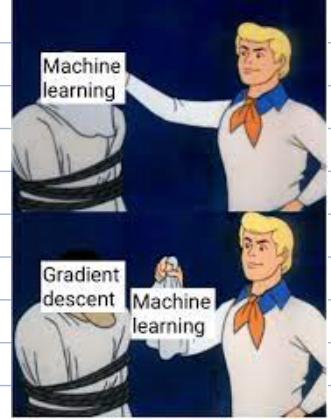


# TỐI ƯU HÓA

MT+56-2425



## . Giới thiệu & động lực

- . Gradient descent ("di chuyển xuống theo gradient")
- . Lagrange multiplier (phương pháp nhân tử Lagrange)

## Giới thiệu

Bài toán học từ dữ liệu hay học máy (ML)

$\theta$ : data

$f$ : giá trị ham

$L$ : ham đánh giá (loss function,  $\rightarrow$  1 giá trị thực)

Tìm  $\underset{f \in \mathcal{H}}{\operatorname{argmin}} L(\theta, f)$

$\underset{x \in D \subset \mathbb{R}^n}{\operatorname{minimize}} f(x)$

## Tối ưu hàm số

$\underset{x \in D \subset \mathbb{R}^n}{\operatorname{minimize}} f(x)$  với  $x \in D \subset \mathbb{R}^n$

Phụ thuộc vào:

1) tính chất của  $f$  (liên tục, khả vi, lồi, ...)

2) đặc điểm (hình học) của  $D$  (mở, compact, lồi, ...)

Ví dụ .  $f$  liên tục &  $D$  compact  $\Rightarrow f$  đạt cực trị toàn cục trên  $D$

. f kha<sup>2</sup> vi & D mo'  $\Rightarrow$  neu  $x^*$  la mat cuc tri cua f thi  $Df(x^*) = 0$

. f lo' & D lo'  $\Rightarrow$  cuc tri dia phuong la cuc tri toan cuc

Van de giao nghiem chinh xac cho  $Df(x) = 0$  rat kho!

Ví dụ  $f(x) = x^6 + x^5 + 2x^3 - x + 1$

$$f'(x) = 6x^5 + 5x^4 + 6x^2 - 1$$

Giai phap cac thuat toan xep xi tai vun

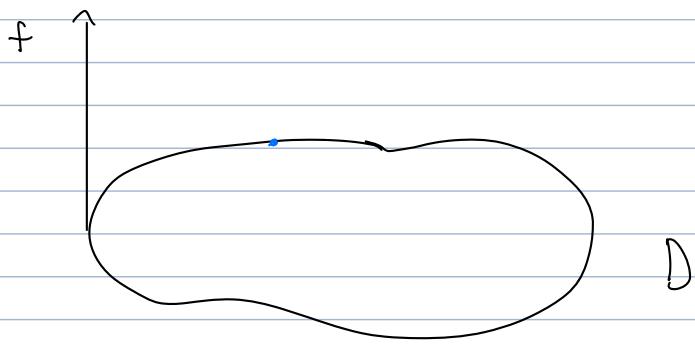
### Gradient descent (G D)

Dieu kien  $f: D \subset \mathbb{R}^n \rightarrow \mathbb{R}$  kha vi,  $x = (x_1, \dots, x_n)^T$

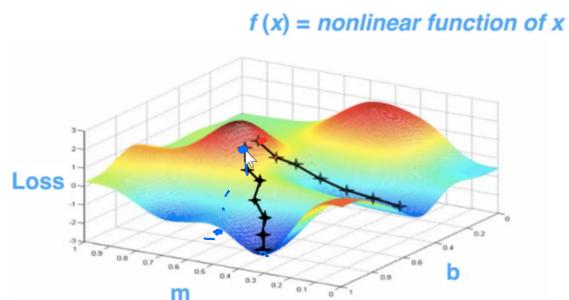
gradient  $\nabla f = Df^T = (\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n})^T$

### Y tuong

$-\nabla f(u)$  chi phuong co toc do giam lon nhat cua f tai u.



