ and θ the angle between $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$.

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

by the cosine formula for dot products and $\|\boldsymbol{v}\| = 1$. $\cos(\theta)$ is maximal if $\theta = 0$, hence if $\boldsymbol{v}$ and $\nabla f(\mathbf{x})$ point in the same direction.
(Alternative proof: Apply Cauchy-Schwarz to $\nabla f(\mathbf{x})^T \boldsymbol{v}$ and look for equality.)

Analogous: Negative gradient $-\nabla f(\mathbf{x})$ points in direction of greatest *de*crease

PROPERTIES OF THE GRADIENT / 2



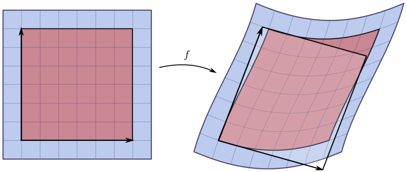
Length of arrows is norm of their gradient

JACOBIAN MATRIX

For vector-valued function $f = (f_1, \dots, f_m)^T$, $f_j : \mathcal{S} \to \mathbb{R}$, the **Jacobian** matrix $J_f : \mathcal{S} \to \mathbb{R}^{m \times d}$ generalizes gradient by placing all ∇f_j in its rows:

$$J_f(\mathbf{x}) = \begin{pmatrix} \nabla f_1(\mathbf{x})^T \\ \vdots \\ \nabla f_m(\mathbf{x})^T \end{pmatrix} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x})}{\partial x_d} \end{pmatrix}$$

- Jacobian gives best linear approximation of distorted volumes



Source: Wikipedia

JACOBIAN DETERMINANT

Let $f \in \mathcal{C}^1$ and $\mathbf{x}_0 \in \mathcal{S}$.

Inverse function theorem: Let $\mathbf{y}_0 = f(\mathbf{x}_0)$. If $\det(J_f(\mathbf{x}_0)) \neq 0$, then

1. f is invertible in a neighborhood of $\mathbf{x}_0$,
2. $f^{-1} \in \mathcal{C}^1$ with $J_{f^{-1}}(\mathbf{y}_0) = J_f(\mathbf{x}_0)^{-1}$.

- $|\det(J_f(\mathbf{x}_0))|$: factor by which f expands/shrinks volumes near $\mathbf{x}_0$
- If $\det(J_f(\mathbf{x}_0)) > 0$, f preserves orientation near $\mathbf{x}_0$
- If $\det(J_f(\mathbf{x}_0)) < 0$, f reverses orientation near $\mathbf{x}_0$

HESSIAN MATRIX

For real-valued function $f : \mathcal{S} \to \mathbb{R}$, the **Hessian** matrix $H : \mathcal{S} \to \mathbb{R}^{d \times d}$ contains all their second derivatives (if they exist):

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

Note: Hessian of f is Jacobian of ∇f

Example: Let $f(\mathbf{x}) = \sin(x_1) \cdot \cos(2x_2)$. Then:

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

- If $f \in \mathcal{C}^2$, then H is symmetric
- Many local properties (geometry, convexity, critical points) are encoded by the Hessian and its spectrum ($\rightarrow$ later)

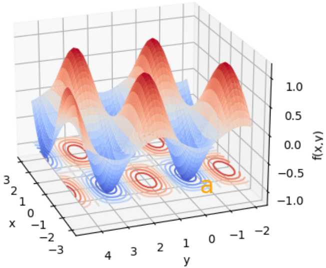
LOCAL CURVATURE BY HESSIAN

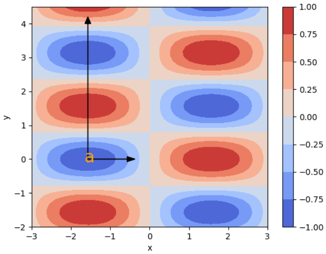
Eigenvector corresponding to largest (resp. smallest) **eigenvalue** of Hessian points in direction of largest (resp. smallest) **curvature**

Example (previous slide)**:** For $\boldsymbol{a} = (-\pi/2, 0)^T$, we have

$$H(\boldsymbol{a}) = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$$

and thus $\lambda_1 = 4$, $\lambda_2 = 1$, $\boldsymbol{v}_1 = (0,1)^T$, and $\boldsymbol{v}_2 = (1,0)^T$.

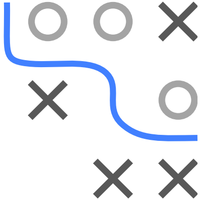




LIPSCHITZ CONTINUITY

Function $h : \mathcal{S} \rightarrow \mathbb{R}^m$ is **Lipschitz continuous** if slopes are bounded:

$$\|h(\mathbf{x}) - h(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\| \quad \text{for each } \mathbf{x}, \mathbf{y} \in \mathcal{S} \text{ and some } L > 0$$



- **Examples** ($d = m = 1$)**:** $\sin(x)$, $|x|$
- **Not** examples: $1/x$ (but *locally* Lipschitz continuous), $\sqrt{x}$
- If $m = d$ and h **differentiable**:

$$h \text{ Lipschitz continuous with constant } L \iff J_h \preccurlyeq L \cdot \mathbf{I}_d$$

Note: $\mathbf{A} \preccurlyeq \mathbf{B} :\iff \mathbf{B} - \mathbf{A}$ is positive semidefinite, i.e., $\mathbf{v}^T(\mathbf{B} - \mathbf{A})\mathbf{v} \geq 0 \;\; \forall \mathbf{v} \neq 0$

Proof of "$\Rightarrow$" for $d = m = 1$**:**

$$h'(x) = \lim_{\epsilon \to 0} \frac{h(x+\epsilon) - h(x)}{\epsilon} \leq \lim_{\epsilon \to 0} \underbrace{\left| \frac{h(x+\epsilon) - h(x)}{\epsilon} \right|}_{\leq L} \leq \lim_{\epsilon \to 0} L = L$$

[**Proof** of "$\Leftarrow$" by mean value theorem: Show that $\lambda_{\max}(J_h) \leq L$.]

LIPSCHITZ GRADIENTS

- Let $f \in \mathcal{C}^2$. Since $\nabla^2 f$ is Jacobian of $h = \nabla f$ $(m = d)$:

$$\nabla f \text{ Lipschitz continuous with constant } L \iff \nabla^2 f \preccurlyeq L \cdot \mathbf{I}_d$$

- Equivalently, eigenvalues of $\nabla^2 f$ are bounded by L
- **Interpretation:** Curvature in any direction is bounded by L
- Lipschitz gradients occur frequently in machine learning
  $\implies$ Fairly **weak assumption**
- Important for analysis of **gradient descent** optimization
  $\implies$ Descent lemma (later)