3. Derivatives

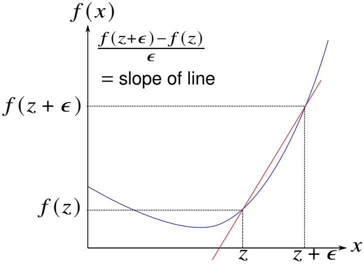
- scalar derivatives
- gradient and hessian
- differentiation rules
- Taylor approximation
- level sets and directional derivative

Derivative definition

the *derivative* of $f(x)$ ($f : \mathbb{R} \to \mathbb{R}$) at a point z is

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

- geometrically, $f'(z)$ is the slope of the tangent line to the graph of f at the point z



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

Common derivatives

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

(we use $\log(\cdot) = \ln(\cdot)$ 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 $f(x) = g(h(x))$:

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

Second derivative

the *second derivative* of $f(x)$ at a point z is the derivative of the first derivative:

$$\begin{aligned} f''(z) = \frac{d^2 f}{dx^2}(z) &= \lim_{\epsilon \to 0} \frac{f'(z+\epsilon) - f'(z)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{f'(z+\epsilon) - 2f'(z) + f'(z-\epsilon)}{\epsilon^2} \end{aligned}$$

- 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**
- differentiation rules
- Taylor approximation
- level sets and directional derivative

Gradient

- the *partial derivative* of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at point z is, with respect to x_i is

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

- quantifies the variation of f concerning x_i, while other variables remain constant

the **gradient** of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at point z is the n-vector

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

f is *differentiable* if its $\mathrm{dom}\, f$ is open and $\nabla f(x)$ exists for every $x \in \mathrm{dom}\, f$

Examples

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

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

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

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

- partial derivatives of

$$f(x) = \|x\|^2 = x_1^2 + \cdots + x_n^2$$

are $\frac{\partial f}{\partial x_i}(x) = 2x_i$; hence

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

Jacobian

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

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix} = \begin{bmatrix} f_1(x_1, \ldots, x_n) \\ \vdots \\ f_m(x_1, \ldots, x_m) \end{bmatrix}, \quad f_i : \mathbb{R}^n \to \mathbb{R}$$

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

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

if $m = 1$, then $Df(z) = \nabla f(z)^T$

Examples

- the Jacobian of

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

is

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

- the derivative matrix or Jacobian of $f(x) = Ax$ is

$$Df(x) = A$$

Hessian

the **Hessian** of a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ at z is the $n \times n$ matrix

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

- f is *twice differentiable* if $\nabla^2 f(x)$ exists for all $x \in \mathrm{dom}\, f$ (with open domain)
- the Hessian is a *symmetric* matrix $\nabla^2 f(z) = \nabla^2 f(z)^T$ since

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

- Jacobian of the gradient of $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is its Hessian: $D\nabla f(x) = \nabla^2 f(x)$

Examples

- for $f(x) = 5x_1 + 8x_2 + x_1x_2 - x_1^2 - 2x_2^2$:

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

- for

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

the gradient is

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

and the Hessian is

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

Linear and quadratic functions

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

$$\begin{aligned} \nabla f(x) &= a \\ \nabla^2 f(x) &= 0 \end{aligned}$$

Quadratic functions: for $f(x) = x^T Q x + r^T x + s$, where $Q = Q^T$ is symmetric:

$$\begin{aligned} \nabla f(x) &= 2Qx + r \\ \nabla^2 f(x) &= 2Q \end{aligned}$$

Least-squares function

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

$$\begin{aligned} f(x) &= \|Ax - b\|^2 \\ &= (Ax - b)^T(Ax - b) \\ &= (x^TA^T - b^T)(Ax - b) \\ &= x^TA^TAx - b^TAx - x^TA^Tb + b^Tb \\ &= x^TA^TAx - 2b^TAx + b^Tb \end{aligned}$$

this means that f is quadratic $f(x) = x^TQx + r^Tx + s$ with

$$Q = A^TA, \quad r^T = -2b^TA, \quad s = b^Tb$$

hence,

$$\nabla f(x) = 2A^TAx - 2A^Tb, \quad \nabla^2 f(x) = 2A^TA$$

Outline

- scalar derivatives
- gradient and hessian
- **differentiation rules**
- Taylor approximation
- level sets and directional derivative

