TOPIC 3

Optimality, convexity, and gradient descent

1 Background: Jacobian and Hessian; introduction to automatic differentiation

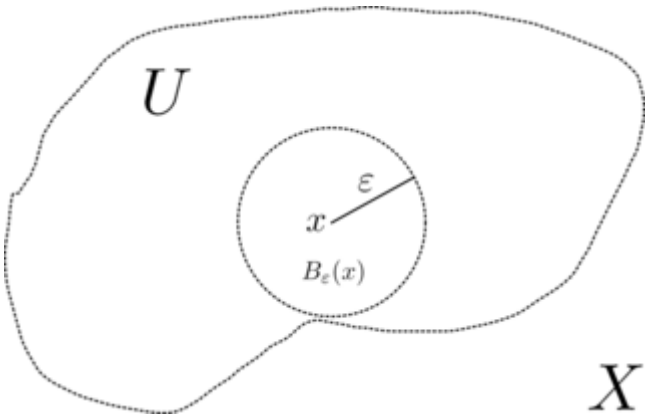
Course: Math 535 (http://www.math.wisc.edu/~roch/mmids/) - Mathematical Methods in Data Science (MMiDS)
Author: Sebastien Roch (http://www.math.wisc.edu/~roch/), Department of Mathematics, University of Wisconsin-Madison
Updated: Oct 10, 2020

We further review the differential calculus of several variables. We highlight a few key results that will play an important role: the Chain Rule and the Mean Value Theorem. We also give a brief introduction to automatic differentiation; we will come back later in this topic to the mathematical theory behind it.

1.1 Mean value theorem

Recalll that a point $\mathbf{x} \in \mathbb{R}^d$ is an interior point of a set $A \subseteq \mathbb{R}^d$ if there exists an $r > 0$ such that $B_r(\mathbf{x}) \subseteq A$, where the open r-ball around $\mathbf{x} \in \mathbb{R}^d$ is the set of points within Euclidean distance r of $\mathbf{x}$, that is,

$$B_r(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^d \,:\, \|\mathbf{y} - \mathbf{x}\| < r\}.$$



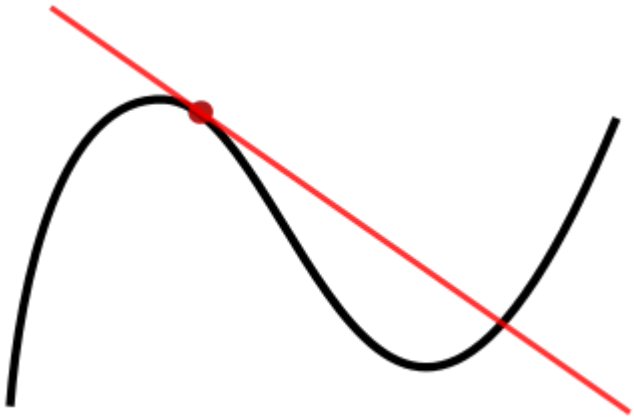
(Source (https://commons.wikimedia.org/wiki/File:Open_set_-_example.png))

Recall that the derivative of a function of a real variable is the rate of change of the function with respect to the change in the variable.

Definition (Derivative): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}$ and let $x_0 \in D$ be an interior point of D. The derivative of f at x_0 is

$$f'(x_0) = \frac{\mathrm{d}f(x_0)}{\mathrm{d}x} = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}$$

provided the limit exists. ⊲



(Source (https://commons.wikimedia.org/wiki/File:Tangent_to_a_curve.svg))

Earlier in the course, we proved a key insight about the derivative of f at x_0: it tells us where to find smaller values.

Lemma (Descent Direction): Let $f : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}$ and let $x_0 \in D$ be an interior point of D where $f'(x_0)$ exists. If $f'(x_0) > 0$, then there is an open ball $B_\delta(x_0) \subseteq D$ around x_0 such that for each x in $B_\delta(x_0)$:

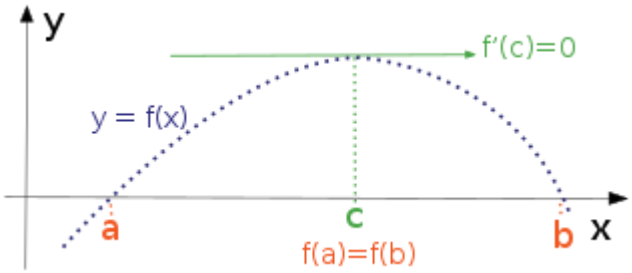
(a) $f(x) > f(x_0)$ if $x > x_0$, (b) $f(x) < f(x_0)$ if $x < x_0$.

