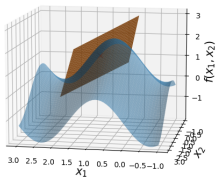


# Optimization in Machine Learning

## Mathematical Concepts

## Taylor Approximation

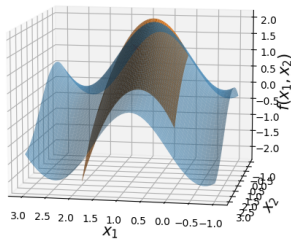
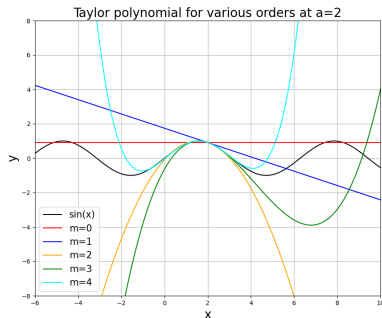


### Learning goals

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

# TAYLOR APPROXIMATIONS

- Mathematically fascinating: **Globally** approximate function by sum of polynomials determined by **local** properties
- Extremely important for **analyzing** optimization algorithms
- Geometry of **linear** and **quadratic** functions very well understood  
⇒ use them for **approximations**



# TAYLOR'S THEOREM (UNIVARIATE)

**Taylor's theorem:** Let  $I \subseteq \mathbb{R}$  be an open interval and  $f \in \mathcal{C}^k(I, \mathbb{R})$ . For each  $a, x \in I$ , it holds that

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

with the  $k$ -th **Taylor polynomial**  $T_k$  and a **remainder term**

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

- There are explicit formulas for the remainder
- Wording: We “expand  $f$  via Taylor around  $a$ ”



# TAYLOR SERIES (UNIVARIATE)

- If  $f \in C^\infty$ , it *might* be expandable around  $a \in I$  as a **Taylor series**

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x - a)^k$$

- If Taylor series converges to  $f$  in an interval  $I_0 \subseteq I$  centered at  $a$  (does not have to), we call  $f$  an *analytic function*
- Convergence if  $R_k(x, a) \rightarrow 0$  as  $k \rightarrow \infty$  for all  $x \in I_0$
- Then, for all  $x \in I_0$ :

$$f(x) = \sum_{j=0}^{\infty} \frac{f^{(j)}(a)}{j!} (x - a)^j$$



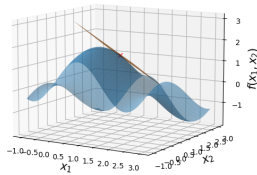
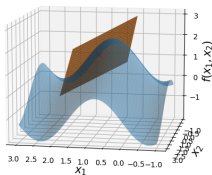
# TAYLOR'S THEOREM (MULTIVARIATE)

Taylor's theorem (1st order): For  $f \in \mathcal{C}^1$ , it holds that

$$f(\mathbf{x}) = \underbrace{f(\mathbf{a}) + \nabla f(\mathbf{a})^T (\mathbf{x} - \mathbf{a})}_{T_1(\mathbf{x}, \mathbf{a})} + R_1(\mathbf{x}, \mathbf{a}).$$

**Example:**  $f(\mathbf{x}) = \sin(2x_1) + \cos(x_2)$ ,  $\mathbf{a} = (1, 1)^T$ . Since  $\nabla f(\mathbf{x}) = \begin{pmatrix} 2 \cos(2x_1) \\ -\sin(x_2) \end{pmatrix}$ ,

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



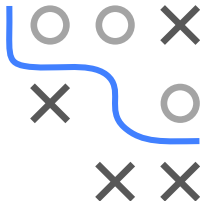
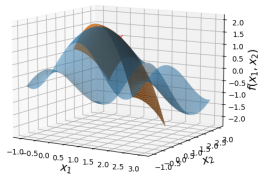
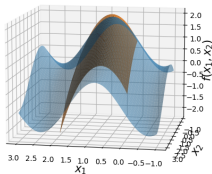
# TAYLOR'S THEOREM (MULTIVARIATE) / 2

Taylor's theorem (2nd order): If  $f \in \mathcal{C}^2$ , it holds that

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

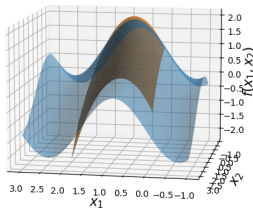
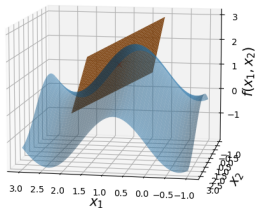
**Example (continued):** Since  $\mathbf{H}(\mathbf{x}) = \begin{pmatrix} -4 \sin(2x_1) & 0 \\ 0 & -\cos(x_2) \end{pmatrix}$ ,

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



# MULTIVARIATE TAYLOR APPROXIMATION

- Higher order  $k$  gives a better approximation
- $T_k(\mathbf{x}, \mathbf{a})$  is the best  $k$ -th order approximation to  $f(\mathbf{x})$  near  $\mathbf{a}$



Consider  $T_2(\mathbf{x}, \mathbf{a}) = f(\mathbf{a}) + \nabla f(\mathbf{a})^T(\mathbf{x} - \mathbf{a}) + \frac{1}{2}(\mathbf{x} - \mathbf{a})^T H(\mathbf{a})(\mathbf{x} - \mathbf{a})$ .  
The first/second/third term ensures the values/slopes/curvatures of  $T_2$  and  $f$  match at  $\mathbf{a}$ .

# TAYLOR'S THEOREM (MULTIVARIATE)

The theorem for general order  $k$  requires a more involved notation.

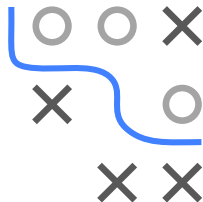
**Taylor's theorem ( $k$ -th order):** If  $f \in \mathcal{C}^k$ , it holds that

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

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

**Notation:** Multi-index  $\alpha \in \mathbb{N}^d$

- $|\alpha| = \alpha_1 + \dots + \alpha_d$
- $\mathbf{x}^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$
- $\alpha! = \alpha_1! \dots \alpha_d!$
- $D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$





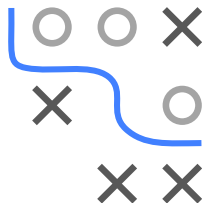
# TAYLOR'S THEOREM (MULTIVARIATE) / 2

Let us check for bivariate  $f$  ( $d = 2$ ). For  $|\alpha| \leq 1$ , we have

$\alpha_1$	$\alpha_2$	$ \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

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



# TAYLOR SERIES (MULTIVARIATE)

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

$$\sum_{|\alpha| \geq 0} \frac{D^\alpha f(\mathbf{a})}{\alpha!} (\mathbf{x} - \mathbf{a})^\alpha$$

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

- 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  $B_R(\mathbf{a})$

