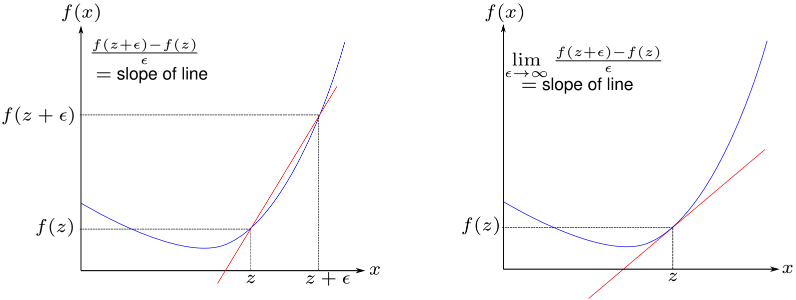
3. Derivatives

- scalar derivatives
- gradient and hessian
- multi-variable differentiation rules

Derivative definition

the *derivative* of $f : \mathbb{R} \to \mathbb{R}$ at the number z is defined as

$$f'(z) = \frac{df}{dz} = \lim_{\epsilon \to 0} \frac{f(z+\epsilon) - f(z)}{\epsilon}$$



- when $f'(x)$ is positive, the function $f(x)$ increases as x does
- $f'(x)$ is negative, $f(x)$ decreases as x increases

Common derivatives

$f(x)$	$f'(x)$
c	0
x^ℓ	$\ell x^{\ell-1}$
e^x	e^x
$\log(x), x > 0$	$\frac{1}{x}$
$\log_c(x), x > 0, c > 0$	$\frac{1}{x \ln(c)}$
$\sin(x)$	$\cos(x)$
$\cos(x)$	$-\sin(x)$

(we use $\log(.) = \ln(.)$ to denote the natural logarithm)

Derivative rules

- **Linearity:** for $f(x) = \alpha g(x) + \beta h(x)$:

$$f'(x) = \alpha g'(x) + \beta h'(x)$$

- **Product rule:** for $f(x) = g(x)h(x)$:

$$f'(x) = g'(x)h(x) + g(x)h'(x)$$

- **Quotient rule:** for $f(x) = \frac{g(x)}{h(x)}$:

$$f'(x) = \frac{g'(x)h(x) - g(x)h'(x)}{h(x)^2}$$

- **Chain rule:** for $f(x) = g(h(x))$ where $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$:

$$f'(x) = h'(x)g'\big(h(x)\big)$$

Second derivative

the *second derivative* of f at a point z is the derivative of the first derivative:

$$f''(z) = \frac{d^2 f}{dz^2} = \lim_{\epsilon \to 0} \frac{f'(z+\epsilon) - f'(z)}{\epsilon},$$

the second derivative conveys information about the curvature of the function

- when $f''(x) > 0$, then $f'(x)$ is increasing, which suggests the slope of the tangent line to f increases as x does yielding a concave-upwards shape
- if $f''(x)$ is negative, the function exhibits a concave-downwards curvature

Outline

- scalar derivatives
- **gradient and hessian**
- multi-variable differentiation rules

Gradient

the *gradient* of $f : \mathbb{R}^n \to \mathbb{R}$ (at $\boldsymbol{z}$) is

$$\nabla f(\boldsymbol{z}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\boldsymbol{z}) \\ \frac{\partial f}{\partial x_2}(\boldsymbol{z}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\boldsymbol{z}) \end{bmatrix}$$

- entries $\frac{\partial f}{\partial x_i}(\boldsymbol{z})$ are *partial derivative* of f at point $\boldsymbol{z}$, with respect to x_i:

$$\frac{\partial f}{\partial x_i}(\boldsymbol{z}) = \lim_{\epsilon \to 0} \frac{f(z_1, \ldots, z_{i-1}, z_i + \epsilon, z_{i+1}, \ldots, z_n) - f(\boldsymbol{z})}{\epsilon}$$

