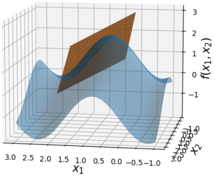
Optimization in Machine Learning

Mathematical Concepts: Taylor Approximations

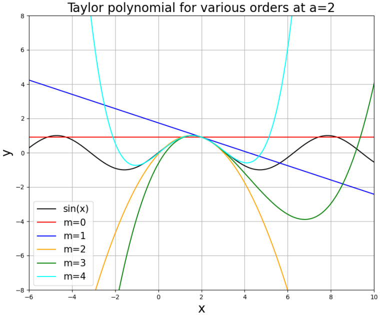


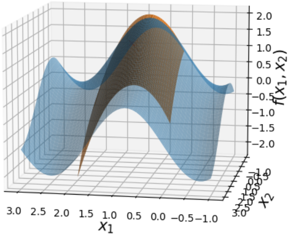
Learning goals

- Taylor Polynomials (Univariate)
- Taylor Series
- Taylor Polynomials (Multivariate)

TAYLOR APPROXIMATIONS

- Mathematically fascinating: We can approx a whole function via a sum of polynomials which are computed based only on properties of one local point.
- Extremely important in the analysis of optimization algorithms. We understand the geometry of linear and quadratic functions very well, so we often approx with them.



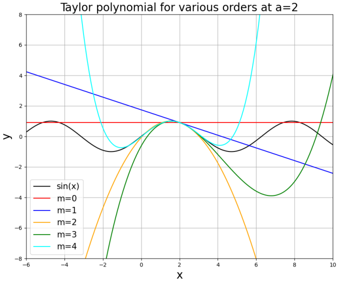


DEFINITION TAYLOR'S THEOREM (UNIVARIATE)

Let $I \subseteq \mathbb{R}$ an open interval and $a, x \in I$ and $f \in \mathcal{C}^{m+1}(I, \mathbb{R})$. Then

$$f(x) = T_m(x, a) + R_m(x, a), \text{with}$$

- We develop via Taylor around point a
- m-th **Taylor polynomial**: $T_m(x, a) \overset{(*)}{=} \sum_{k=0}^{m} \frac{f^{(k)}(a)}{k!}(x-a)^k$
- **Remainder term**: $R_m(x, a)$



$^{(*)}$ $T_m(x, a) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \ldots + \frac{f^{(m)}(a)}{m!}(x-a)^m$

TAYLOR SERIES

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

- If the Taylor series converges (does not have to) and it converges to f (does not have to), we call f an *analytic function*
- Convergence happens if $R_m(x, a) \to 0$ as $m \to \infty$ for all x
- Then:

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

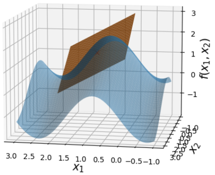
MULTIVARIATE TAYLOR POLYNOMIALS

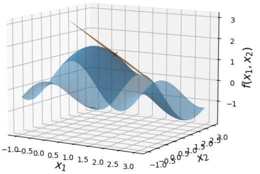
Taylor's theorem (1st order):

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

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

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





MULTIVARIATE TAYLOR POLYNOMIALS

Taylor's theorem (2nd order):

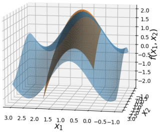
$$f(\mathbf{x}) = \underbrace{f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^\top(\mathbf{x} - \boldsymbol{a}) + \frac{1}{2}(\mathbf{x} - \boldsymbol{a})^\top \boldsymbol{H}(\boldsymbol{a})(\mathbf{x} - \boldsymbol{a})}_{T_2(\mathbf{x},\boldsymbol{a})} + R_2(\mathbf{x}, a)$$