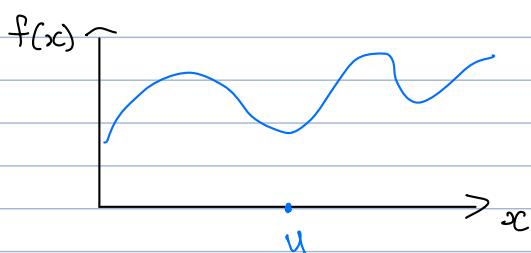
### Gradient Descent



### Thuat toan gradient descent (van tat)

Tai mot diem u bat ky, di theo huong  $-\nabla f$  de giam f voi toc do lon nhat.

Kí dụ (1 chiều)  $f: \mathbb{R} \rightarrow \mathbb{R}$



Gradient của  $f$  là gì?

Câu hỏi . xuất phát từ điểm nào?

- . di chuyển một đoạn dài như nào?
- . dừng khi nào?

### Thiết kế gradient descent

1) Chọn  $u_0 \in D$ ,  $\alpha > 0$  và  $\varepsilon > 0$

2)  $i = 0, 1, 2, \dots$

. If  $\|\nabla f(u_i)\| < \varepsilon$  : output  $u_i$  và dừng

. Else  $u_{i+1} := u_i - \alpha \nabla f(u_i)$

Kí dụ: gradient descent cho  $f(x) = x^2$ ,  $u_0 = 2$ ,  $\varepsilon = 1/2$

với  $\alpha = 1$  và  $\alpha = 1/2$

### Bài tập 1 (GD cho hồi quy tuyến tính)

Cho  $X \in \mathbb{R}^{n \times m}$ ,  $Y \in \mathbb{R}^{n \times 1}$

( $n$  điểm dữ liệu gồm 1 nhãn và  $m$  thông tin đặc trưng)

Hãy tìm vector hệ số  $w \in \mathbb{R}^m$  để minimize

$$f(w) = \|Xw - Y\|^2.$$

a) Chứng minh:  $\nabla f(w) = 2(X^T X w - X^T Y)$ .

b) Viết công thức cập nhật khi áp dụng GD cho  $f$ .

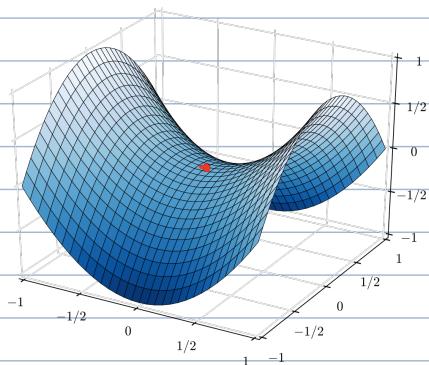
Vấn đề. 1 điểm xuất phát bất kỳ có thể ko cho kết quả đúng

Chỉ descent về cực trị điểm phẳng;

(local minimum)

tէ hơn, dùng  $\delta$  điểm yên ngửa.

(saddle point)



- tốc độ học phù hợp ?
- thuật toán có dùng ko ?

### GD có quán tính (momentum GD)

Ý tưởng thêm quán tính để thoát được những điểm yên ngửa hoặc cực trị đia phẳng (vùng lõm ko sâu)

+ Cụ thể, thêm biến  $\frac{m}{\tau}$  và  $\beta$  momentum

đóng quan trọng của momentum.

+  $m_0 = 0$  và  $\delta$  bước thứ i :

$$m_i := \beta m_{i-1} + \nabla f(u_i)$$

$$u_i := u_{i-1} - \alpha m_i$$

## Điều chỉnh tốc độ học

. Thay đổi theo thời gian:  $\alpha(i)$  cho bước thứ  $i$

Ví dụ .  $\alpha(i) = \alpha_0 c^i$  với  $c < 1$ : hàn mòn

.  $\alpha(i) = \alpha_0 (1 + \frac{i}{L})^{-N}$ : hàn lũy thừa

$\Rightarrow$  nhanh lúc đầu để tìm vùng có cực trị và chậm lúc sau để hội tụ về cực trị.

