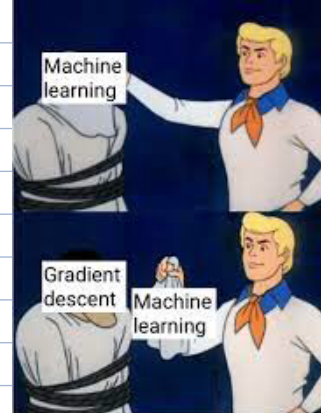


TỐI ƯU HÓA

TTMA 2021

Tung Cui, Caltech



- . Giới thiệu & tổng lược
- . Gradient descent ("di chuyển xuống theo gradient")
- . Lagrange multiplier (phương pháp nhân tử Lagrange)

Giới thiệu

Previously on PiMA: ML trong DS

Θ : data

L : hàm đánh giá (loss function) = 1 giá trị thực

\mathcal{H} : tập hàm ứng viên

$$\text{Tìm } \underset{F \in \mathcal{H}}{\operatorname{argmin}} L(\Theta, F) \iff \underset{x \in D \subset \mathbb{R}^n}{\operatorname{minimize}} f(x)$$

Tối ưu hàm số

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

Phụ thuộc vào:

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

2) đặc điểm (hình học) của D (mở, ^{đóng + bị chặn} compact, lồi, ...)

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

. f khả vi & D mở \Rightarrow nếu x^* là một cực trị của f thì $Df(x^*) = 0$

. f lồi & D lồi \Rightarrow cực trị địa phương là cực trị toàn cầu

Vấn đề giải nghiệm chính xác cho $Df(x) = 0$ rất khó!

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

$f'(x) = 6x^5 + 5x^4 + 6x^2 - 1$: không có công thức nghiệm tổng quát

Giải pháp các thuật toán xấp xỉ tối ưu : gradient descent

Gradient descent (GD) Tham khảo Tối ưu hóa PIMA 18 & 19

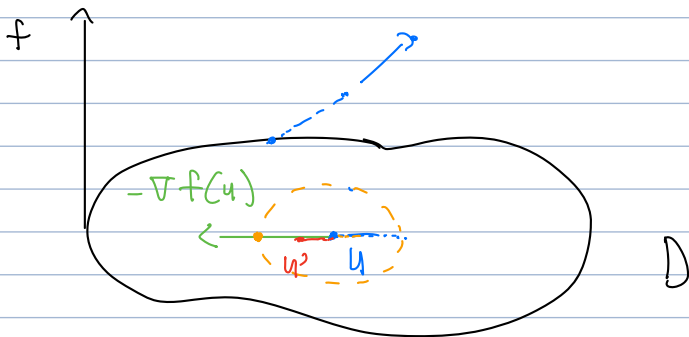
- Nhật Phạm, Thiên Lê

Điều kiện $f: D \subset \mathbb{R}^n \rightarrow \mathbb{R}$ khả vi, $x = (x_1, \dots, x_n)^T$

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

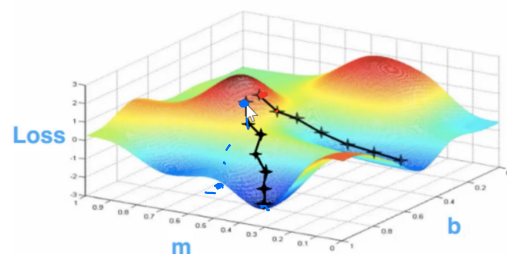
Ý tưởng

$-\nabla f(u)$ chỉ phương có tốc độ giảm lớn nhất của f tại u .



Gradient Descent

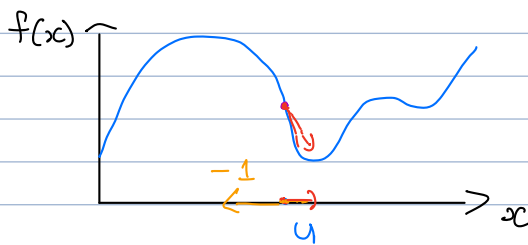
$f(x) = \text{nonlinear function of } x$



Thuật toán gradient descent (vấn tắt)

Tại một điểm u bất kỳ, đi theo hướng $-\nabla f$ để giảm f với tốc độ lớn nhất.