If instead $f'(x_0) < 0$, the opposite holds.

One implication of the *Descent Direction Lemma* is the *Mean Value Theorem*, which will lead us later to *Taylor's Theorem*. First, an important special case:

Theorem (Rolle): Let $f : [a, b] \to \mathbb{R}$ be a continuous function and assume that its derivative exists on (a, b). If $f(a) = f(b)$ then there is $a < c < b$ such that $f'(c) = 0$.

Proof idea: Look at an extremum and use the *Descent Direction Lemma* to get a contradiction.



(Source (https://commons.wikimedia.org/wiki/File:RTCalc.svg))

Proof: If $f(x) = f(a)$ for all $x \in (a, b)$, then $f'(x) = 0$ on (a, b) and we are done. So assume there is $y \in (a, b)$ such that $f(y) \neq f(a)$. Assume without loss of generality that $f(y) > f(a)$ (otherwise consider the function $-f$). By the *Extreme Value Theorem*, f attains a maximum value at some $c \in [a, b]$. By our assumption, a and b cannot be the location of the maximum and it must be that $c \in (a, b)$.

We claim that $f'(c) = 0$. We argue by contradiction. Suppose $f'(c) > 0$. By the *Descent Direction Lemma*, there is a $\delta > 0$ such that $f(x) > f(c)$ for all $x \in B_\delta(c)$, a contradiction. A similar argument holds if $f'(c) < 0$. That concludes the proof. □

Theorem (Mean Value): Let $f : [a, b] \to \mathbb{R}$ be a continuous function and assume that its derivative exists on (a, b). Then there is $a < c < b$ such that

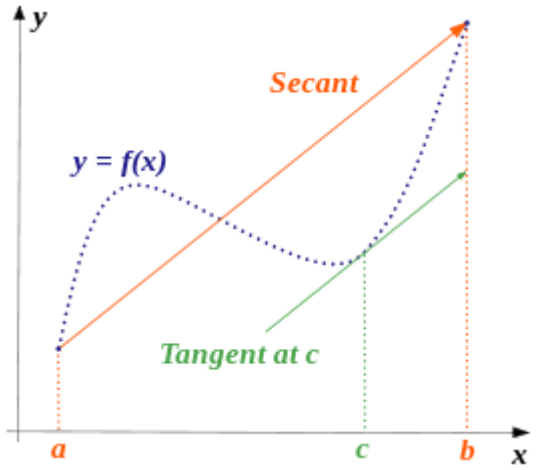
$$f(b) = f(a) + (b - a)f'(c),$$

or put differently

$$\frac{f(b) - f(a)}{b - a} = f'(c).$$

Proof idea: Apply *Rolle* to

$$\phi(x) = f(x) - \left[f(a) + \frac{f(b) - f(a)}{b - a}(x - a)\right].$$



(Source (https://commons.wikimedia.org/wiki/File:Mvt2.svg))

Proof: Let $\phi(x) = f(x) - f(a) - \frac{f(b)-f(a)}{b-a}(x - a)$. Note that $\phi(a) = \phi(b) = 0$ and $\phi'(x) = f'(x) - \frac{f(b)-f(a)}{b-a}$ for all $x \in (a, b)$. Thus, by *Rolle*, there is $c \in (a, b)$ such that $\phi'(c) = 0$. That implies $\frac{f(b)-f(a)}{b-a} = \phi'(c)$ and plugging into $\phi(b)$ gives the result. □

1.2 Jacobian

Recall the partial derivative and the gradient:

