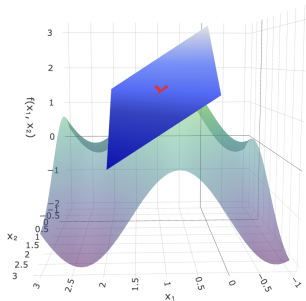


# Optimization in Machine Learning

## Mathematical Concepts

## Taylor Approximation

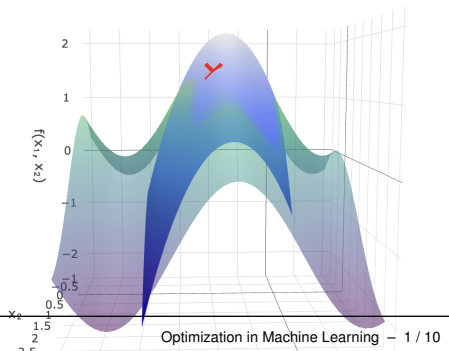
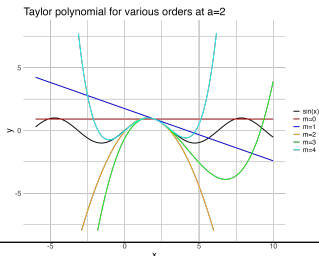
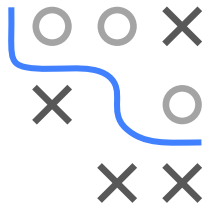


### Learning goals

- Taylor's theorem (univariate)
- Taylor series (univariate)
- Taylor's theorem (multivariate)
- Taylor series (multivariate)

# TAYLOR APPROXIMATIONS: OVERVIEW

- To optimize (find minima and maxima) it can be extremely helpful to approximate nonlinear functions locally
- We can use **Taylor polynomials** to approximate functions and
- **Taylor's theorem** provides us with the tools to estimate the error of this approximation  $\implies$  helpful for analyzing optimization algorithms
- Some functions can locally or even globally equal their **Taylor series**, i.e. the limit of Taylor polynomials



# TAYLOR APPROXIMATIONS: MOTIVATION

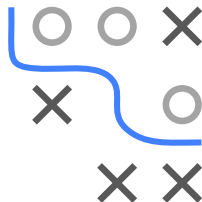
- Since the geometry of linear and quadratic functions is very well understood we will often want to use those for approximations
- For example, for a function  $f : \mathcal{S} \subseteq \mathbb{R}^d \rightarrow \mathbb{R}$

$$f(\mathbf{x} + \mathbf{h}) = f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}) \mathbf{h} + o(\mathbf{h})$$

- You might also often see an approximation via the gradient and Hessian of a function:

$$f(\mathbf{x} + \mathbf{h}) \approx f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}) \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \nabla_{\mathbf{x}}^2 f(\mathbf{x}) \mathbf{h}$$

- In fact,  $f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}) \mathbf{h}$  and  $f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x}) \mathbf{h} + \frac{1}{2} \mathbf{h}^\top \nabla_{\mathbf{x}}^2 f(\mathbf{x}) \mathbf{h}$  are, respectively, the first and second **Taylor polynomial of  $f$  at  $\mathbf{x}$** , evaluated at  $\mathbf{x} + \mathbf{h}$



# TAYLOR POLYNOMIALS

- Idea: Find a polynomial that locally behaves like a function  $f$  at point  $\mathbf{a}$ , i.e. matches  $f$ 's value ( $f$ ), slope ( $f'$ ), curvature ( $f''$ ), etc.
- $\Leftrightarrow$  Find polynomial so that

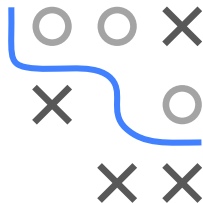
$$f(x) \approx T_k(x, \mathbf{a}) \quad \text{for all } x \text{ near } \mathbf{a}$$

where  $k$  denotes the highest order of derivative of  $f$  used in  $T_k$

- Wording: We “*expand  $f$  (via Taylor) around  $\mathbf{a}$* ”

**Definition of Taylor polynomial (univariate):** Let  $I \subseteq \mathbb{R}$  be an open interval and  $f \in \mathcal{C}^k(I, \mathbb{R})$ . For each  $a, x \in I$ , the  $k$ th order Taylor polynomial for  $f$  at  $a$  is defined as

$$T_k(x, a) := \sum_{j=0}^k \frac{f^{(j)}(a)}{j!} (x - a)^j$$



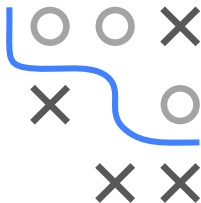
# MULTIVARIATE TAYLOR POLYNOMIALS

For the multivariate version, we need a concise way to express derivatives and powers involving several variables

- A **multi-index** is a vector  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_d) \in \mathbb{N}^d$ .
- Its **order** is the sum of its components:  $|\alpha| = \alpha_1 + \alpha_2 + \dots + \alpha_d$ .
- Partial derivative is written as  $D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$
- Factorials generalize componentwise:  $\alpha! = \alpha_1! \alpha_2! \dots \alpha_d!$ .
- For  $\mathbf{x}, \mathbf{a} \in \mathbb{R}^d$ :  $(\mathbf{x} - \mathbf{a})^\alpha = (x_1 - a_1)^{\alpha_1} \dots (x_d - a_d)^{\alpha_d}$ .