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



(?) Gradient của f là gì?

$$\nabla f = f' = -1$$

$$f'(u) < 0,$$

- Câu hỏi
- xuất phát từ điểm nào?
 - di chuyển một đoạn dài như nào? α
 - dừng khi nào? $-\|\nabla f\| \leq \dots \leq \|\nabla f\|$

Thuật toán 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 $u_{i+1} := u_i - \underbrace{\alpha}_{\text{tốc độ học (learning rate)}} \nabla f(u_i)$

tốc độ học (learning rate)

Ví 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)

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

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

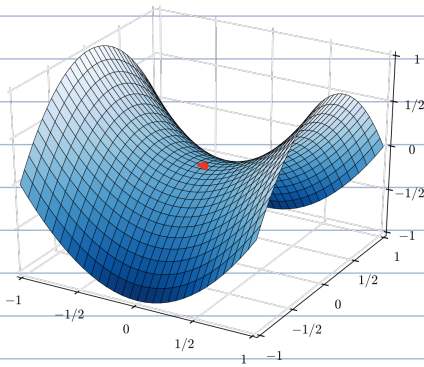
a) Chứng minh: $\nabla f(w) = 2(X^T X w - X^T Y)$.

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

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

Chỉ descent về cực trị địa phương ;
(local minimum)

tệ hơn, dừng ở điểm yên ngựa.
(saddle point)



- tốc độ học phù hợp ?
- thuật toán có dừng ko ? ($\nabla f \rightarrow 0$?)

GD có quán tính (momentum GD)

Ý tưởng thêm quán tính để thoát được những điểm yên ngựa hoặc cực trị địa phương (vùng lõm ko sâu)

+ Cụ thể, thêm biến \underline{m} và $\underline{\beta}$ \leftarrow
momentum độ quan trọng của momentum.

+ $m_0 = 0$ và ở bước thứ i :

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

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

Đã sửa

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 $c \leq 1$: hàm mũ

. $\alpha(i) = \alpha_0 \left(1 + \frac{i}{x}\right)^{-N}$: hàm lũy thừa

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

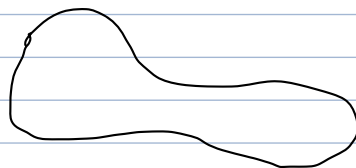
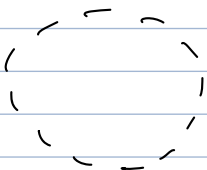
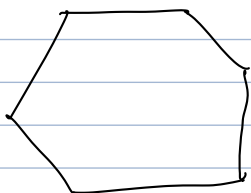
. Ngoài ra, có thể cập nhật $\alpha(i)$ dựa trên $\nabla f(u_i)$, u_i , $\nabla f(u_{i-1})$, u_{i-1} (ví dụ: $\|\nabla f(u_i)\| \gg 0$ thì $\alpha(i) \sim 0$)

Kết quả về tính đúng

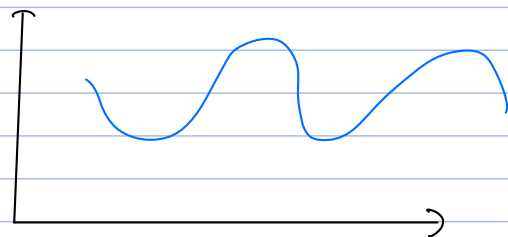
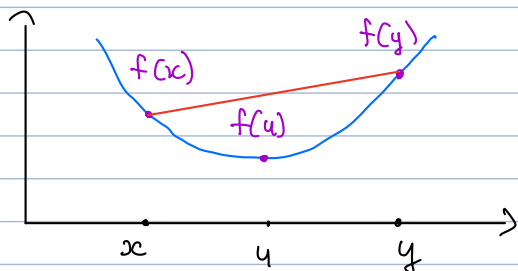
Định lý GD sẽ hội tụ về cực trị địa phương nếu:
toàn cục.

