Function of Many Variables

(a crash course)

Dmytro Velychko

Theoretical Neuroscience lab
Philipps-Universität Marburg

Last edited: December 14, 2016

Essential basic elements

Multivariate calculus operates with:

- sets (with special properties)
- functional mappings (with special properties)

$f : \mathbb{R} \to \mathbb{R}$ - classical calculus of function of single variable

$f : \mathbb{R}^n \to \mathbb{R}^m$ - multivariable calculus

Domain and codomain

$f : \mathbb{R} \to \mathbb{R}$ - classical calculus of function of single variable

$f : \mathbb{R}^n \to \mathbb{R}$ - function defines a scalar field

$f : \mathbb{R}^n \to \mathbb{R}^m$ - function defines a vector field

Mappings of interest

We are interested in continuous functional mappings $f : \mathbb{R}^n \to \mathbb{R}^m$

$$\lim_{\mathbf{a}\to\mathbf{x}} f(\mathbf{a}) = f(\mathbf{x}) \tag{1}$$

which are also differentiable:

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a}) - Df_{\mathbf{a}} \cdot \mathbf{h}}{||\mathbf{h}||} = 0 \tag{2}$$

Here $Df_{\mathbf{a}} \cdot \mathbf{h}$ is a linear approximation of $f(\mathbf{a}+\mathbf{h}) - f(\mathbf{a})$.
$Df_{\mathbf{a}}$ is called the derivative of f at $\mathbf{a}$.

All the following properties follow from this idea of linear approximation (derivative)!

Directional derivative

But the $Df_{\mathbf{a}} \cdot \mathbf{h}$ approximation depends on both the length and the direction of $\mathbf{h}$ vector!
Let $f : (x_1, \ldots, x_n) \in \mathbb{R}^n \to \mathbb{R}$. Directional derivative of a scalar function along a vector $\mathbf{v}$ is defined as the limit:

$$\nabla_{\mathbf{v}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{v}) - f(\mathbf{x})}{h} \tag{3}$$

Partial derivative

Partial derivative is a special case of the directional derivative, where direction is taken along the axis of some variable.
Let $f : (x_1, \ldots, x_n) \in \mathbb{R}^n \to \mathbb{R}$. If we fix all x_j but x_i, we get a section of the function f:

$$f^* : (x_i) \in \mathbb{R} \to \mathbb{R} \tag{4}$$

We may take the derivative of this function:

$$\frac{df^*}{dx_i} : x_i \in \mathbb{R} \to \mathbb{R} \tag{5}$$

If we extend this derivative and make it a function of other variables, which were fixed previously, we get a partial derivative:

$$\frac{\partial f}{\partial x_i} : (x_1, \ldots, x_n) \in \mathbb{R}^n \to \mathbb{R} \tag{6}$$

which indicate the values of of the f derivative along the x_i direction.

Differentiability implies existence of partial derivatives

Theorem. Let U be an open subset of $\mathbb{R}^m$, $f : U \to \mathbb{R}^n, f(\mathbf{a}) = [f_1(\mathbf{a}), \ldots, f_n(\mathbf{a})]^T, \mathbf{a} \in U$. If f is differentiable at $\mathbf{a}$, then all of the partial derivatives exist at $\mathbf{a}$, and

$$\frac{\partial f(\mathbf{a})}{\partial x_i} = Df_{\mathbf{a}} \cdot \mathbf{e}_i \tag{7}$$

Jacobian matrix

Corollary. The derivative of a multivariable function differentiable at $\mathbf{a}$ is a matrix comprised of partial derivatives:

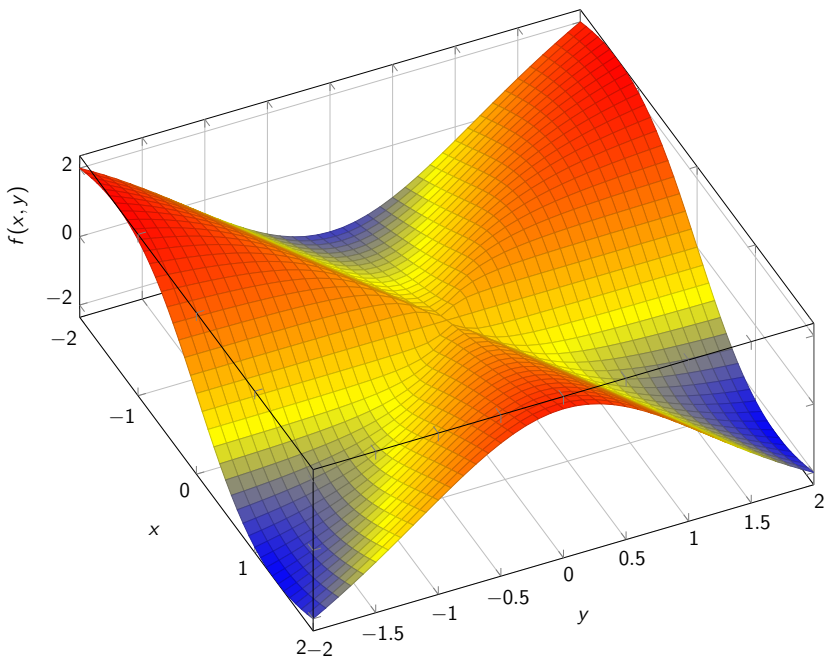
$$Df_{\mathbf{a}} = \begin{bmatrix} \frac{\partial f_1(\mathbf{a})}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{a})}{\partial x_m} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n(\mathbf{a})}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{a})}{\partial x_m} \end{bmatrix} \tag{8}$$