- the gradient $\nabla f(\boldsymbol{z})$ is a vector that is pointing to the direction where f increases the fastest at $\boldsymbol{z}$

Example 3.1

a) partial derivatives of $f(\boldsymbol{x}) = \|\boldsymbol{x}\|^2 = x_1^2 + \cdots + x_n^2$ are $\frac{\partial f}{\partial x_i}(\boldsymbol{x}) = 2x_i$; hence

$$\nabla f(\boldsymbol{x}) = (2x_1, \ldots, 2x_n) = 2\boldsymbol{x}$$

b) gradient of the function $f(\boldsymbol{x}) = 5x_1 + 8x_2 + x_1x_2 - x_1^2 - 2x_2^2$ is

$$\nabla f(\boldsymbol{x}) = (5 + x_2 - 2x_1, 8 + x_1 - 4x_2)$$

c) gradient of $f(x) = x_1^2 + e^{-x_1} + \sin(x_2)$ is

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 2x_1 - e^{-x_1} \\ \cos(x_2) \end{bmatrix}$$

Hessian

the *Hessian* of a function $f : \mathbb{R}^n \to \mathbb{R}$ at $\boldsymbol{z}$ is is defined as

$$\nabla^2 f(\boldsymbol{z}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2}(\boldsymbol{z}) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(\boldsymbol{z}) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{z}) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(\boldsymbol{z}) & \frac{\partial^2 f}{\partial x_2^2}(\boldsymbol{z}) & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(\boldsymbol{z}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{z}) & \frac{\partial^2 f}{\partial x_n \partial x_2}(\boldsymbol{z}) & \cdots & \frac{\partial^2 f}{\partial x_n^2}(\boldsymbol{z}) \end{bmatrix}$$

- we say that f is *twice differentiable* if $\nabla^2 f(\boldsymbol{x})$ exists for all $\boldsymbol{x} \in \mathbb{R}^n$
- the Hessian is a symmetric matrix $\nabla^2 f(\boldsymbol{z}) = \nabla^2 f(\boldsymbol{z})^T$ since

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\boldsymbol{z}) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\boldsymbol{z}), \quad \text{for all } i, j = 1, \ldots, n$$

Example 3.2

a) for $f(\boldsymbol{x}) = 5x_1 + 8x_2 + x_1x_2 - x_1^2 - 2x_2^2$:

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} 5 + x_2 - 2x_1 \\ 8 + x_1 - 4x_2 \end{bmatrix}, \quad \nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} -2 & 1 \\ 1 & -4 \end{bmatrix}$$

b) for

$$f(\boldsymbol{x}) = e^{x_1+x_2-1} + e^{x_1-x_2-1} + e^{-x_1-1}$$

the gradient is

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} e^{x_1+x_2-1} + e^{x_1-x_2-1} - e^{-x_1-1} \\ e^{x_1+x_2-1} - e^{x_1-x_2-1} \end{bmatrix}$$

and the Hessian is

$$\nabla^2 f(\boldsymbol{x}) = \begin{bmatrix} e^{x_1+x_2-1} + e^{x_1-x_2-1} + e^{-x_1-1} & e^{x_1+x_2-1} - e^{x_1-x_2-1} \\ e^{x_1+x_2-1} - e^{x_1-x_2-1} & e^{x_1+x_2-1} + e^{x_1-x_2-1} \end{bmatrix}$$

Linear and quadratic functions

Linear and affine functions: for $f(\boldsymbol{x}) = \boldsymbol{a}^T\boldsymbol{x} + b$:

$$\nabla f(\boldsymbol{x}) = \boldsymbol{a}$$
$$\nabla^2 f(\boldsymbol{x}) = 0$$

Quadratic functions: for $f(\boldsymbol{x}) = \boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{r}^T\boldsymbol{x} + c$, where $Q = Q^T$ is symmetric:

$$\nabla f(\boldsymbol{x}) = 2Q\boldsymbol{x} + \boldsymbol{r}$$
$$\nabla^2 f(\boldsymbol{x}) = 2Q$$