Definition (Partial Derivative): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D. The partial derivative of f at $\mathbf{x}_0$ with respect to x_i is

$$\frac{\partial f(\mathbf{x}_0)}{\partial x_i} = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h}$$

provided the limit exists. If $\frac{\partial f(\mathbf{x}_0)}{\partial x_i}$ exists and is continuous in an open ball around $\mathbf{x}_0$ for all i, then we say that f continuously differentiable at $\mathbf{x}_0$. ◁

Definition (Gradient): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D. Assume f is continuously differentiable at $\mathbf{x}_0$. The vector

$$\nabla f(\mathbf{x}_0) = \left(\frac{\partial f(\mathbf{x}_0)}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x}_0)}{\partial x_d}\right)^T$$

is called the gradient of f at $\mathbf{x}_0$. ◁

For vector-valued functions, we have the following generalization.

Definition (Jacobian): Let $\mathbf{f} = (f_1, \ldots, f_m) : D \to \mathbb{R}^m$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D where $\frac{\partial f_j(\mathbf{x}_0)}{\partial x_i}$ exists for all i, j. The Jacobian of $\mathbf{f}$ at $\mathbf{x}_0$ is the $d \times m$ matrix

$$\mathbf{J}_{\mathbf{f}}(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_m(\mathbf{x}_0)}{\partial x_d} \end{pmatrix}.$$

For a real-valued function $f : D \to \mathbb{R}$, the Jacobian reduces to the row vector

$$\mathbf{J}_f(\mathbf{x}_0) = \nabla f(\mathbf{x}_0)^T$$

where $\nabla f(\mathbf{x}_0)$ is the gradient of f at $\mathbf{x}_0$. ◁

Functions are often obtained from the composition of simpler ones. We will use the standard notation $h = g \circ f$ for the function $h(\mathbf{x}) = g(f(\mathbf{x}))$.

Lemma (Composition of Continuous Functions): Let $\mathbf{f} : D_1 \to \mathbb{R}^m$, where $D_1 \subseteq \mathbb{R}^d$, and let $\mathbf{g} : D_2 \to \mathbb{R}^p$, where $D_2 \subseteq \mathbb{R}^m$. Assume that $\mathbf{f}$ is continuous at $\mathbf{x}_0$ and that $\mathbf{g}$ is continuous $\mathbf{f}(\mathbf{x}_0)$. Then $g \circ f$ is continuous at $\mathbf{x}_0$.

Exercise: Prove the Composition of Continuous Functions Lemma. ◁

The chain rule gives a formula for the Jacobian of a composition. We will use the vector notation $\mathbf{h} = \mathbf{g} \circ \mathbf{f}$ for the function $\mathbf{h}(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x}))$.

Theorem (Chain Rule): Let $\mathbf{f} : D_1 \to \mathbb{R}^m$, where $D_1 \subseteq \mathbb{R}^d$, and let $\mathbf{g} : D_2 \to \mathbb{R}^p$, where $D_2 \subseteq \mathbb{R}^m$. Assume that $\mathbf{f}$ is continuously differentiable at $\mathbf{x}_0$, an interior point of D_1, and that $\mathbf{g}$ is continuously differentiable at $\mathbf{f}(\mathbf{x}_0)$, an interior point of D_2. Then

$$\mathbf{J}_{\mathbf{g}\circ\mathbf{f}}(\mathbf{x}_0) = \mathbf{J}_{\mathbf{g}}(\mathbf{f}(\mathbf{x}_0))\, \mathbf{J}_{\mathbf{f}}(\mathbf{x}_0)$$

as a product of matrices.

Proof: To simplify the notation, we begin with a special case. Suppose that f is a real-valued function of $\mathbf{x} = (x_1, \ldots, x_m)$ whose components are themselves functions of $t \in \mathbb{R}$. Assume f is continuously differentiable at $\mathbf{x}(t)$. To compute the [total derivative (https://en.wikipedia.org/wiki/Total_derivative)] $\frac{\mathrm{d}f(t)}{\mathrm{d}t}$, let $\Delta x_k = x_k(t + \Delta t) - x_k(t)$, $x_k = x_k(t)$ and

$$\Delta f = f(x_1 + \Delta x_1, \ldots, x_m + \Delta x_m) - f(x_1, \ldots, x_m).$$