**Definition of Taylor polynomial (multivariate):** Let  $I$  be an open subset of  $\mathbb{R}^n$  and  $f \in \mathcal{C}^k(I, \mathbb{R})$ . For each  $\mathbf{a}, \mathbf{x} \in I$ , the  $k$ th order Taylor polynomial for  $f$  at  $\mathbf{a}$  is defined as

$$T_k(\mathbf{x}, \mathbf{a}) := \sum_{|\alpha| \leq k} \frac{D^\alpha f(\mathbf{a})}{\alpha!} (\mathbf{x} - \mathbf{a})^\alpha$$

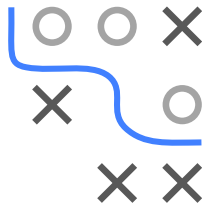


# MULTIVARIATE TAYLOR POLYNOMIAL IDENTITIES

For  $f \in \mathcal{C}^k(I, \mathbb{R})$  as before, we will often use the following identities:

- $T_1(\mathbf{x}, \mathbf{a}) = f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a})$
- $T_2(\mathbf{x}, \mathbf{a}) = f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T \mathbf{H}(\mathbf{a})(\mathbf{x} - \mathbf{a})$

(Which squares with the notation of the motivation slide by setting  $\mathbf{a} \hat{=} \mathbf{x}$  and  $\mathbf{x} \hat{=} \mathbf{x} + \mathbf{h}$ )



**Example: Specifically deriving  $T_1$  for bivariate  $f$  ( $d = 2$ )**

$\alpha_1$	$\alpha_2$	$ \alpha $	$D^\alpha f$	$\alpha!$	$(\mathbf{x} - \mathbf{a})^\alpha$
0	0	0	$f$	1	1
1	0	1	$\partial f / \partial x_1$	1	$x_1 - a_1$
0	1	1	$\partial f / \partial x_2$	1	$x_2 - a_2$

 and, therefore

$$\begin{aligned} T_1(\mathbf{x}, \mathbf{a}) &= \frac{f(\mathbf{a})}{1} \cdot 1 + \frac{\partial f(\mathbf{a})}{\partial x_1} (x_1 - a_1) + \frac{\partial f(\mathbf{a})}{\partial x_2} (x_2 - a_2) \\ &= f(\mathbf{a}) + \begin{pmatrix} \frac{\partial f(\mathbf{a})}{\partial x_1} \\ \frac{\partial f(\mathbf{a})}{\partial x_2} \end{pmatrix}^T \begin{pmatrix} x_1 - a_1 \\ x_2 - a_2 \end{pmatrix} = f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a}) \end{aligned}$$

# TAYLOR'S THEOREM

**General version for both univariate and multivariate functions:**

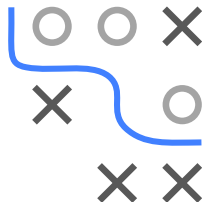
Let  $I$  be an open subset of  $\mathbb{R}^n$ ,  $n \in \mathbb{N}_{>0}$ , and  $f \in \mathcal{C}^k(I, \mathbb{R})$ . There exists a function  $R_k : I \times I \rightarrow \mathbb{R}$  so that for each  $\mathbf{a}, \mathbf{x} \in I$

$$R_k(\mathbf{x}, \mathbf{a}) = o(\|\mathbf{x} - \mathbf{a}\|^k) \text{ as } \mathbf{x} \rightarrow \mathbf{a}$$

and

$$f(\mathbf{x}) = T_k(\mathbf{x}, \mathbf{a}) + R_k(\mathbf{x}, \mathbf{a})$$

- $R_k(\mathbf{x}, \mathbf{a})$  is called **remainder term** and different specific forms have been established
- However, we will usually focus on the property  $R_k(\mathbf{x}, \mathbf{a}) = o(\|\mathbf{x} - \mathbf{a}\|^k)$  as  $\mathbf{x} \rightarrow \mathbf{a}$  or upper bounds derived for specific function classes when analyzing optimization algorithms