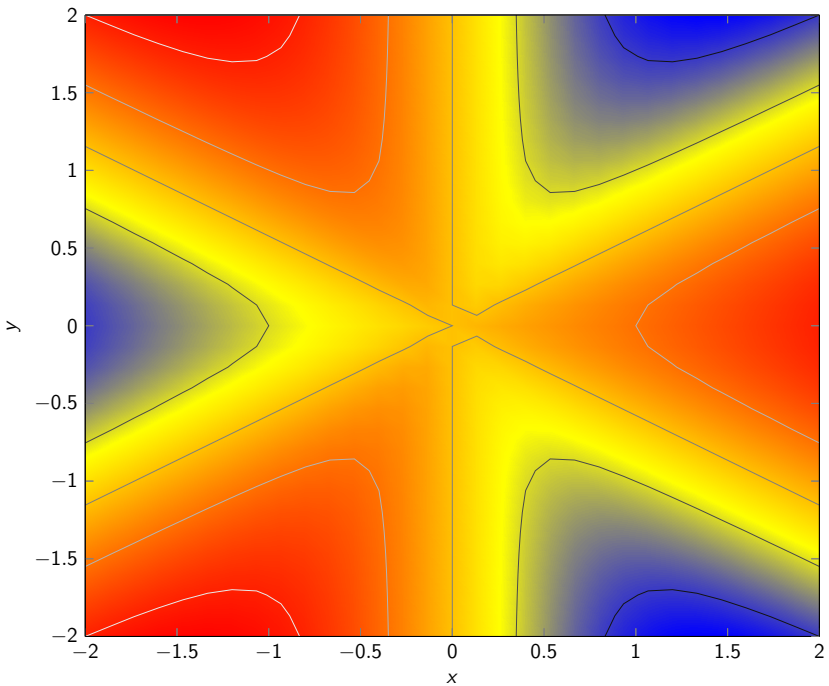
We call such matrix the **Jacobian matrix**. It may exists even for non-differential points.

Non-differentiable function example

Example. Here is a non-differentiable function at $\mathbf{a} = (0, 0)$, but partial derivatives exist at $\mathbf{a}$:

$$f(x, y) = \sqrt{x^2 + y^2} cos(3arccos(\frac{x}{\sqrt{x^2 + y^2}})) = \frac{x^3 - 3xy^2}{x^2 + y^2} \quad (9)$$





Check that partial derivatives are not continuous at $\mathbf{a} = (0, 0)$

Differentiability condition

Theorem. Multivariable function f is differentiable at $\mathbf{a}$ iff all partial derivatives $\partial f_i/\partial x_j$ exist and are continuous at the point $\mathbf{a}$.

Gradient

When $f : \mathbb{R}^n \to \mathbb{R}$ is a scalar field, the derivative is a vector comprised of partial derivatives w.r.t each variable:

$$\nabla f(x_1, \ldots, x_n) = (Df_{\mathbf{a}})^T = \left[\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n} \right]^T \tag{10}$$

With the gradient we can get the linear approximation of a scalar field function:

$$f(\mathbf{x}_0 + \delta\mathbf{x}) \approx f(\mathbf{x}_0) + (\nabla f)(\mathbf{x}_0) \cdot \delta\mathbf{x} \tag{11}$$

Which $\delta\mathbf{x}$ maximizes the $(\nabla f)(\mathbf{x}_0) \cdot \delta\mathbf{x}$?

Chain rule and Jacobian

Chain rule for composed functions

Let $g : t \in \mathbb{R} \rightarrow \mathbf{x} \in \mathbb{R}^n, \mathbf{x} = g(t), f : \mathbf{x} \in \mathbb{R}^n \rightarrow \mathbb{R}, f = f(\mathbf{x}) = f(g(t))$, then the total derivative of f is:

$$\frac{df(t)}{dt} = \frac{\partial f(g(t))}{\partial g(t)} \frac{\partial g(t)}{\partial t} = \nabla f(g(t)) \cdot J_g(t) = \sum_{i=1}^{n} \frac{\partial f(g(t))}{\partial x_i} \frac{dx_i}{dt} \tag{12}$$

