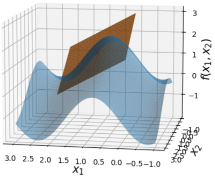
Optimization in Machine Learning

Mathematical Concepts: Taylor Approximations

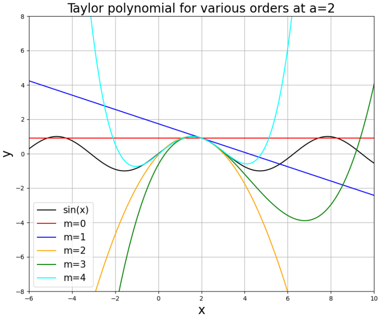


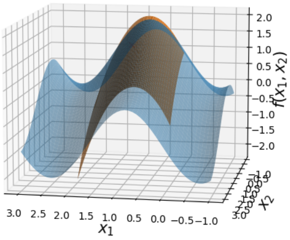
Learning goals

- 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
  $\Longrightarrow$ 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**

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

- 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 $x \in I_0$:

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

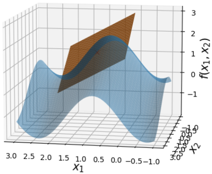
TAYLOR'S THEOREM (MULTIVARIATE)

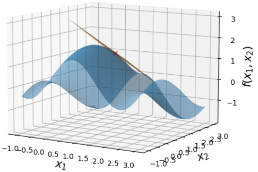
Taylor's theorem (1st order): For $f \in \mathcal{C}^1$, it holds that

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

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

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





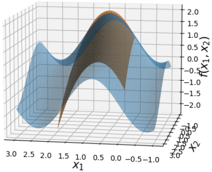
TAYLOR'S THEOREM (MULTIVARIATE)

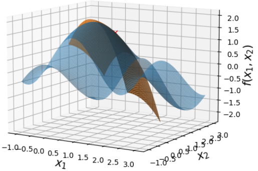
Taylor's theorem (2nd order): If $f \in \mathcal{C}^2$, it holds that

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

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

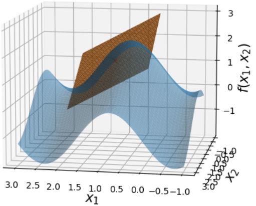
$$f(\mathbf{x}) = T_1(\mathbf{x}, \boldsymbol{a}) + \frac{1}{2}\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}^T \begin{pmatrix} -4\sin(2) & 0 \\ 0 & -\cos(1) \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} + R_2(\mathbf{x}, \boldsymbol{a})$$

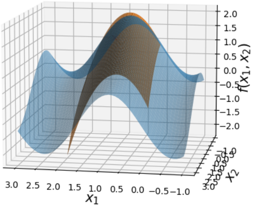




MULTIVARIATE TAYLOR APPROXIMATION

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





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

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

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

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

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

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

TAYLOR'S THEOREM (MULTIVARIATE)

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

α_1	α_2	$\lvert\boldsymbol{\alpha}\rvert$	$D^{\boldsymbol{\alpha}} f$	$\boldsymbol{\alpha}!$	$(\mathbf{x} - \boldsymbol{a})^{\boldsymbol{\alpha}}$
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

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

TAYLOR SERIES (MULTIVARIATE)

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

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

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