Least-squares function

the *least-squares function* $f(\boldsymbol{x}) = \|A\boldsymbol{x} - \boldsymbol{b}\|^2$ can be expressed as

$$\begin{aligned} f(\boldsymbol{x}) &= \|A\boldsymbol{x} - \boldsymbol{b}\|^2 \\ &= (A\boldsymbol{x} - \boldsymbol{b})^T(A\boldsymbol{x} - \boldsymbol{b}) \\ &= (\boldsymbol{x}^T A^T - \boldsymbol{b}^T)(A\boldsymbol{x} - \boldsymbol{b}) \\ &= \boldsymbol{x}^T A^T A\boldsymbol{x} - \boldsymbol{b}^T A\boldsymbol{x} - \boldsymbol{x}^T A^T \boldsymbol{b} + \boldsymbol{b}^T\boldsymbol{b} \\ &= \boldsymbol{x}^T A^T A\boldsymbol{x} - 2\boldsymbol{b}^T A\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{b} \end{aligned}$$

this means that f is quadratic $f(\boldsymbol{x}) = \boldsymbol{x}^T Q\boldsymbol{x} + \boldsymbol{r}^T\boldsymbol{x} + c$ with

$$Q = A^T A, \quad \boldsymbol{r}^T = -2\boldsymbol{b}^T A, \quad c = \boldsymbol{b}^T\boldsymbol{b}$$

hence,

$$\nabla f(\boldsymbol{x}) = 2A^T A\boldsymbol{x} - 2A^T\boldsymbol{b}, \quad \nabla^2 f(\boldsymbol{x}) = 2A^T A$$

Outline

- scalar derivatives
- gradient and hessian
- **multi-variable differentiation rules**

Composition with affine function

$$f(\boldsymbol{x}) = g(A\boldsymbol{x} + \boldsymbol{b})$$

- $f : \mathbb{R}^n \rightarrow$, $g : \mathbb{R}^m \rightarrow \mathbb{R}$
- A is an $m \times n$ matrix
- $\boldsymbol{b}$ is an m vector

the gradient and Hessian are

$$\nabla f(\boldsymbol{x}) = A^T \nabla g(A\boldsymbol{x} + \boldsymbol{b})$$

and

$$\nabla^2 f(\boldsymbol{x}) = A^T \nabla^2 g(A\boldsymbol{x} + \boldsymbol{b}) A$$

Example 3.3

use the composition with affine function property to find the gradient and Hessian of

$$f(\boldsymbol{x}) = e^{x_1+x_2-1} + e^{x_1-x_2-1} + e^{-x_1-1}$$

we can express f as $f(\boldsymbol{x}) = g(A\boldsymbol{x} + \boldsymbol{b})$, where $g(\boldsymbol{y}) = e^{y_1} + e^{y_2} + e^{y_3}$, and

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 0 \end{bmatrix}, \quad \boldsymbol{b} = \begin{bmatrix} -1 \\ -1 \\ -1 \end{bmatrix}$$

the gradient and Hessian of g are

$$\nabla g(\boldsymbol{y}) = \begin{bmatrix} e^{y_1} \\ e^{y_2} \\ e^{y_3} \end{bmatrix}, \quad \nabla^2 g(\boldsymbol{y}) = \begin{bmatrix} e^{y_1} & 0 & 0 \\ 0 & e^{y_2} & 0 \\ 0 & 0 & e^{y_3} \end{bmatrix}$$

hence

$$\nabla f(\boldsymbol{x}) = A^T \nabla g(A\boldsymbol{x}+\boldsymbol{b}) = \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} e^{x_1+x_2-1} \\ e^{x_1-x_2-1} \\ e^{-x_1-1} \end{bmatrix}$$

$$= \begin{bmatrix} e^{x_1+x_2-1} + e^{x_1-x_2-1} - e^{-x_1-1} \\ e^{x_1+x_2-1} - e^{x_1-x_2-1} \end{bmatrix}$$