Chain rule for composed functions

Example. Let $g(t) = [t^2, t^3]^T, f(g(t)) = 3g_1 + 2g_2^2 + 4$. Find the derivative df/dt. First get the partial derivatives and gradients:

$$\nabla f(g) = \begin{bmatrix} \frac{\partial f}{\partial g_1} \\ \frac{\partial f}{\partial g_2} \end{bmatrix} = \begin{bmatrix} 3 \\ 4g_2 \end{bmatrix} \tag{13}$$

$$J_g(t) = \frac{dg}{dt} = \begin{bmatrix} 2t \\ 3t^2 \end{bmatrix} \tag{14}$$

$$\frac{df}{dt} = \nabla f(g) \cdot J_g(t) = 3 * 2t + 4g_2 * 3t^2 = 6t + 4t^3 * 3t^2 = 12t^5 + 6t \tag{15}$$

Chain rule

In general case, for $g : \mathbf{t} \in \mathbb{R}^m \to \mathbf{x} \in \mathbb{R}^n, \mathbf{x} = g(\mathbf{t}), f : \mathbf{x} \in \mathbb{R}^n \to \mathbb{R}^k, f = f(\mathbf{x}) = f(g(\mathbf{t}))$, the derivative of the composed function f is:

$$Df_{\mathbf{t}} = J_f(g(t)) \cdot J_g(t) \tag{16}$$

Hessian and Critical points of a cost function

Hessian

For a continuous differentiable $f : \mathbb{R}^n \to \mathbb{R}$ first-order Taylor expansion at a point gave us the definition of the derivative of the multivariable function. Second-order Taylor series expansion adds quadratic terms:

$$T^2_{\mathbf{x}_0} = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0) \cdot (\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T (H(f)(\mathbf{x}_0))(\mathbf{x} - \mathbf{x}_0) \tag{17}$$

The matrix $Hf(\mathbf{x}_0)$ represents all quadratic coefficients and basically comprises second order partial derivatives:

$$H(f)_{i,j}(\mathbf{x}) = \frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j} \tag{18}$$

Hessian

Hessian matrix is symmetric, $H_{i,j} = H_{j,i}$. As it is a matrix of second order partial derivatives, it can be expressed as a Jacobian of gradient:

$$H(f)(\mathbf{x}) = J(\nabla f)(\mathbf{x}) \tag{19}$$

Critical points

To find critical points of a continuous differentiable $f : \mathbb{R}^n \to \mathbb{R}$, one has to find the points, where the gradient vanishes:

$$\nabla f(\mathbf{x}) = 0 \tag{20}$$

To classify, whether some critical point is a max, min, or of other type (saddle point), eigenvalues of the Hessian matrix have to be found:

- all eigenvalues are positive $\to$ the critical point is a minimum
- all eigenvalues are negative $\to$ the critical point is a maximum
- some eigenvalues are positive and some are negative $\to$ the critical point is a saddle point

Convex functions

Convex functions

Definition. Set S in some vector space V is called convex if the line segment connecting any two points in S lies entirely in S:

$$\forall \mathbf{s}_1, \mathbf{s}_2 \in S, \forall \alpha \in [0 \ldots 1] : \alpha \mathbf{s}_1 + (1 - \alpha)\mathbf{s}_2 \in S \qquad (21)$$

Convex functions

Definition. Function $f : \mathbb{R}^N \to \mathbb{R}$ is called convex if its set of points above the graph is convex. *Alternatively:* function is convex if the line segment connecting any two points of the function lies above the function:

$$\forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N, \forall \alpha \in [0 \dots 1] : \tag{22}$$

$$f(\alpha \mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha) f(\mathbf{x}_2) \tag{23}$$

Correspondingly, if $f(\mathbf{x})$ is convex, then $-f(\mathbf{x})$ is a concave function.
Some examples of convex functions:

- $f(x) = x^2$, $f(x) = x^2 n$
- $f(x) = e^x$

Convexity test

One of the tests on function convexity is positive-(semi)definiteness of its Hessian:

$$\forall \mathbf{x} \in X : \mathbf{x}^T H_f(\mathbf{x}) \mathbf{x} \geq 0 \tag{24}$$

Convexity is good property in optimization

Why are convex functions so special?

- it has no local minima; if it has a minimum - then it is global
- if it has multiple minima - then the set of all minima is convex

This gives guarantees in optimization results.

Jensen's inequality

Form the convex function definition inequality (22) holds for any two points. This inequality remains valid if we take the sum of any number of points weighted to 1:

$$\forall \mathbf{x}_1 \dots \mathbf{x}_n \in \mathbb{R}^N, \forall \alpha_1 \dots \alpha_n \in [0 \dots 1] | \sum_{i=1 \dots n} \alpha_i = 1 : \tag{25}$$

$$f(\sum_{i=1 \dots n} \alpha_i \mathbf{x}_i) \leq \sum_{i=1 \dots n} \alpha_i f(\mathbf{x}_i) \tag{26}$$

This inequality can be proved by induction employing the basic case (22).

The Rosenbrock function - the guinea pig of continuous optimization

Exercise. The Rosenbrock function (a.k.a Rosenbrock's banana function) is defined on $\mathbb{R}^2$ by:

$$f(x, y) = (a - x)^2 + b(y - x^2)^2 \tag{27}$$

The coefficients are usually set to $a = 1, b = 100$. Find the Jacobian and the Hessian matrices. Find the critical point(s). Test various numerical optimization routines on this function.