We seek to compute the limit $\lim_{\Delta t \to 0} \frac{\Delta f}{\Delta t}$. To relate this limit to partial derivatives of f, we re-write Δf as a telescoping sum where each term involves variation of a single variable x_k. That is,

$$\begin{aligned}\Delta f =& [f(x_1 + \Delta x_1, \ldots, x_m + \Delta x_m) - f(x_1, x_2 + \Delta x_2, \ldots, x_m + \Delta x_m)] \\ &+ [f(x_1, x_2 + \Delta x_2, \ldots, x_m + \Delta x_m) - f(x_1, x_2, x_3 + \Delta x_3, \ldots, x_m + \Delta x_m)] \\ &+ \cdots + [f(x_1, \cdots, x_{m-1}, x_m + \Delta x_m) - f(x_1, \ldots, x_m)].\end{aligned}$$

Applying the *Mean Value Theorem* to each term gives

$$\begin{aligned}\Delta f =& \Delta x_1 \frac{\partial f(x_1 + \theta_1 \Delta x_1, x_2 + \Delta x_2, \ldots, x_m + \Delta x_m)}{\partial x_1} \\ &+ \Delta x_2 \frac{\partial f(x_1, x_2 + \theta_2 \Delta x_2, x_3 + \Delta x_3, \ldots, x_m + \Delta x_m)}{\partial x_2} \\ &+ \cdots + \Delta x_m \frac{\partial f(x_1, \cdots, x_{m-1}, x_m + \theta_m \Delta x_m)}{\partial x_m}\end{aligned}$$

where $0 < \theta_k < 1$ for $k = 1, \ldots, m$. Dividing by Δt, taking the limit $\Delta t \to 0$ and using the fact that f is continuously differentiable, we get

$$\frac{\mathrm{d}f(t)}{\mathrm{d}t} = \sum_{k=1}^{m} \frac{\partial f(\mathbf{x}(t))}{\partial x_k} \frac{\mathrm{d}x_k(t)}{\mathrm{d}t}.$$

Going back to the general case, the same argument shows that

$$\frac{\partial h_i(\mathbf{x}_0)}{\partial x_j} = \sum_{k=1}^{m} \frac{\partial g_i(\mathbf{f}(\mathbf{x}_0))}{\partial f_k} \frac{\partial f_k(\mathbf{x}_0)}{\partial x_j}$$

where the notation $\frac{\partial g}{\partial f_k}$ indicates the partial derivative of g with respect to its k-th component. In matrix form, the claim follows. □

NUMERICAL CORNER We illustrate the use of [automatic differentiation (https://en.wikipedia.org/wiki/Automatic_differentiation)] to compute gradients.

Quoting Wikipedia (https://en.wikipedia.org/wiki/Automatic_differentiation):

> In mathematics and computer algebra, automatic differentiation (AD), also called algorithmic differentiation or computational differentiation, is a set of techniques to numerically evaluate the derivative of a function specified by a computer program. AD exploits the fact that every computer program, no matter how complicated, executes a sequence of elementary arithmetic operations (addition, subtraction, multiplication, division, etc.) and elementary functions (exp, log, sin, cos, etc.). By applying the chain rule repeatedly to these operations, derivatives of arbitrary order can be computed automatically, accurately to working precision, and using at most a small constant factor more arithmetic operations than the original program. Automatic differentiation is distinct from symbolic differentiation and numerical differentiation (the method of finite differences). Symbolic differentiation can lead to inefficient code and faces the difficulty of converting a computer program into a single expression, while numerical differentiation can introduce round-off errors in the discretization process and cancellation.

We will use the Flux.jl (https://github.com/FluxML/Flux.jl) package. The `gradient` function takes another Julia function `f` and a set of arguments, and returns the gradient with respect to each argument. Here is an example.

In [1]:
```
# Julia version: 1.5.1
ENV["JULIA_CUDA_SILENT"] = true # silences warning about GPUs

using Flux
```

In [2]:
```
f(x, y) = 3x^2 + exp(x) + y;
```

In [3]:
```
df(x, y) = gradient(f, x, y);
```

In [4]:
```
df(1., 2.) # answer is (6 + e, 1)
```

Out[4]:
```
(8.718281828459045, 1.0)
```

We get the same answer by setting each parameter in the call to `gradient`.