and

$$\begin{aligned} \nabla^2 f(\boldsymbol{x}) &= A^T \nabla^2 g(A\boldsymbol{x}+\boldsymbol{b}) A \\ &= \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} e^{x_1+x_2-1} & 0 & 0 \\ 0 & e^{x_1-x_2-1} & 0 \\ 0 & 0 & e^{-x_1-1} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 0 \end{bmatrix} \\ &= \begin{bmatrix} e^{x_1+x_2-1} + e^{x_1-x_2-1} + e^{-x_1-1} & e^{x_1+x_2-1} - e^{x_1-x_2-1} \\ e^{x_1+x_2-1} - e^{x_1-x_2-1} & e^{x_1+x_2-1} + e^{x_1-x_2-1} \end{bmatrix} \end{aligned}$$

Jacobian

let $f : \mathbb{R}^n \to \mathbb{R}^m$:

$$f(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_m) \end{bmatrix}$$

where f_i is a scalar-valued function of $\boldsymbol{x}$

the **Jacobian** or **derivative matrix** of f at $\boldsymbol{z}$ is the $m \times n$ matrix:

$$Df(\boldsymbol{z}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\boldsymbol{z}) & \frac{\partial f_1}{\partial x_2}(\boldsymbol{z}) & \cdots & \frac{\partial f_1}{\partial x_n}(\boldsymbol{z}) \\ \frac{\partial f_2}{\partial x_1}(\boldsymbol{z}) & \frac{\partial f_2}{\partial x_2}(\boldsymbol{z}) & \cdots & \frac{\partial f_2}{\partial x_n}(\boldsymbol{z}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\boldsymbol{z}) & \frac{\partial f_m}{\partial x_2}(\boldsymbol{z}) & \cdots & \frac{\partial f_m}{\partial x_n}(\boldsymbol{z}) \end{bmatrix} = \begin{bmatrix} \nabla f_1(\boldsymbol{z})^T \\ \nabla f_2(\boldsymbol{z})^T \\ \vdots \\ \nabla f_m(\boldsymbol{z})^T \end{bmatrix}$$

- if $m = 1$, then $Df(\boldsymbol{z}) = \nabla f(\boldsymbol{z})^T$
- Jacobian of the gradient of $f : \mathbb{R}^n \to \mathbb{R}$ is its Hessian

Example 3.4

a) the Jacobian of

$$f(\boldsymbol{x}) = \begin{bmatrix} x_1 + x_2^2 \\ -x_1 + x_1 x_2 \end{bmatrix}$$

is

$$Df(\boldsymbol{x}) = \begin{bmatrix} 1 & 2x_2 \\ -1 + x_2 & x_1 \end{bmatrix}$$

b) the derivative matrix or Jacobian of $f(\boldsymbol{x}) = A\boldsymbol{x}$ is

$$Df(\boldsymbol{x}) = A$$

Rules

Sum of two functions: if $f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$, then

$$\nabla f(\boldsymbol{x}) = \nabla f_1(\boldsymbol{x}) + \nabla f_2(\boldsymbol{x}), \quad \nabla^2 f(\boldsymbol{x}) = \nabla^2 f_1(\boldsymbol{x}) + \nabla^2 f_2(\boldsymbol{x})$$

Scalar multiplication: if $f(\boldsymbol{x}) = \alpha g(\boldsymbol{x})$, where α is a scalar, then

$$\nabla f(\boldsymbol{x}) = \alpha \nabla g(\boldsymbol{x}), \quad \nabla^2 f(\boldsymbol{x}) = \alpha \nabla^2 g(\boldsymbol{x})$$

Multivariable product rule: let $f : \mathbb{R}^n \to \mathbb{R}$ be

$$f(\boldsymbol{x}) = g(\boldsymbol{x})^T h(\boldsymbol{x}),$$

where $g : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^n \to \mathbb{R}^m$, then

$$\nabla f(\boldsymbol{x}) = Df(\boldsymbol{x})^T = Dg(\boldsymbol{x})^T h(\boldsymbol{x}) + Dh(\boldsymbol{x})^T g(\boldsymbol{x})$$

