Session 14 : Multivariate function approx and optimization

Example of Taylor polynomial

$$f(x) = e^x$$

General Taylor approximation of $f(x)$ at $a = 0$

$$P_{n,a}(x) = f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2 + \cdots$$
$$= \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(a)(x-a)^k$$

$$f(x) = e^x, \; f'(x) = e^x, \; f''(x) = e^x, \ldots$$
$$f^{(k)}(x) = e^x \quad \text{for all } K \geq 0$$

$$P_{n,0}(x) = 1 + x + \frac{1}{2}x^2 + \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \cdots$$
$$= \sum_{k=0}^{n} \frac{1}{k!} x^k \quad \lrcorner$$

Another example : $f(x) = \log(1-x)$

General Taylor approx. of $f(x)$ at $a = 0$

$$f(a) = f(0) = \log(1-0) = \log(1) = 0$$

$$f'(x) = \frac{-1}{1-x} \Rightarrow f'(0) = -1$$

$(-1)\cdot(1-x)^{-1}$ ↑ think of it as...

$$f''(x) = (-1)(-1)(1-x)^{-2}(-1) = (-1)^3(1-x)^{-2}$$
$$\Rightarrow f''(0) = -1$$

$$f'''(x) = (-1)^3(-2)(1-x)^{-3}(-1)$$
$$= (-1)^5 2(1-x)^{-3}$$
$$\Rightarrow f'''(0) = -2$$

odd $\rightarrow 2p+1 = 1,3,5,\ldots$

even $\rightarrow 2p = 2,4,6,\ldots$

$$f^{(4)}(x) = (-1)^5 2(-3)(1-x)^{-4}(-1)$$
$$= (-1)^7 3\cdot 2\,(1-x)^{-4}$$
$$\Rightarrow f^{(4)}(0) = -3!$$

$$f^{(k)}(0) = -(k-1)!$$

↳ seen by induction

$$f^{(k)}(x) = (-1)^{2(k-1)+1}(k-1)!\,(1-x)^{-k}$$

$$f^{(k+1)}(x) = \left(f^{(k)}(x)\right)' = (-1)^{2(k-1)+1}(-1)k\cdot(k-1)!\,(1-x)^{-(k+1)}(-1)$$
$$= (-1)^{2k+1}k!\,(1-x)^{-(k+1)}$$

In conclusion,

$$P_{n,0}(x) = f(0) + f'(0)x + \frac{1}{2}f''(0)x^2 + \frac{1}{3!}f'''(0)x^3 + \cdots$$
$$= 0 - x - \frac{1}{2}x^2 - \frac{1}{3}x^3 \cdots - \frac{1}{n}x^n$$
$$= \sum_{k=1}^{n} \frac{-1}{k} x^k \quad \lrcorner$$

single-variable

$f'(x)$
rate of change locally at a

multi-variable

partial derivatives

- Jacobian matrix $\left[\frac{\partial f}{\partial x_1}(a) \cdots \frac{\partial f}{\partial x_n}(a)\right]$
  - Linear transformation it encodes
- Differential $\mathbb{R}^n \xrightarrow{Df(a)} \mathbb{R}$, $v \mapsto Df(a)(v)$

linear transformation

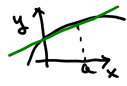
$Df(a)(v)$ the rate of change locally at a when we move from a along the vector v.

- $\text{grad} f(a) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(a) \\ \vdots \\ \frac{\partial f}{\partial x_n}(a) \end{bmatrix}$

"gradient vector"

tangent line

$y = f(a) + f'(a)(x-a)$



tangent space hyperplane

$f: \mathbb{R}^2 \to \mathbb{R}$

↳ tangent plane

$z = f(a) + \underbrace{Jf(a)}_{\text{matrix}} \begin{bmatrix} x-a_1 \\ y-a_2 \end{bmatrix}$

tangent plane equation — matrix-vector multiplication

in general ...

$z = f(a) + Jf(a)(x-a)$

tangent space $\cap$ $\mathbb{R}^{n+1}$

where

$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ $a = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$

$Jf(a) = \left[\frac{\partial f}{\partial x_1}(a), \cdots, \frac{\partial f}{\partial x_n}(a)\right]$

Approximation theorems

$f(x) \approx f(a) + f'(a)(x-a) + o(x-a)$

In general,

Taylor approximation

$P(x) = \sum \frac{1}{k!} f^{(k)}(a)(x-a)^k$

$\mathbb{R}^n \to \mathbb{R}$

$f(x) \approx f(a) + Jf(a)(x-a)$

↓ but we can go further

Quadratic approximation

$f(x) = f(a) + Jf(a)(x-a) + (x-a)^t (Hf(a))(x-a)$

quadratic term

Hessian matrix ($n \times n$)

↳ conveys the information of "second derivatives" of f ... but what are those?