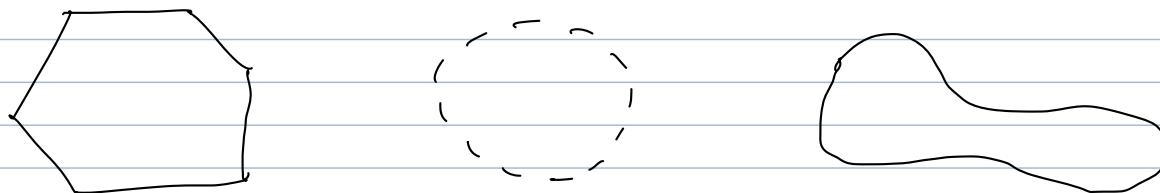
. Ngoài ra, có thể cập nhật  $\alpha(i)$  dựa trên  $\nabla f(u_i)$ ,  $u_i$ ,  $\nabla f(u_{i-1})$ ,  $u_{i-1}$  (ví dụ:  $\|\nabla f(u_i)\| > 0$  thì  $\alpha(i) \sim 0$ )

## Kết quả về tính đồng

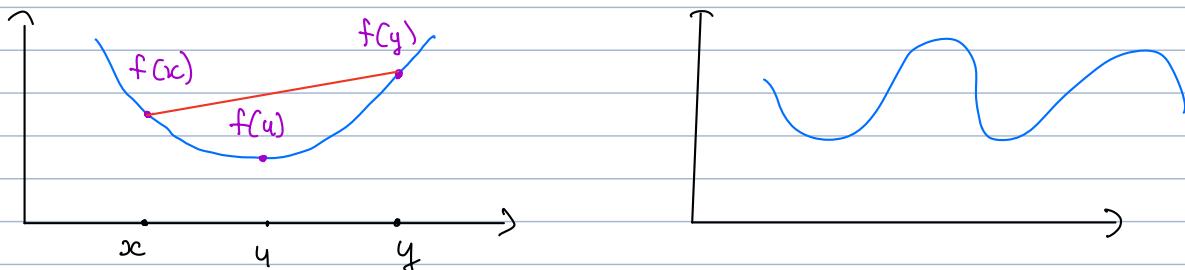
**Dinh lý** GD sẽ hội tụ về cực trị địa phương nếu:

1) D lồi (convex):  $\forall x, y \in D, t \in [0, 1]$

$$\Rightarrow tx + (1-t)y \in D$$



2) f lồi:  $f(tx + (1-t)y) \leq t f(x) + (1-t) f(y)$



3) f Lipschitz:  $\exists K > 0$  sao cho  $\forall x, y \in D$

$$\Rightarrow \|f(x) - f(y)\| \leq K \|x - y\|$$

$$\left( \frac{\|f(x) - f(y)\|}{\|x - y\|} \leq K \Rightarrow |\nabla f| \text{ bị chặn} \right)$$