Multivariable chain rule: let $f : \mathbb{R}^n \to \mathbb{R}$ be the composition

$$f(\boldsymbol{x}) = g(h(\boldsymbol{x})) = g(h_1(\boldsymbol{x}), \ldots, h_p(\boldsymbol{x}))$$

where $g : \mathbb{R}^p \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}^p$ are differentiable functions

- using the chain rule, the partial derivatives of f are

$$\frac{\partial f}{\partial x_j}(\boldsymbol{x}) = \frac{\partial h_1}{\partial x_j}(\boldsymbol{x})\, \frac{\partial g}{\partial y_1}\big(h(\boldsymbol{x})\big) + \cdots + \frac{\partial h_p}{\partial x_j}(\boldsymbol{x})\, \frac{\partial g}{\partial y_p}\big(h(\boldsymbol{x})\big)$$

  for $j = 1, \ldots, n$

- the gradient can be compactly represented as the vector-matrix product:

$$\nabla f(\boldsymbol{x}) = Df(\boldsymbol{x})^T = Dh(\boldsymbol{x})^T \nabla g\big(h(\boldsymbol{x})\big)$$

Example 3.5

a) use the chain-rule to find the gradient of

$$f(\boldsymbol{x}) = \big(\sin(x_1) + x_2^2\big)^2 + \big(\sin(x_1) + x_2^2\big)(x_1 + x_2)^2$$

- we can write f as $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$ where

$$g(\boldsymbol{y}) = y_1^2 + y_1 y_2^2, \quad h(\boldsymbol{x}) = \begin{bmatrix} \sin(x_1) + x_2^2 \\ x_1 + x_2 \end{bmatrix}$$

- the gradient of g is $\nabla g(\boldsymbol{y}) = \begin{bmatrix} 2y_1 + y_2^2 \\ 2y_1 y_2 \end{bmatrix}$ and the derivative of h is $Dh(\boldsymbol{x}) = \begin{bmatrix} \cos(x_1) & 2x_2 \\ 1 & 1 \end{bmatrix}$
- hence,

$$\begin{aligned} \nabla f(\boldsymbol{x}) &= Dh(\boldsymbol{x})^T \nabla g\big(h(\boldsymbol{x})\big) \\ &= \begin{bmatrix} \cos(x_1) & 1 \\ 2x_2 & 1 \end{bmatrix}^T \begin{bmatrix} 2\sin(x_1) + 2x_2^2 + (x_1 + x_2)^2 \\ 2(\sin(x_1) + x_2^2)(x_1 + x_2) \end{bmatrix} \end{aligned}$$

b) *nonlinear least-squares function:*

$$f(\boldsymbol{x}) = \|h(\boldsymbol{x})\|^2 = \sum_{j=1}^{p} h_j(\boldsymbol{x})^2$$

we have $f(\boldsymbol{x}) = g(h(\boldsymbol{x}))$ where $g(\boldsymbol{y}) = \|\boldsymbol{y}\|^2$

using $\nabla g(\boldsymbol{y}) = 2\boldsymbol{y}$ and the chain rule, we get

$$\nabla f(\boldsymbol{x}) = Dh(\boldsymbol{x})^T \nabla g\big(h(\boldsymbol{x})\big) = 2Dh(\boldsymbol{x})^T h(\boldsymbol{x})$$

References and further readings

- L. Vandenberghe. *EE133A Lecture Notes,* UCLA. (`http://www.seas.ucla.edu/~vandenbe/ee133a.html`)
- Edwin KP Chong and Stanislaw H Zak. *An Introduction to Optimization,* John Wiley & Sons, 2013, chapters 2,3.
- Stephen Boyd and Lieven Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares,* Cambridge University Press, 2018, chapters 3,5,8.
- Stephen Boyd and Lieven Vandenberghe. *Convex Optimization,* Cambridge University Press, 2004, appendices A.1, A.5, C.5.