$\frac{\partial}{\partial x_j}\left(\frac{\partial f}{\partial x_i}\right) =: \boxed{\frac{\partial^2 f}{\partial x_j \partial x_i}}$

↳ by definition

Definition

If all the second partial derivatives of $f: \mathbb{R}^n \to \mathbb{R}$ at a are defined the Hessian matrix of f at a is given by:

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

Under additional mild conditions

[e.g. all the second derivatives of f are continuous functions]

[e.g. all higher order derivatives of f are defined and continuous]

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \frac{\partial^2 f}{\partial x_j \partial x_i}(a)$$

In other words, $Hf(a)$ is symmetric.

Remark:

"mountain height"

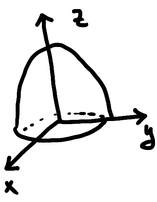
Level sets

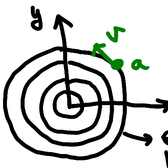
$f: \mathbb{R}^n \to \mathbb{R}$

$S_c = \{(x_1, \dots, x_n) \mid f(x_1, \dots, x_n) = c\}$

Level curves

$f: \mathbb{R}^2 \to \mathbb{R}$

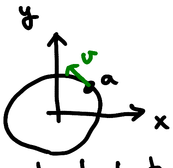




each curve has constant height

$S_c = \{(x,y) \mid f(x,y) = c\}$

Let $v \in \mathbb{R}^n$ tangent to the level set passing through a



$$v = \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix}$$

v is tangent to level set, but moving along the level set the value of the function does not change (*)

$$\underbrace{Df(a)}_{\text{differential of } f \text{ at } a}(v) = 0$$

$$\Leftrightarrow \left[\frac{\partial f}{\partial x_1}(a) \cdots \frac{\partial f}{\partial x_n}(a) \right] \begin{bmatrix} v_1 \\ \vdots \\ v_n \end{bmatrix} =$$

$$= \frac{\partial f}{\partial x_1}(a) v_1 + \cdots + \frac{\partial f}{\partial x_n}(a) v_n$$

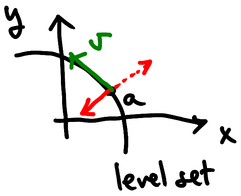
$$= \text{grad} f(a) \cdot v = 0$$

$\cdot$: dot product; $= 0$: by (*)

The gradient vector of at a, $\text{grad} f(a)$ is either

i) $\text{grad} f(a) = \vec{0}$

ii) perpendicular to v



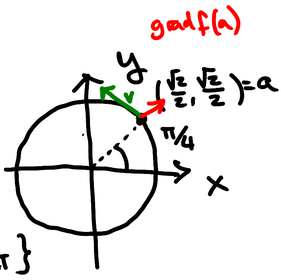
Example:

$$f(x,y) = x^2 + y^2$$

Level curves

Level curve of f through a is $S = \{(x,y) \mid x^2 + y^2 = 1\}$

$$= \{(\cos\alpha, \sin\alpha) \mid 0 \le \alpha < 2\pi\}$$



Take a tangent vector of S at a

$(-\sin\alpha, \cos\alpha)$ ← derivative or velocity vector

$$v = \left(-\sin\frac{\pi}{4}, \cos\frac{\pi}{4}\right) = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$$

Compute $\text{grad} f(a) = \left[2\frac{\sqrt{2}}{2}, 2\frac{\sqrt{2}}{2}\right] = [\sqrt{2}, \sqrt{2}]$

We can see:

$\text{grad} f(a) \perp v$

Quadratic approximations of $f: \mathbb{R}^n \to \mathbb{R}$

$$f(x) = f(a) + Jf(a)(x-a) + (x-a)^t Hf(a)(x-a)$$

Use this expression to reason about optimization

I want to find out $\hat{a} \in \mathbb{R}^n$ such that f attains a local max / min at $\hat{a}$.

Our candidates have to satisfy

$$Jf(\hat{a}) = [0 \ \dots \ 0] \quad \text{or equivalently}$$

$$\operatorname{grad} f(\hat{a}) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix} = \vec{0}$$

"critical point"

$$f(x) = f(a) + (x-a)^t Hf(a)(x-a)$$

Ej $f: \mathbb{R}^2 \to \mathbb{R}$ $\quad f(a) = 1$

$a = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\quad \operatorname{grad} f(a) = \vec{0}$

$$Hf(a) = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}$$

$$f(x) = 1 + [x \ y] \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$$

$$= 1 - x^2 + 2xy - y^2$$

is this representing some max / min / other?

Quadratic classification of critical points of $f: \mathbb{R}^n \to \mathbb{R}$

Let $\hat{a}$ be a critical point for f.

$Hf(a)$ has $\lambda_1, \dots, \lambda_n$ eigenvalues (possibly repeated)

Then the quadratic approximation allows us to conclude the following rules:

i) $\lambda_1, \dots, \lambda_n < 0 \Rightarrow$ local max

ii) $\lambda_1, \dots, \lambda_n > 0 \Rightarrow$ local min

iii) $\lambda_1 \cdots \lambda_n < 0 \Rightarrow$ "saddle point"

iv) $\lambda_1 \cdots \lambda_n = 0 \Rightarrow$ we do not know!