In [5]:
```
gradient(f, 1., 2.)[1]
```

Out[5]:
```
8.718281828459045
```

The input parameters can also be vectors, which allows to consider function of large numbers of variables.

In [6]:
```
g(z) = sum(z.^2);
```

In [7]:
```
gradient(g, [1., 2., 3.])[1] # gradient is (2 z_1, 2 z_2, 2 z_3)
```

Out[7]:
```
3-element Array{Float64,1}:
 2.0
 4.0
 6.0
```

Finally, the function `params` can also be used to specify the parameters with respect to which derivatives are to be taken. Here is a typical example.

In [8]:
```
predict(X) = X*theta # classifier with parameter vector θ
loss(X, y) = sum((predict(X) - y).^2) # loss function
```

Out[8]:
```
loss (generic function with 1 method)
```

In [9]:
```
(X, y) = (randn(3, 2), [1., 0., 1.]); # dataset (features, labels)
theta = ones(2) # parameter assignment
gradient(() -> loss(X, y), params(theta))[theta]
```

Out[9]:
```
2-element Array{Float64,1}:
 1.9560808628127127
 1.6881129851496863
```

1.3 Further derivatives

Partial derivatives measure the rate of change of a function along the axes. More generally:

Definition (Directional Derivative): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$, let $\mathbf{x}_0 \in D$ be an interior point of D and let $\mathbf{v} \in \mathbb{R}^d$ be a unit vector. The directional derivative of f at $\mathbf{x}_0$ in the direction $\mathbf{v}$ is

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}} = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h}$$

provided the limit exists. ◁

Note that taking $\mathbf{v} = \mathbf{e}_i$ recovers the i-th partial derivative

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{e}_i} = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{e}_i) - f(\mathbf{x}_0)}{h} = \frac{\partial f(\mathbf{x}_0)}{\partial x_i}.$$

Conversely, a general directional derivative can be expressed in terms of the partial derivatives.

Theorem (Directional Derivative from Gradient): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$, let $\mathbf{x}_0 \in D$ be an interior point of D and let $\mathbf{v} \in \mathbb{R}^d$ be a unit vector. Assume that f is continuously differentiable at $\mathbf{x}_0$. Then the directional derivative of f at $\mathbf{x}_0$ in the direction $\mathbf{v}$ is given by

$$\frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}} = \mathbf{J}_f(\mathbf{x}_0)\,\mathbf{v} = \nabla f(\mathbf{x}_0)^T \mathbf{v}.$$

Proof idea: To bring out the partial derivatives, we re-write the directional derivative as the derivative of a composition of f with an affine function. We then use the chain rule.

Proof: Consider the composition $\beta(h) = f(\boldsymbol{\alpha}(h))$ where $\boldsymbol{\alpha}(h) = \mathbf{x}_0 + h\mathbf{v}$. Observe that $\boldsymbol{\alpha}(0) = \mathbf{x}_0$ and $\beta(0) = f(\mathbf{x}_0)$. Then, by definition of the derivative,

$$\frac{\mathrm{d}\beta(0)}{\mathrm{d}h} = \lim_{h \to 0} \frac{\beta(h) - \beta(0)}{h} = \lim_{h \to 0} \frac{f(\mathbf{x}_0 + h\mathbf{v}) - f(\mathbf{x}_0)}{h} = \frac{\partial f(\mathbf{x}_0)}{\partial \mathbf{v}}.$$

Applying the *Chain Rule*, we arrive at

$$\frac{\mathrm{d}\beta(0)}{\mathrm{d}h} = \mathbf{J}_\beta(0) = \mathbf{J}_f(\boldsymbol{\alpha}(0))\,\mathbf{J}_{\boldsymbol{\alpha}}(0) = \mathbf{J}_f(\mathbf{x}_0)\,\mathbf{J}_{\boldsymbol{\alpha}}(0).$$