Ví dụ: ko lipschitz  $f(x) = -e^x$



4)  $\alpha(i)$  thỏa mãn điều kiện Wolfe

### Phương pháp nhân tử Lagrange

#### Tối ưu hàm số

minimize  $f(x)$  với  $x \in D \subset \mathbb{R}^n$

.  $D$  mồi :  $D f = 0$  hoặc  $G D$

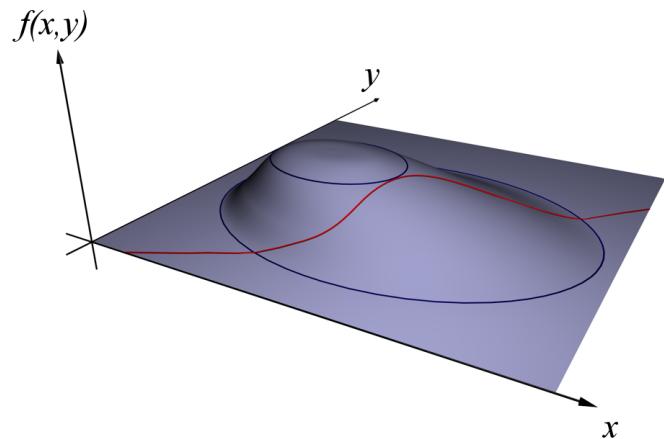
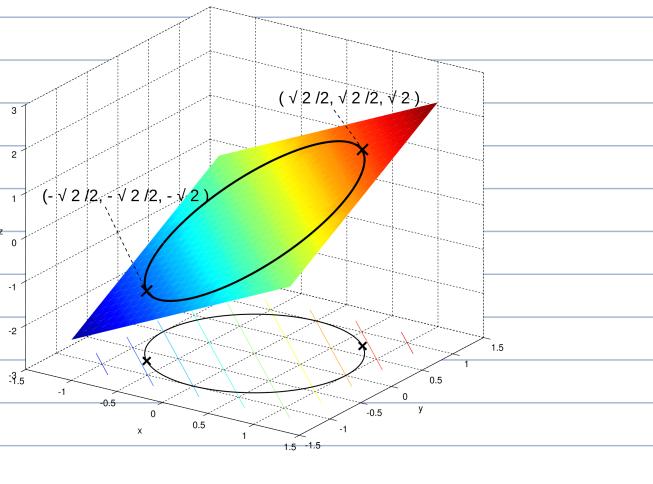
.  $D$  đóng : ko còn động

Cụ thể:  $D$  được xác định bởi các điều kiện (constraints)

$$g_1(x) = g_2(x) = \dots = g_m(x) = 0.$$

Ví dụ: minimize  $f(x, y) = x + y$  với  $\text{đk } x^2 + y^2 = 1$ .

(?)  $g(x, y)$  là hàm số nào? xác định  $D$ ?



Tương hợp riêng  $n = 2, m = 1$ .

Tìm cực trị của  $f(x, y) : \mathbb{R}^2 \rightarrow \mathbb{R}$  với điều kiện

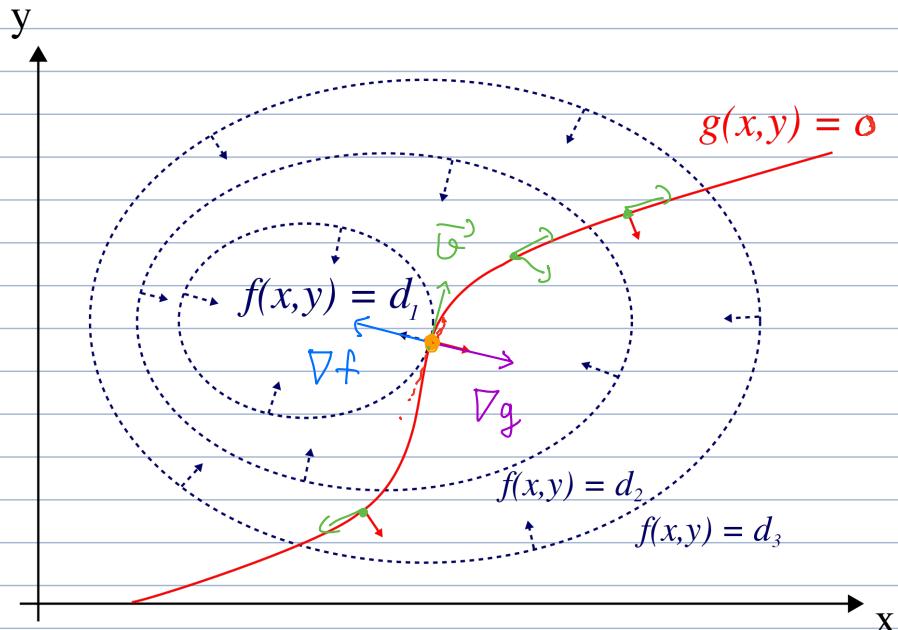
$$g(x, y) = 0.$$

Bố đề 1 Giả sử  $(x_0, y_0)$  là 1 điểm cực trị của  $f$  trên D

và  $\nabla g(x_0, y_0) \neq 0$ . Khi đó  $\exists \lambda \in \mathbb{R}$  sao cho :