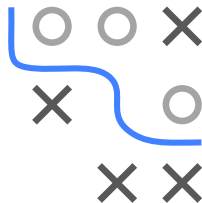
# TAYLOR SERIES

- For  $f \in \mathcal{C}^\infty$ , there might exist an open ball  $B_r(\mathbf{a})$  with radius  $r > 0$  around  $\mathbf{a}$  such that the **Taylor series**

$$T_\infty(\mathbf{x}, \mathbf{a}) = \begin{cases} \sum_{k=0}^{\infty} \frac{f^{(k)}(\mathbf{a})}{k!} (\mathbf{x} - \mathbf{a})^k & \text{if } f \text{ is univariate} \\ \sum_{|\alpha| \geq 0} \frac{D^\alpha f(\mathbf{a})}{\alpha!} (\mathbf{x} - \mathbf{a})^\alpha & \text{if } f \text{ is multivariate} \end{cases}$$

converges to  $f$  on  $B_r(\mathbf{a})$

- If such an open Ball exists for all  $\mathbf{a}$  in the domain of  $f$ ,  $f$  is called an **analytic function**
- Even if Taylor series converges, it might not converge to  $f$
- Upper bound  $R = \sup \{r \mid \text{Taylor series converges on } B_r(\mathbf{a})\}$  is called the radius of convergence of Taylor series around  $\mathbf{a}$
- If  $R > 0$  and  $f$  analytic, Taylor series converges absolutely and uniformly to  $f$  on compact sets inside  $B_R(\mathbf{a})$
- No general convergence behaviour on boundary of  $B_R(\mathbf{a})$





# EXAMPLES OF ANALYTIC FUNCTIONS

For analytic functions the remainder term eventually vanishes, i.e.

$R_k(\mathbf{x}, \mathbf{a}) \rightarrow 0$  as  $k \rightarrow \infty$  for all  $\mathbf{x} \in B_r(\mathbf{a})$ .

Important examples are

- Polynomials
- Exponential function (exp)
- Trigonometric functions (sin, cos)

And important rules are

- Any analytic function of a polynomial is again an analytic function
- Analytic functions are closed under sum and product (due to the properties of series)
- The derivative of an analytic function is again an analytic function

**One specific example:**  $f : \mathbb{R}^2 \longrightarrow \mathbb{R} \quad \mathbf{x} \mapsto \sin(2x_1) + \cos(x_2)$



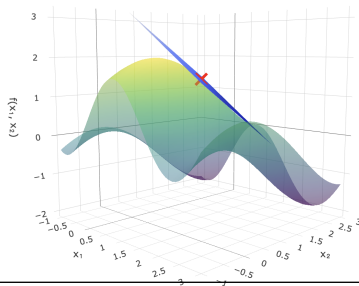
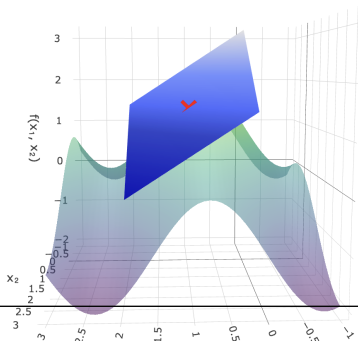
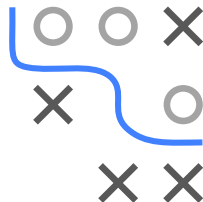
# EXAMPLE: TAYLOR APPROXIMATION OF

$f(\mathbf{x}) = \sin(2x_1) + \cos(x_2)$  AT  $\mathbf{a} = (1, 1)^T$

**1st order:** we know that  $f(\mathbf{x}) = \underbrace{f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a})}_{T_1(\mathbf{x}, \mathbf{a})} + R_1(\mathbf{x}, \mathbf{a})$  and since

$$\nabla f(\mathbf{x}) = (2 \cos(2x_1), -\sin(x_2)),$$

$$\begin{aligned} f(\mathbf{x}) &= T_1(\mathbf{x}) + R_1(\mathbf{x}, \mathbf{a}) = f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + R_1(\mathbf{x}, \mathbf{a}) \\ &= \sin(2) + \cos(1) + (2 \cos(2), -\sin(1)) \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} + R_1(\mathbf{x}, \mathbf{a}) \end{aligned}$$



$$f(\mathbf{x}) = \sin(2x_1) + \cos(x_2) \quad \mathbf{A}^T \mathbf{a} = (1, 1)^T$$

**2nd order:** we know that

$$f(\mathbf{x}) = \underbrace{f(\mathbf{a}) + \nabla f(\mathbf{a})(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T \mathbf{H}(\mathbf{a})(\mathbf{x} - \mathbf{a})}_{T_2(\mathbf{x}, \mathbf{a})} + R_2(\mathbf{x}, \mathbf{a})$$

and since  $H(\mathbf{x}) = \begin{pmatrix} -4 \sin(2x_1) & 0 \\ 0 & -\cos(x_2) \end{pmatrix}$ ,

$$f(\mathbf{x}) = T_1(\mathbf{x}, \mathbf{a}) + \frac{1}{2} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}^T \begin{pmatrix} -4 \sin(2) & 0 \\ 0 & -\cos(1) \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} + R_2(\mathbf{x}, \mathbf{a})$$