It remains to compute $\mathbf{J}_{\boldsymbol{\alpha}}(0)$. By the linearity of derivatives (proved in a preivous exercise),

$$\frac{\partial \alpha_i(h)}{\partial h} = [x_{0,i} + hv_i]' = v_i$$

where we denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$ and $\mathbf{x}_0 = (x_{0,1}, \ldots, x_{0,d})^T$. So $\mathbf{J}_{\boldsymbol{\alpha}}(0) = \mathbf{v}$ and we get finally

$$\frac{\mathrm{d}\beta(0)}{\mathrm{d}h} = \mathbf{J}_f(\mathbf{x}_0)\,\mathbf{v},$$

as claimed. □

One can also define higher-order derivatives. We start with the single-variable case, where $f : D \to \mathbb{R}$ with $D \subseteq \mathbb{R}$ and $x_0 \in D$ is an interior point of D. . Note that, if f' exists in D, then it is itself a function of x. Then the second derivative at x_0 is

$$f''(x_0) = \frac{\mathrm{d}^2 f(x_0)}{\mathrm{d}x^2} = \lim_{h \to 0} \frac{f'(x_0 + h) - f'(x_0)}{h}$$

provided the limit exists.

We can also take higher-order derivatives. We will restrict ourselves to the second order.

Definition (Second Partial Derivatives and Hessian): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D. Assume that f is continuously differentiable in an open ball around $\mathbf{x}_0$. Then $\partial f(\mathbf{x})/\partial x_i$ is itself a function of $\mathbf{x}$ and its partial derivative with respect to x_j, if it exists, is denoted by

$$\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_i} = \lim_{h \to 0} \frac{\partial f(\mathbf{x}_0 + h\mathbf{e}_j)/\partial x_i - \partial f(\mathbf{x}_0)/\partial x_i}{h}.$$

To simplify the notation, we write this as $\partial^2 f(\mathbf{x}_0)/\partial x_i^2$ when $j = i$. If $\partial^2 f(\mathbf{x})/\partial x_j \partial x_i$ and $\partial^2 f(\mathbf{x})/\partial x_i^2$ exist and are continuous in an open ball around $\mathbf{x}_0$ for all i, j, we say that f is twice continuously differentiable at $\mathbf{x}_0$.

The Jacobian of the gradient ∇f is called the Hessian and is denoted by

$$\mathbf{H}_f(\mathbf{x}_0) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_d \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_1 \partial x_d} & \cdots & \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_d^2} \end{pmatrix}.$$

◁

When f is twice continuously differentiable at $\mathbf{x}_0$, its Hessian is a symmetric matrix.

Theorem (Symmetry of the Hessian): Let $f : D \to \mathbb{R}$ where $D \subseteq \mathbb{R}^d$ and let $\mathbf{x}_0 \in D$ be an interior point of D. Assume that f is twice continuously differentiable at $\mathbf{x}_0$. Then for all $i \neq j$

$$\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_i} = \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j}.$$

Proof idea: Two applications of the *Mean Value Theorem* show that the limits can be interchanged.

Proof: By definition of the partial derivative,

$$\begin{aligned}
\frac{\partial^2 f(\mathbf{x}_0)}{\partial x_j \partial x_i} &= \lim_{h_j \to 0} \frac{\partial f(\mathbf{x}_0 + h_j \mathbf{e}_j)/\partial x_i - \partial f(\mathbf{x}_0)/\partial x_i}{h_j} \\
&= \lim_{h_j \to 0} \lim_{h_i \to 0} \frac{1}{h_j h_i} \left\{ [f(\mathbf{x}_0 + h_j \mathbf{e}_j + h_i \mathbf{e}_i) - f(\mathbf{x}_0 + h_j \mathbf{e}_j)] - [f(\mathbf{x}_0 + h_i \mathbf{e}_i) - f(\mathbf{x}_0)] \right\} \\
&= \lim_{h_j \to 0} \lim_{h_i \to 0} \frac{1}{h_i} \left\{ \frac{[f(\mathbf{x}_0 + h_i \mathbf{e}_i + h_j \mathbf{e}_j) - f(\mathbf{x}_0 + h_i \mathbf{e}_i)] - [f(\mathbf{x}_0 + h_j \mathbf{e}_j) - f(\mathbf{x}_0)]}{h_j} \right\} \\
&= \lim_{h_j \to 0} \lim_{h_i \to 0} \frac{1}{h_i} \left\{ \frac{\partial}{\partial x_j} [f(\mathbf{x}_0 + h_i \mathbf{e}_i + \theta_j h_j \mathbf{e}_j) - f(\mathbf{x}_0 + \theta_j h_j \mathbf{e}_j)] \right\} \\
&= \lim_{h_j \to 0} \lim_{h_i \to 0} \frac{1}{h_i} \left\{ \partial f(\mathbf{x}_0 + h_i \mathbf{e}_i + \theta_j h_j \mathbf{e}_j)/\partial x_j - \partial f(\mathbf{x}_0 + \theta_j h_j \mathbf{e}_j)/\partial x_j \right\}
\end{aligned}$$