$$\nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0).$$

C/M



Có  $\vec{v}$  là vector tiếp tuyến với D tại  $(x_0, y_0)$

$$\nabla f := \nabla f(x_0, y_0)$$

$$\nabla g := \nabla g(x_0, y_0)$$

$$\cdot \text{Bước 1: } \nabla g \cdot \vec{v} = 0 \Rightarrow \nabla g \perp \vec{v}$$

$$\cdot \text{Bước 2: } \nabla f \cdot \vec{v} = 0 \Rightarrow \nabla f \perp \vec{v}$$

$$\cdot \text{Bước 3: } \Rightarrow \nabla f(x_0, y_0) = \lambda \nabla g(x_0, y_0) \text{ for some } \lambda.$$

Hết quá

$$\left\{ \begin{array}{l} \nabla f(x_0, y_0) - \lambda \nabla g(x_0, y_0) = 0 \\ g(x_0, y_0) = 0 \end{array} \right.$$

Nhân tử Lagrange -

$$L(x, y, \lambda) = f(x, y) - \lambda g(x, y).$$

Nhân xét  $(x_0, y_0, \lambda)$  là điểm dừng của  $L$

$$\Rightarrow \text{là nghiệm của } \nabla L(x, y, \lambda) = 0$$

Nhân xét trên vẫn đúng cho  $n > 2$

Cho  $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$  khả vi và  $D = \{x \in \mathbb{R}^n : g(x) = 0\}$

Định lý 1 (Nhân tử Lagrange)

Giả sử  $x_0$  là một cực trị của  $f$  trên  $D$  và  $\nabla g(x_0) \neq 0$

$$\text{Đặt } L(x, \lambda) = f(x) - \lambda g(x).$$

Khi đó,  $\exists \lambda_0$  sao cho  $\nabla L(x_0, \lambda_0) = 0$ .

Bài tập 2 Tìm giá trị nhỏ nhất của  $f(x, y) = x + y$  trên

$$D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}$$

(Ghi ý: chia  $D$  thành một miền mở và một miền đóng)

Tối ưu hóa trong ML & tì riêng và vectơ riêng

Đạng 1 Cho ma trận  $A_{n \times n}$  đối xứng,  $x \in \mathbb{R}^n$

$$\text{Maximize } f(x) = x^T A x \quad (1)$$

$$\text{với điều kiện } x^T x = 1.$$

Bố đề Nhân tử Lagrange  $\lambda$  và nghiệm tối ưu  $x$  của (1)

là tì riêng và vectơ riêng của  $A$ .

$$\text{C/M } \mathcal{L}(x, \lambda) = x^T A x - \lambda(x^T x - 1)$$

$$\frac{\partial \mathcal{L}}{\partial x} = 2Ax - 2\lambda x$$

$$\frac{\partial \mathcal{L}}{\partial x} = 0 \Rightarrow Ax = \lambda x$$

□

Chuẩn ma trận  $\|A\|_{op} := \max_{\|x\|=1} \|Ax\|$   
 (Matrix norm)

Hệ quả  $\|A\|_{op} := \sqrt{\text{giá trị riêng lớn nhất của } A^T A}$

$$\begin{aligned} \text{C/M } \max_{\|x\|=1} \|Ax\| &= \max \sqrt{\|Ax\|^2} \text{ s.t. } \|x\|^2 = 1 \\ &= \max \sqrt{Ax \cdot Ax} \text{ s.t. } x^T x = 1 \\ &= \max \sqrt{(Ax)^T Ax} \text{ s.t. } x^T x = 1 \\ &= \max \sqrt{x^T A^T A x} \text{ s.t. } x^T x = 1 \end{aligned}$$

Giống dạng  $\hat{\sigma}(1) \Rightarrow$  2 véc x là tri & vector riêng của  $A^T A$ .