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

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



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

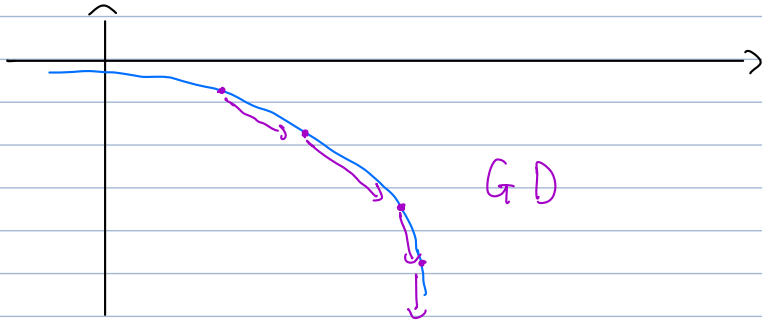


3) f Lipschitz: $\exists K > 0$ sao cho $\forall x, y \in D$

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

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

Ví dụ: không Lipschitz $f(x) = -e^x$



4) $x(i)$ thỏa mãn điều kiện KKT

Phương pháp nhân tử Lagrange

Tối ưu hàm số

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

• D mở: $D \neq \emptyset$ hoặc $G.D$

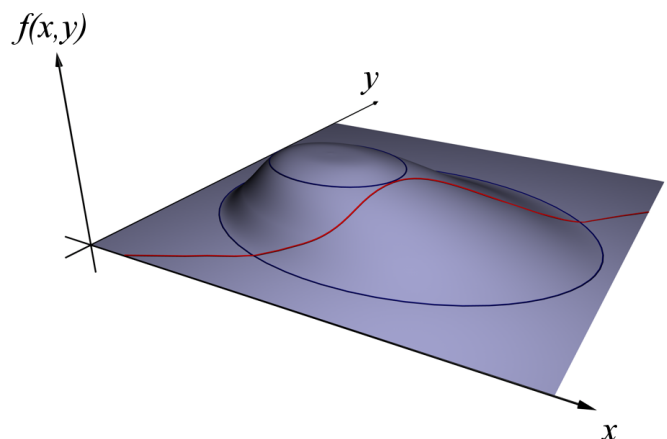
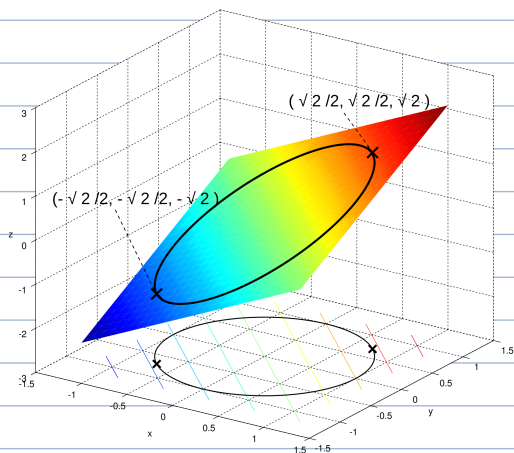
• D đóng: không còn đúng

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 đk $x^2 + y^2 = 1$.

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



Tương hợp Lagrange $n = 2$, $m = 1$.

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

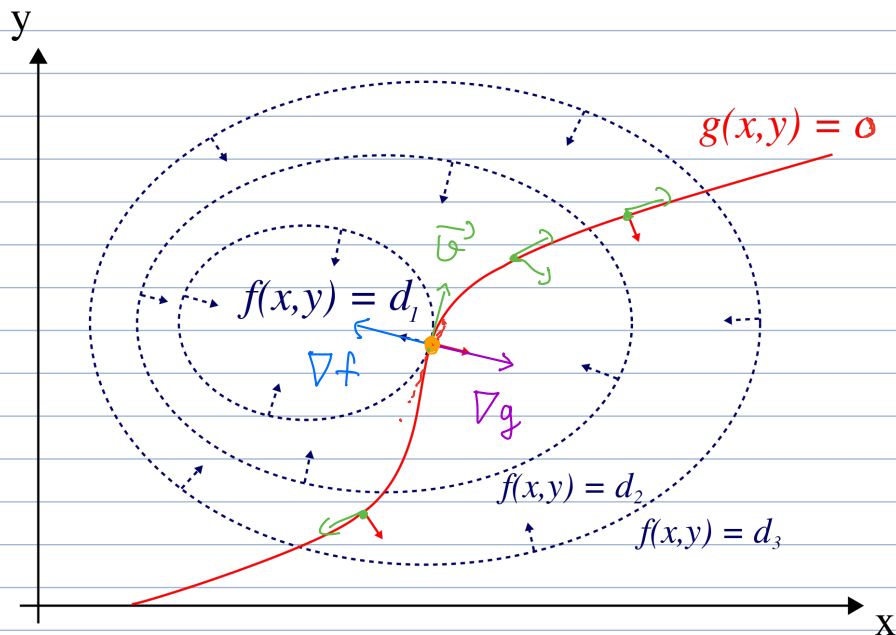
$$g(x, y) = 0.$$

Bổ 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



Chọn \vec{w} là vector tiếp tuyến với D tại (x_0, y_0)