for some $\theta_j \in (0, 1)$. Note that, on the third line, we rearranged the terms and, on the fourth line, we applied the Mean Value Theorem to $f(\mathbf{x}_0 + h_i \mathbf{e}_i + h_j \mathbf{e}_j) - f(\mathbf{x}_0 + h_j \mathbf{e}_j)$ as a continuously differentiable function of h_j.

Because $\partial f/\partial x_j$ is continuously differentiable in an open ball around $\mathbf{x}_0$, a second application of the Mean Value Theorem gives for some $\theta_i \in (0, 1)$

$$\begin{aligned}
&\lim_{h_j \to 0} \lim_{h_i \to 0} \frac{1}{h_i} \left\{ \partial f(\mathbf{x}_0 + h_i \mathbf{e}_i + \theta_j h_j \mathbf{e}_j)/\partial x_j - \partial f(\mathbf{x}_0 + \theta_j h_j \mathbf{e}_j)/\partial x_j \right\} \\
&= \lim_{h_j \to 0} \lim_{h_i \to 0} \frac{\partial}{\partial x_i} \left[\partial f(\mathbf{x}_0 + \theta_j h_j \mathbf{e}_j + \theta_i h_i \mathbf{e}_i)/\partial x_j \right] \\
&= \lim_{h_j \to 0} \lim_{h_i \to 0} \frac{\partial^2 f(\mathbf{x}_0 + \theta_j h_j \mathbf{e}_j + \theta_i h_i \mathbf{e}_i)}{\partial x_i \partial x_j}.
\end{aligned}$$

The claim then follows from the continuity of $\partial^2 f/\partial x_i \partial x_j$. □

Example: Consider again the quadratic function

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T P \mathbf{x} + \mathbf{q}^T \mathbf{x} + r.$$

Recall that the gradient of f is

$$\nabla f(\mathbf{x}) = \frac{1}{2} [P + P^T] \mathbf{x} + \mathbf{q}.$$

Each component of ∇f is an affine function of $\mathbf{x}$, so by our previous result the gradient of ∇f is

$$\mathbf{H}_f = \frac{1}{2} [P + P^T].$$

Observe that this is indeed a symmetric matrix. ◁

NUMERICAL CORNER We return to [automatic differentiation (https://en.wikipedia.org/wiki/Automatic_differentiation)](https://en.wikipedia.org/wiki/Automatic_differentiation).

Each component of the output of `gradient(f, x, y)` is itself a function and can also be differentiated to obtain the second derivative.

In [10]:
```
f(x, y) = x * y + x^2 + exp(x) * cos(y)
```

Out[10]:
```
f (generic function with 1 method)
```

In [11]:
```
dfdx(x, y) = gradient(f, x, y)[1]
dfdy(x, y) = gradient(f, x, y)[2];
```

In [12]:
```
dfdx(0., 0.) # answer is 1 (see example is next notebook)
```

Out[12]:
```
1.0
```

In [13]:
```
df2dx2(x, y) = gradient(dfdx, x, y)[1]
```

Out[13]:
```
df2dx2 (generic function with 1 method)
```

In [14]:
```
df2dx2(0., 0.) # answer is 3 (see example is next notebook)
```

Out[14]:
```
3.0
```