Lưu ý:  $A^T A$  là ma trận đối xứng & xác định dương:  $\lambda \geq 0$

$\Rightarrow$  Khi  $\|Ax\|$  đạt max:  $x^T A^T A x = x^T \lambda x = \lambda$

□

Dạng 2 Cho ma trận vuông  $n \times n$   $X$  và  $n \times n$   $A$  đối xứng  
 maximize trace( $X^T A X$ )  
 với điều  $X^T X = I_n$  (2)

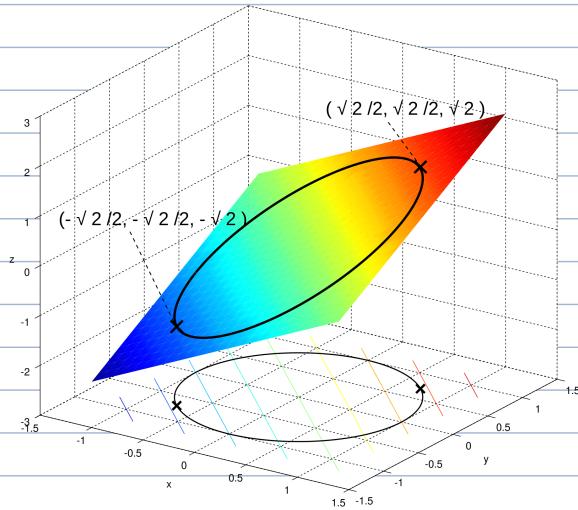
Hệ quả Nếu  $X$  là nghiệm tối ưu thì  $AX = XA$ .

Bài tập 3 Chứng minh tính chất trên.

## Sửa BT về nhân tử Lagrange

Bài tập 2 Tìm giá trị nhỏ nhất của  $f(x, y) = x + y$  trên

$$D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$



(Nhắc lại) Cực trị hàm  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  trên  $D \subset \mathbb{R}^n$

phù thuộc vào tắc điểm hình học của  $D$ . Cụ thể:

(1)  $D$  mở: dùng  $Df(x) = \nabla f(x)^T = 0$  - first derivative test

và Hessian matrix  $(Hf)_{i,j}$  - second derivative test.

đóng →

(2)  $D = \{x \mid g(x) = 0\}$ : dùng nhân tử Lagrange

ĐLý  $x^*$  là cực trị của  $f$  trên  $D$  &  $\nabla g(x^*) \neq 0$

$\Rightarrow \exists \lambda^*$  sao cho  $(x^*, \lambda^*)$  là điểm đóng của  
 $L(x, \lambda) = f(x) - \lambda g(x)$ .

(3)  $D$  compact:  $f(x)$  đạt GTNN và GTLN trên  $D$

đóng và bị chặn

Lời giải:  $D_1 = \{x \mid g(x) = 0\}$

↓

$D_2$  mở

↙

Bước 1:  $D = \{(x, y) \mid x^2 + y^2 = 1\} \cup \{(x, y) \mid x^2 + y^2 < 1\}$  là tập

compact. Tùy (3)  $\Rightarrow$  f đạt GTNN tại  $(x^*, y^*)$  nào đó trên D. Vậy  $(x^*, y^*) \in D_1$  hoặc  $D_2$ .

Bước 2: Nếu  $(x^*, y^*) \in D_2$ , tùy (1)  $\Rightarrow \nabla f(x^*, y^*) = 0$  (vì lý do  $\nabla f = (1, 1)$ ). Vậy  $(x^*, y^*) \in D_2$

Bước 3: Ta biết  $(x^*, y^*) \in D_1 = \{(x, y) \mid g(x, y) = 0\}$  là một điểm cực trị vì GTNN trên D cũng phải là GTNN trên  $D_1$ . Bên cạnh đó,  $\nabla g(x, y) = 2 \begin{pmatrix} x \\ y \end{pmatrix} \neq 0$  trên  $D_1$   $\Rightarrow$  áp dụng tính lý (2): tồn tại  $\lambda^*$  sao cho  $(x^*, y^*, \lambda^*)$  là nghiệm của  $\nabla L(x, y, \lambda) = 0$ .