Sum and scalar multiplication

Sum of two functions: if $f(x) = g(x) + h(x)$, then

$$\nabla f(x) = \nabla g(x) + \nabla h(x), \quad \nabla^2 f(x) = \nabla^2 g(x) + \nabla^2 h(x)$$

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

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

Product rule

Product rule: let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be

$$f(x) = g(x)^T h(x),$$

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

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

Product rule for second derivative

- if $f(x) = g(x)h(x)$ where $g : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $h : \mathbb{R}^n \longrightarrow \mathbb{R}$
- the Hessian is

$$\nabla^2 f(x) = \nabla^2 g(x)h(x) + \nabla^2 h(x)g(x) + \nabla g(x)\nabla h(x)^T + \nabla h(x)\nabla g(x)^T$$

Example: pure quadratic function

$$f(x) = x^T A x \quad \text{where } A \text{ is not symmetric}$$

- since $f(x) = x^T(0.5A + 0.5A^T)x$, we know from before that $\nabla f(x) = (A + A^T)x$
- we can also derive the gradient using the product rule
- express f as $f(x) = g(x)^T h(x)$ where $g(x) = x$ and $h(x) = Ax$
- we have

$$Dg(x) = I \quad \text{and} \quad Dh(x) = A$$

- applying the product rule we obtain:

$$\begin{aligned} \nabla f(x) &= Dg(x)^T h(x) + Dh(x)^T g(x) \\ &= Ax + A^T x \\ &= (A + A^T)x \end{aligned}$$

Example: nonlinear least squares

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

- each term of the sum is the product of two identical function $h_j(x)h_j(x)$
- so we can apply the product rule to each term find the gradient as:

$$\nabla f(x) = \sum_{j=1}^{p} 2Dh_j(x)^T h_j(x) = 2\sum_{j=1}^{p} 2\nabla h_j(x)h_j(x) = 2Dh(x)h(x)$$

- the Hessian can also be found using the product rule and is given by:

$$\begin{aligned}\nabla^2 f(x) &= 2\sum_{j=1}^{p}\left(\nabla h_j(x)\nabla h_j(x)^T + h_j(x)\nabla^2 h_j(x)\right)\\ &= 2Dh(x)^T Dh(x) + 2\sum_{j=1}^{p} h_j(x)\nabla^2 h_j(x)\end{aligned}$$

Chain rule

let $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ be the composition

$$f(x) = g(h(x)) = g(h_1(x), \dots, h_p(x))$$

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

Chain rule

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

Chain rule for second derivative

- let $f : \mathbb{R}^n \longrightarrow$ be $f(x) = g(h(x))$ with $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ and $g : \mathbb{R} \longrightarrow \mathbb{R}$
- the Hessian is

$$\nabla^2 f(x) = g'(h(x))\nabla^2 h(x) + g''(h(x))\nabla h(x)\nabla h(x)^T$$

Example

we use the chain-rule to find the gradient of

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

- we can write f as $f(x) = g(h(x))$ where

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

- we have $\nabla g(y) = \begin{bmatrix} 2y_1 + y_2^2 \\ 2y_1 y_2 \end{bmatrix}$ and $Dh(x) = \begin{bmatrix} \cos(x_1) & 2x_2 \\ 1 & 1 \end{bmatrix}$
- hence,

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

Example: nonlinear least-squares

consider again the function $f(x) = \|h(x)\|^2 = \sum_{j=1}^{p} h_j(x)^2$

- we have $f(x) = g(h(x))$ where $g(y) = \|y\|^2$
- using $\nabla g(y) = 2y$ and the chain rule, we get

$$\nabla f(x) = Dh(x)^T \nabla g(h(x)) = 2Dh(x)^T h(x)$$

- the Hessian can be found using the chain rule applied to each term

$$f_j(x) = g(h_j(x)) \quad \text{where} \quad g(y) = y^2$$

- with $g'(y) = 2y$ and $g''(y) = 2$, we get

$$\begin{aligned}\nabla^2 f(x) &= \sum_{j=1}^{p} 2h_j(x)\nabla^2 h_j(x) + 2\nabla h_j(x)\nabla h_j(x)^T \\ &= 2\sum_{j=1}^{p} h_j(x)\nabla^2 h_j(x) = 2Dh(x)^T Dh(x)\end{aligned}$$