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

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

• Bước 1: $\nabla g \cdot \vec{w} = 0 \Rightarrow \nabla g \perp \vec{w}$

• Bước 2: $\nabla f \cdot \vec{w} = 0 \Rightarrow \nabla f \perp \vec{w}$

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

Hệ quả

$$\begin{cases} \nabla f(x_0, y_0) - \lambda \nabla g(x_0, y_0) = 0 \\ g(x_0, y_0) = 0 \end{cases}$$

Nhân tử Lagrange — $L(x, y, \lambda) = f(x, y) - \lambda g(x, y).$

Nhận xét (x_0, y_0, λ) là điểm dừng của L
 \Rightarrow là nghiệm của $\nabla L(x, y, \lambda) = 0$

Nhận xét tiên 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 $\nabla g(x_0) \neq 0$

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

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

(x_0, λ_0) là điểm dừng của $L(x, \lambda)$

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

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

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

Tối ưu hóa trong ML & trị riêng và vector riêng

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

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

$$\text{với đk } x^T x = 1.$$

Bổ đề Nhân tử Lagrange λ và nghiệm tối ưu x của (1)
là trị riêng và vector riêng của A .

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

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

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

□

(Nhắc lại) $\|A\|_{op} = \max_{\|x\|=1} \|Ax\| \rightarrow$ khoảng cách Euclidean

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

C/M

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

Giống dạng B (1) $\Rightarrow \lambda$ và x là trị & vector riêng của $A^T A$.

Nhắc lại $A^T A$ là ma trận xác định dương: $\lambda > 0$

\Rightarrow Khi $\|Ax\| \max$: $x^T A^T A x = x^T \lambda x = \lambda$ → hệ song □

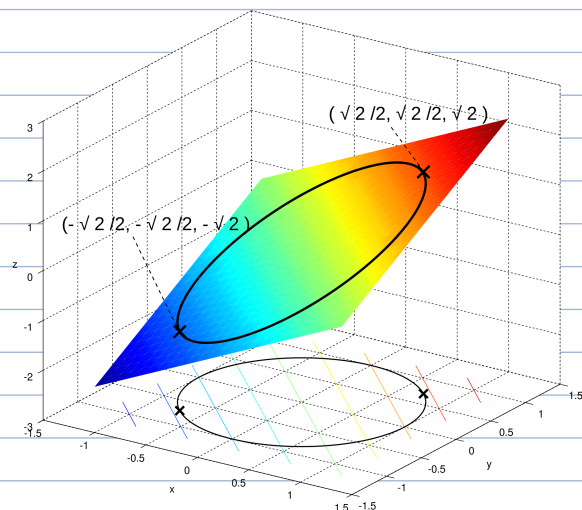
Dạng 2 Cho ma trận vuông $n \times n$ X và $n \times n$ A đối xứng
 maximize trace($X^T A X$)
 với đk $X^T X = I_n$ (2)

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



(Nhắc lại) Cực trị hàm khả vi $f: \mathbb{R}^n \rightarrow \mathbb{R}$ trên $D \subset \mathbb{R}^n$ phụ thuộc vào đặc 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 \rightarrow

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

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

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

(3) D compact : $f(x)$ đạt GTNN và GTLN trên D
 \leftarrow đóng và bị chặn

Lời giải:

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

\searrow

D_2 mở
 \swarrow

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

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

Bước 2: Nếu $(x^*, y^*) \in D_2$, từ (1) $\Rightarrow \nabla f(x^*, y^*) = 0$ (vô lý do $\nabla f = (1, 1)$). Vậy $(x^*, y^*) \in D_1$.

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 định lý (2): tồn tại λ^* sao cho (x^*, y^*, λ^*) 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 $S(x) := \text{span} \{ \nabla g_1(x), \nabla g_2(x), \dots, \nabla g_m(x) \}$
 $\quad \quad \quad \nwarrow$ phụ thuộc x .

Bổ 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: Bổ 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 mọi $\nabla g_i(x^*)$ nên g_i sẽ không đổi)

Bước 2 Những phương di chuyển được từ x^* trong D phải \perp với $\nabla f(x^*)$ vì $f(x^*)$ ko thể tăng hoặc do 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

Nhân 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 nhân tử Lagrange với nhiều điều kiện

Định lý 2 (Nhân 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 :

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

□

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