Ta có hệ pt:  $\begin{cases} \nabla f(x, y) - \lambda \nabla g(x, y) = 0 \\ g(x, y) = 0 \end{cases}$

$$\Leftrightarrow \begin{cases} (1, 1)^T - \lambda (2x, 2y)^T = 0 \\ x^2 + y^2 = 1 \end{cases}$$

$$\Leftrightarrow \begin{cases} x = y = \lambda = \frac{1}{\sqrt{2}} \\ x = y = \lambda = -\frac{1}{\sqrt{2}} \end{cases}$$

GTNN phải là điểm có  $f(x, y)$  nhỏ hơn nên GTNN là  $-\sqrt{2}$  tại  $(x, y) = (-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ .  $\square$

Nhiều hơn một điều kiện

Cho các hàm khả vi  $f, g_1, \dots, g_m: \mathbb{R}^n \rightarrow \mathbb{R}$  và

tập  $D = \{x \in \mathbb{R}^n \mid g_1(x) = \dots = g_m(x) = 0\}$ .

Đặt  $S(x) := \text{span} \{ \nabla g_1(x), \nabla g_2(x), \dots, \nabla g_m(x) \}$

phù thuộc  $x$ .

Bố đề 2 Nếu  $x^*$  là một điểm cực trị của  $f$  trên  $D$  và  $S(x^*) \neq 0$  thì  $\nabla f \in S(x^*)$ .  
(lưu ý: Bố đề 1 là trường hợp riêng).

Chứng minh

Bước 1 Những phương di chuyển từ  $x^*$  mà vẫn nằm trong  $D$  là những phương trong  $S(x^*)^\perp$   
(vì  $\perp$  với mỗi  $\nabla g_i(x^*)$  nên  $g_i$  sẽ không đổi?)

Bước 2 Những phương di chuyển bước từ  $x^*$  trong  $D$  phải  $\perp$  với  $\nabla f(x^*)$  vì  $f(x^*)$  ko thể tăng hoặc  
đó  $x^*$  là cực trị. Do đó,  $\nabla f(x^*) \in (S(x^*)^\perp)^\perp$ .

Ta có:  $(S(x^*)^\perp)^\perp = S(x^*)$

(một kết quả ko hiển nhiên trong đại số tuyến tính).

Từ đó suy ra  $\nabla f(x^*) \in S(x^*)$ .  $\square$

Nhận xét  $\nabla f(x^*) \in S(x^*) \Rightarrow \exists \lambda_1, \dots, \lambda_m \in \mathbb{R}$

sao cho:  $\nabla f(x^*) = \lambda_1 \nabla g_1(x^*) + \dots + \lambda_m \nabla g_m(x^*)$ .

Ta có định lý cho nhận tử Lagrange với nhiều điều kiện

Định lý 2 (Nhận tử Lagrange)

Đặt  $L(x, \lambda_1, \dots, \lambda_m) = f(x) - \sum_{i=1}^m \lambda_i g_i(x)$

Giả sử  $x^*$  là một cực trị của  $f$  trên  $D$  và

$S(x^*) \neq 0$ . Khi đó  $\exists \{\lambda_i^*\}$  sao cho

$(x^*, \lambda_1^*, \dots, \lambda_m^*)$  là điểm dừng của  $L$ :

$$\nabla L(x^*, \lambda_1^*, \dots, \lambda_m^*) = 0. \quad \square$$

Lưu ý Phương pháp nhân tử Lagrange thường dùng để  
tìm ra các ứng viên cho các từ (tiêu kiện cần);  
tiêu kiện đủ thường cần thêm một số kết quả và  
suy luận (như Bài tập 2 ở trên)