Composition with affine function

$$f(x) = g(Ax + b)$$

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

the gradient and Hessian are

$$\nabla f(x) = A^T \nabla g(Ax + b)$$

and

$$\nabla^2 f(x) = A^T \nabla^2 g(Ax + b) A$$

Example

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

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

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

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

the gradient and Hessian of g are

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

hence

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

and

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

Example

$$f(x) = \log \sum_{i=1}^{m} \exp(a_i^T x + b_i)$$

where $a_1, \ldots, a_m \in \mathbb{R}^n$ and $b_1, \ldots, b_m \in \mathbb{R}$

- this is the composition of the affine function $Ax + b$ and the function:

$$g(y) = \log \left(\sum_{i=1}^{m} \exp y_i \right)$$

where $A \in \mathbb{R}^{m \times n}$ is a matrix whose rows are $a_1^T, \ldots, a_m^T$

- differentiating $g(y)$ gives:

$$\nabla g(y) = \frac{1}{\sum_{i=1}^{m} \exp y_i} \begin{bmatrix} \exp y_1 \\ \vdots \\ \exp y_m \end{bmatrix}$$

- using the composition rule for gradients, we find:

$$\nabla f(x) = \frac{1}{\mathbf{1}^T z} A^T z$$

where $z_i = \exp(a_i^T x + b_i)$ for $i = 1, \ldots, m$

- for the Hessian, taking the partial derivatives of $g(y)$ yields:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \begin{cases} \frac{\exp(y_i) \sum_{i=1}^m \exp y_i - \exp(y_i)^2}{(\sum_{i=1}^m \exp y_i)^2} & i = j \\ -\frac{\exp(y_i) \exp(y_j)}{(\sum_{i=1}^m \exp y_i)^2} & i \neq j \end{cases}$$

or in matrix form:

$$\nabla^2 g(y) = \mathrm{diag}(\nabla g(y)) - \nabla g(y) \nabla g(y)^T$$

- applying the composition formula, the Hessian of $f(x)$ becomes:

$$\nabla^2 f(x) = A^T \left(\frac{1}{\mathbf{1}^T z} \mathrm{diag}(z) - \frac{1}{(\mathbf{1}^T z)^2} z z^T \right) A$$

where $z_i = \exp(a_i^T x + b_i)$ for $i = 1, \ldots, m$

Outline

- scalar derivatives
- gradient and hessian
- differentiation rules
- **Taylor approximation**
- level sets and directional derivative

First-order Taylor (affine) approximation

first-order *Taylor approximation* of $f : \mathbb{R}^n \to \mathbb{R}$, near point z:

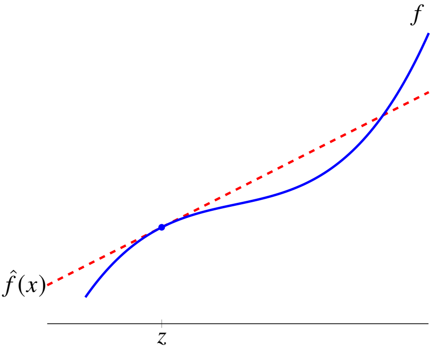
$$\begin{aligned}\hat{f}(x) &= f(z) + \frac{\partial f}{\partial x_1}(z)\,(x_1 - z_1) + \cdots + \frac{\partial f}{\partial x_n}(z)\,(x_n - z_n) \\ &= f(z) + \nabla f(z)^T(x - z)\end{aligned}$$

first-order Taylor approximation of differentiable $f : \mathbb{R}^n \to \mathbb{R}^m$ around z:

$$\hat{f}(x) = f(z) + Df(z)(x - z)$$

- $\hat{f}(x)$ is very close to $f(x)$ when x_i are all near z_i
- sometimes written $\hat{f}(x; z)$, to indicate that z where the approximation appear
- $\hat{f}$ is an *affine* function of x (often called *linear approximation* of f near z)
- useful in deriving and analyzing algorithms (we will see later)

Illustration with one variable



$$\hat{f}(x) = f(z) + f'(z)(x - z)$$

Example for scalar valued functions

$$f(x_1, x_2) = x_1 - 3x_2 + e^{2x_1+x_2-1}$$

- gradient:

$$\nabla f(x) = \begin{bmatrix} 1 + 2e^{2x_1+x_2-1} \\ -3 + e^{2x_1+x_2-1} \end{bmatrix}$$

- Taylor approximation around $z = 0$:

$$\begin{aligned} \hat{f}(x) &= f(0) + \nabla f(0)^T(x - 0) \\ &= e^{-1} + (1 + 2e^{-1})x_1 + (-3 + e^{-1})x_2 \end{aligned}$$

