II. Calculus
2. multivariate Taylor expansion

Lecture based on

https://github.com/gwthomas/math4ml (Garrett Thomas, 2018)
https://mml-book.github.io/ (Deisenroth et al. 2020, Mathematics for Machine Learning)

Prof. Dr. Christoph Lippert
Digital Health & Machine Learning

Let $y = f(x_1, x_2, \ldots, x_n)$ be a **multivariate** function.

The **partial derivative** of y with respect to x_i is

$$\frac{\partial y}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_i, \ldots, x_n)}{h}.$$

- For notation of partial derivatives, the following are equivalent:

$$\frac{\partial y}{\partial x_i} = \frac{\partial f}{\partial x_i} = f_{x_i} = f_i = D_i f = D_{x_i} f.$$

Let $f : \mathbb{R}^n \to \mathbb{R}$.

The input vector is $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top$.

The **gradient** vector of $f(\mathbf{x})$ with respect to $\mathbf{x}$ is

$$D_{\mathbf{x}} f = \quad \nabla_{\mathbf{x}} f(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right]^\top,$$

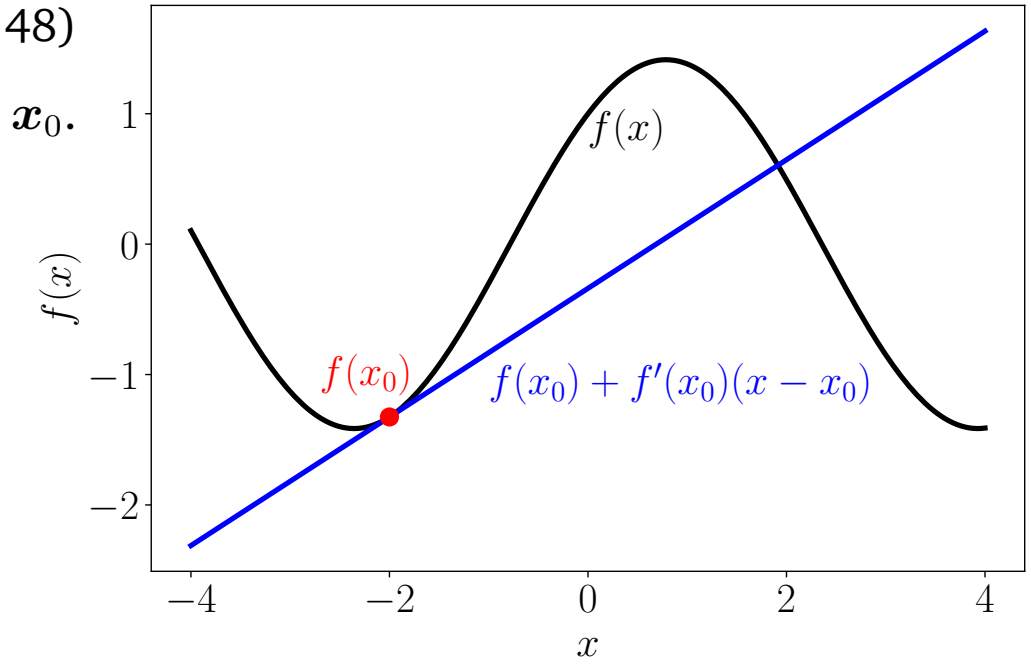
where $\nabla_{\mathbf{x}} f(\mathbf{x})$ is often replaced by $\nabla f(\mathbf{x})$ when there is no ambiguity.

Local Linearization of a multivariate function

The gradient ∇f of a function f is often used for a locally linear approximation of f around $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)^\top (\boldsymbol{x} - \boldsymbol{x}_0)\,. \tag{5.148}$$

Here $(\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)$ is the gradient of f with respect to $\boldsymbol{x}$, evaluated at $\boldsymbol{x}_0$.



