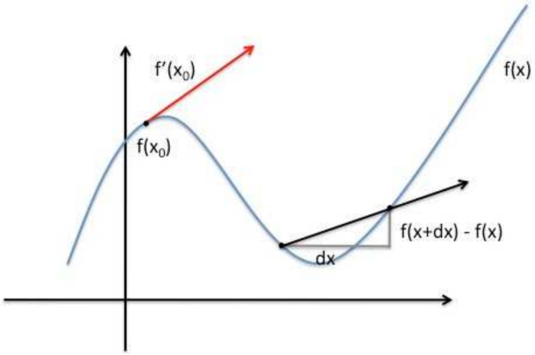
Multivariate Calculus[1]

Technische Universität München

[1]These slides contain material from David Barber's book *Bayesian Reasoning and Machine Learning*, see `www.cs.ucl.ac.uk/staff/D.Barber/brml` for more information.

1d case – Derivative

Consider a function $f : \mathbb{R} \to \mathbb{R}$. Its *derivative* $\frac{1}{dx} f(x) \equiv f'(x)$ can be though of as defining the slope of $f(x)$.



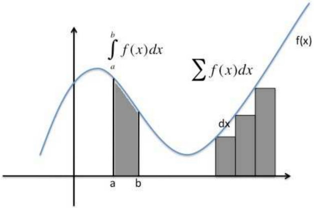
1d case – Integral

The integral can be considered the inverse operation of the derivative

$$\int f(x)dx = F(x) + c$$

$$F'(x) = f(x)$$

An integral resembles the calculation of the area under the curve defined by $f(x)$



Rules for Derivatives

- $g(x) = f_0(x) + f_1(x) \rightarrow g'(x) = f_0'(x) + f_1'(x)$
- $g(x) = f_0(x)f_1(x) \rightarrow g'(x) = f_0'(x)f_1(x) + f_0(x)f_1'(x)$
- $g(x) = \frac{f_0(x)}{f_1(x)} \rightarrow g'(x) = \frac{f_0'(x)f_1(x) - f_0(x)f_1'(x)}{f_1(x)^2}$
- $g(x) = f_1(f_0(x)) \rightarrow g'(x) = f_1'(f_0(x))f_0'(x)$

Rules for Integrals

Integration by parts

$$\int f(x)g'(x)dx = f(x)g(x) - \int f'(x)g(x)dx$$

Change of variable

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(t)dt$$

The multivariate case

Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ of n variables, $f(x_1, x_2, \ldots, x_n)$ or $f(\boldsymbol{x})$.
What is the derivative of f?

As it turns out with the appropriate definition of f' the summation, product, quotient and chain rule are still valid.

The gradient

The partial derivative of $f(\boldsymbol{x})$ with respect to x_i is defined as the following limit (when it exists)

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

For function f the *gradient* is denoted ∇f:

$$\nabla f(\boldsymbol{x}) \equiv \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}$$

The gradient points along the direction in which the function increases most rapidly. Why?

Interpreting the gradient vector

Consider a function $f(\boldsymbol{x})$ that depends on a vector $\boldsymbol{x}$. We are interested in how the function changes when the vector $\boldsymbol{x}$ changes by a small amount :

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + \sum_i \boldsymbol{\delta}_i \frac{f_i}{x_i} + O(\boldsymbol{\delta}^2)$$

We can interpret the summation above as the scalar product between the vector ∇f with components $[\nabla f]_i = \frac{f}{x_i}$ and $\boldsymbol{\delta}$

$$f(\boldsymbol{x}+\boldsymbol{\delta}) = f(\boldsymbol{x}) + (\nabla f)^T \boldsymbol{\delta} + O(\boldsymbol{\delta}^2)$$

Interpreting the gradient vector

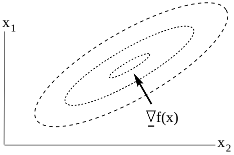


Figure : The ellipses are contours of constant function value, $f = c$ At any point $\boldsymbol{x}$, the gradient vector $\nabla f(\boldsymbol{x})$ points along the direction of maximal increase of the function.

Let $\boldsymbol{\delta} \equiv \delta \boldsymbol{p}$, with $\boldsymbol{p}$ a unit length vector and $\delta \ll 1$. Then

$$f(\boldsymbol{x} + \delta \boldsymbol{p}) \approx f(\boldsymbol{x}) + \delta \nabla f(\boldsymbol{x})^T \boldsymbol{p}$$

The scalar product is defined as (θ is the angle between $\nabla f(\boldsymbol{x})$ and $\boldsymbol{p}$)

$$\nabla f(\boldsymbol{x})^T \boldsymbol{p} = \mid \nabla f(\boldsymbol{x}) \mid\mid \boldsymbol{p} \mid \cos(\theta) = \mid \nabla f(\boldsymbol{x}) \mid \cos(\theta)$$

which is maximised if $\theta \equiv 0$. Hence, the direction along which the function changes the most rapidly is along $\nabla f(\boldsymbol{x})$.

Steepest gradient ascent

Given this nice interpretation of $\nabla f(\boldsymbol{x})$, maximizing $f(\boldsymbol{x})$ (i.e. find a $\boldsymbol{x'}$ such that $f(\boldsymbol{x'})$ is maximal, at least locally) can be done with the following *iterative* algorithm:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + \delta_n \nabla f(\boldsymbol{x}_n)$$

with $\boldsymbol{x}_0$ choosen arbitrarly. δ_n is denoted the *stepsize* and needs to be choosen small enough (otherwise the Taylor approximation does not hold).

The derivative of $f(\boldsymbol{x})$

Given the gradient $\nabla f(\boldsymbol{x})$ the derivative of $f(\boldsymbol{x})$, $f'(\boldsymbol{x})$, is

$$f'(\boldsymbol{x}) = \left(\nabla f(\boldsymbol{x})\right)^T$$

More generally, if $f : \mathbb{R}^n \to \mathbb{R}^m$ then the *Jacobian* of f is given by

$$\frac{\partial f}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

The chain rule

With the previous definition the summation rule and the chain rule generalize to arbitrary vector valued functions (the product rule and quotient rule only generalize to scalar functions f without any additional restrictions).

Let $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^m \to \mathbb{R}^n$, then

$$(f(g(\boldsymbol{x})))' = (\nabla f(g(\boldsymbol{x})))^T g'(\boldsymbol{x})$$