Example for vector valued functions

$$f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \end{bmatrix} = \begin{bmatrix} e^{2x_1+x_2} - x_1 \\ x_1^2 - x_2 \end{bmatrix}$$

- derivative matrix

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

- first order approximation of f around $z = 0$:

$$\hat{f}(x) = \begin{bmatrix} \hat{f}_1(x) \\ \hat{f}_2(x) \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

Second-order approximation

for $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, the second-order Taylor approximation of f near z is given by:

$$f(x) \approx \hat{f}(x) = f(z) + \nabla f(z)^T(x-z) + (1/2)(x-z)^T\nabla^2 f(z)(x-z)$$

- for $n = 1$ reduces to

$$f(x) \approx \hat{f}(x) = f(z) + f'(z)(x-z) + \frac{f''(z)}{2}(x-z)^2$$

- a quadratic function of x; hence, called also quadratic approximation
- useful in deriving and analyzing algorithms (we will see later)

Outline

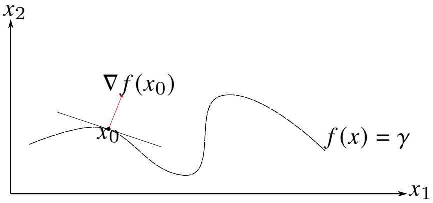
- scalar derivatives
- gradient and hessian
- differentiation rules
- Taylor approximation
- **level sets and directional derivative**

Gradient and level sets

- gradient $\nabla f(x_0)$ is orthogonal to the level sets $f(x) = \gamma$ at $\gamma = f(x_0)$
- to see this,, consider a curve within $\mathcal{S}_\gamma$ parametrized by $r : \mathbb{R} \to \mathbb{R}^n$
- for $r(t_0) = x_0$ and $Dr(t_0) = r' \neq 0$, r' is the tangent vector to the curve at x_0
- the derivative of the function $h(t) = f(r(t)) = \gamma$ yields

$$0 = h'(t_0) = \nabla f(r(t_0))^T Dr(t_0) = \nabla f(x_0)^T r'$$

- this implies $\nabla f(x_0)$ is perpendicular to r'



Directional derivative

let $f : \mathbb{R}^n \to \mathbb{R}$ and consider the function $h(\alpha) = f(x + \alpha v)$ restricted to a line

- using the chain rule (composition with affine function), we have

$$h'(\alpha) = v^T \nabla f(x + \alpha v)$$

- for $\alpha = 0$, this value is

$$f'(x; v) = h'(0) = \lim_{\alpha \to 0} \frac{f(x + \alpha v) - f(x)}{\alpha}$$

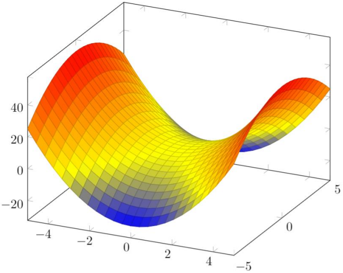
  and called the *directional derivative* of f in the direction of v

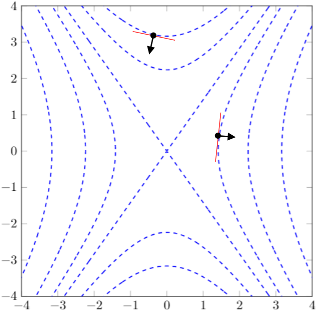
- when $\nabla f(x)^T v > 0$, we have $f(x + \alpha v) > f(x)$ for sufficiently small positive α
- when $\nabla f(x)^T v < 0$, we have $f(x + \alpha v) < f(x)$
- using Cauchy-Schwarz,

$$\nabla f(x)^T v \leq \|\nabla f(x)\| \|v\|$$

  making the directional derivative maximized when $v = \nabla f(x)$

Example





$\nabla f(x)$ is a vector pointing to the direction where f increases the fastest at x

References and further readings

- E. K.P. Chong, Wu-S. Lu, and S. H. Zak, *An Introduction to Optimization: With Applications to Machine Learning*. John Wiley & Sons, 2023. (Ch. 5)

- S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, 2004. (Appendix A.4)

- S. Boyd and L. Vandenberghe. *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares.* Cambridge University Press, 2018. (Appendix C.1)

- L. Vandenberghe, *EE133A Lecture Notes*, UCLA.