Taylor Series for single variable functions

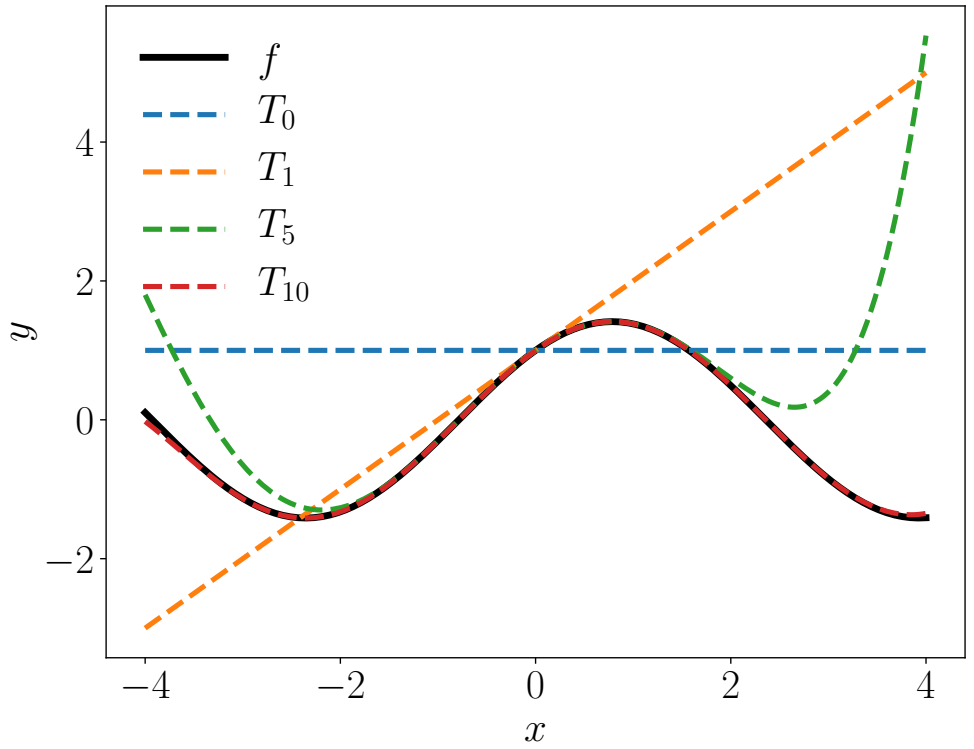
Definition 5.3 (Taylor Polynomial)**.** The *Taylor polynomial* of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k , \tag{5.7}$$

Definition 5.4 (Taylor Series)**.** For a smooth function $f \in \mathcal{C}^\infty$, $f : \mathbb{R} \to \mathbb{R}$, the *Taylor series* of f at x_0 is defined as

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k . \tag{5.8}$$

If $f(x) = T_\infty(x)$, then f is called *analytic*.



Higher order derivatives of multivariate functions

The **Hessian** matrix of $f : \mathbb{R}^d \to \mathbb{R}$ is a matrix of second-order partial derivatives:

$$D^2_{\mathbf{x}} f = \quad \nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix} \quad \text{i.e.} \quad [\nabla^2 f]_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$$

$$D^k_{\mathbf{x}} f$$

The k-th derivative Is a k-th order tensor of size $\underbrace{d \times \cdots \times d}_{k \text{ times}}$

Parabolic approximation of a function

If f is twice continuously differentiable, then

$$f(\mathbf{x}_0 + \mathbf{h}) \approx f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x}_0)\mathbf{h}$$

is a parabolic approximation to f that has the same Gradient and Hessian as f at $\mathbf{x}_0$.

Definition 5.7 (Multivariate Taylor Series)**.** We consider a function

$$f : \mathbb{R}^D \to \mathbb{R} \tag{5.149}$$

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^D, \tag{5.150}$$

that is smooth at $\boldsymbol{x}_0$. When we define the difference vector $\boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}_0$, the *multivariate Taylor series* of f at $(\boldsymbol{x}_0)$ is defined as

$$f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \frac{D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)}{k!} \boldsymbol{\delta}^k, \tag{5.151}$$

where $D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)$ is the k-th (total) derivative of f with respect to $\boldsymbol{x}$, evaluated at $\boldsymbol{x}_0$.

$$D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)\boldsymbol{\delta}^k = \sum_{i_1=1}^{D} \cdots \sum_{i_k=1}^{D} D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)[i_1, \ldots, i_k]\delta[i_1] \cdots \delta[i_k] \tag{5.155}$$

kth-order tensor $\boldsymbol{\delta}^k \in \mathbb{R}^{\overbrace{D \times D \times \ldots \times D}^{k \text{ times}}}$ is obtained as a k-fold outer product, denoted by $\otimes$, of the vector $\boldsymbol{\delta} \in \mathbb{R}^D$. For example,

$$\boldsymbol{\delta}^2 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\delta}^\top, \quad \boldsymbol{\delta}^2[i, j] = \delta[i]\delta[j] \tag{5.153}$$

Summary

- The Hessian is the matrix of second order partial derivatives
- We can approximate a function by an affine function using the first order Taylor expansion
- We can approximate a function by a parabolic using the second order Taylor expansion