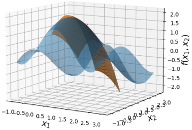
Example (continued): $f(\mathbf{x}) = \sin(2x_1) + \cos(x_2)$, $\boldsymbol{a} = (1,1)^\top$. Since

$$\nabla f(\mathbf{x}) = \begin{pmatrix} 2 \cdot \cos(2x_1) \\ -\sin(x_2) \end{pmatrix} \text{ and } H(\mathbf{x}) = \begin{pmatrix} -4\sin(2x_1) & 0 \\ 0 & -\cos(x_2) \end{pmatrix}$$

we get

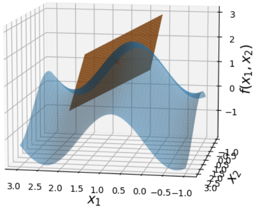
$$f(\mathbf{x}) \quad = \quad T_1(\mathbf{x}, \boldsymbol{a}) + \frac{1}{2}\begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix}^\top \begin{pmatrix} -4\sin(2) & 0 \\ 0 & -\cos(1) \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 1 \end{pmatrix} + R_2(\mathbf{x}, \boldsymbol{a})$$

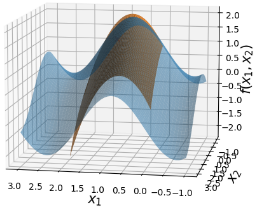




MULTIVARIATE TAYLOR APPROXIMATION

- Higher m gives a better approximation
- The m^{th} order Taylor term is the best m^{th} order approximation to $f(\mathbf{x})$ near $\boldsymbol{a}$





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

MULTIVARIATE TAYLOR POLYNOMIALS

What can be written down nicely for first and second order Taylor polynomial is (notationally) a bit more cumbersome for general k.

Let $f : \mathbb{R}^d \to \mathbb{R}$, $f \in \mathcal{C}^k$ at $\boldsymbol{a} \in \mathbb{R}^d$. Then

$$f(x) = T_m(\mathbf{x}, \boldsymbol{a}) + R_m(\mathbf{x}, \boldsymbol{a}), \text{ with}$$

$$T_m(\mathbf{x}, \boldsymbol{a}) = \sum_{|\boldsymbol{\alpha}| \leq k} \frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{a})}{\boldsymbol{\alpha}!} (\mathbf{x} - \boldsymbol{a})^{\boldsymbol{\alpha}} \text{ and } \lim_{\mathbf{x} \to \boldsymbol{a}} R_m(\mathbf{x}, \boldsymbol{a}) = 0$$

with $\boldsymbol{\alpha} \in \mathbb{N}^d$ and the multi-index notation

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

MULTIVARIATE TAYLOR POLYNOMIALS

Let's check for $f : \mathbb{R}^2 \to \mathbb{R}$ and $k = 1$. We have for $|\alpha| \leq 1$:

- $\alpha_1 = 0, \alpha_2 = 0$: $|\boldsymbol{\alpha}| = 0, \boldsymbol{\alpha}! = 1, \mathbf{x}^{\boldsymbol{\alpha}} = 1, D^{\boldsymbol{\alpha}} f = 1$
- $\alpha_1 = 1, \alpha_2 = 0$: $|\boldsymbol{\alpha}| = 1, \boldsymbol{\alpha}! = 1, \mathbf{x}^{\boldsymbol{\alpha}} = x_1, D^{\boldsymbol{\alpha}} f = \frac{\partial f}{\partial x_1}$
- $\alpha_1 = 0, \alpha_2 = 1$: $|\boldsymbol{\alpha}| = 1, \boldsymbol{\alpha}! = 1, \mathbf{x}^{\boldsymbol{\alpha}} = x_2, D^{\boldsymbol{\alpha}} f = \frac{\partial f}{\partial x_2}$

and therefore:

$$
\begin{aligned}
T_m(\mathbf{x}, \boldsymbol{a}) &= \sum_{|\boldsymbol{\alpha}| \leq k} \frac{D^{\boldsymbol{\alpha}} f(\boldsymbol{a})}{\boldsymbol{\alpha}!} (\mathbf{x} - \boldsymbol{a})^{\boldsymbol{\alpha}} \\
&= \frac{1 \cdot f(\boldsymbol{a})}{1} \cdot 1 + \frac{\partial f}{\partial x_1}(\boldsymbol{a})(x_1 - a_1) + \frac{\partial f}{\partial x_2}(\boldsymbol{a})(x_2 - a_2) \\
&= f(a) + \begin{pmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{a}) \\ \frac{\partial f}{\partial x_2}(\boldsymbol{a}) \end{pmatrix}^\top \begin{pmatrix} x_1 - a_1 \\ x_2 - a_2 \end{pmatrix} = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^\top (\mathbf{x} - \boldsymbol{a}).
\end{aligned}
$$