Cơ sở Toán học cho Machine Learning

Thân Quang Khoát

Nguyễn Văn Sơn

Viện Công nghệ thông tin và Truyền thông
Trường Đại Học Bách Khoa Hà Nội

Năm 2020

Phần 1
Đại số tuyến tính

Chuyển vị và Hermitian

- Cho $A \in R^{m \times n}$, ta nói $B \in R^{n \times m}$ là chuyển vị của A nếu:

$$b_{ij} = a_{ji} \; \forall 1 \le i \le n, 1 \le j \le m$$

Ký hiệu: $B = A^T$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \Rightarrow \mathbf{x}^T = \begin{bmatrix} x_1 \; x_2 \; \dots \; x_m \end{bmatrix}; \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \Rightarrow \mathbf{A}^T = \begin{bmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ a_{12} & a_{22} & \dots & a_{m2} \\ \dots & \dots & \ddots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{bmatrix}$$

Nếu $A = A^T$ thì ta gọi A là ma trận đối xứng

- Cho $A \in R^{m \times n}$, ta nói $B \in R^{n \times m}$ là chuyển vị liên hợp của A nếu:

$$b_{ij} = \overline{a_{ji}} \; \forall 1 \le i \le n, 1 \le j \le m$$

Ký hiệu: $B = A^H$

Nếu $A = A^H$ thì ta gọi A là ma trận Hermitian

Phép nhân hai ma trận

- Cho hai ma trận $A \in R^{m \times n}, B \in R^{n \times p}$, tích của hai ma trận được ký hiệu là $C \in R^{m \times p}$ với:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, 1 \leq i \leq m, 1 \leq j \leq p$$

Tính chất:

- Phép nhân hai ma trận không có tính giao hoán: $AB \neq BA$
- Tính kết hợp: $ABC = (AB)C = A(BC)$
- Tính phân phối đối với phép cộng: $A(B + C) = AB + AC$
- $(AB)^T = A^T B^T$

Ma trận đơn vị, Ma trận nghịch đảo

- Một ma trận vuông với các phần tử trên đường chéo chính bằng 1, còn lại bằng 0 được gọi là ma trận đơn vị, và ký hiệu là I_n.

- Cho một ma trận vuông $A \in R^{n \times n}$, nếu tồn tại ma trận vuông $B \in R^{n \times n}$ sao cho: $AB = I_n$ thì ta nói A là khả nghịch và B được gọi là ma trận nghịch đảo của A.

Ký hiệu $B = A^{-1}$.

Tính chất:

- $A.A^{-1} = I_n$
- $(AB)^{-1} = B^{-1}A^{-1}$

Định thức

- **Định nghĩa:** Định thức của một ma trận vuông A được ký hiệu là $detA$
  - Với $n = 1$, detA chính là phần tử duy nhất của ma trận đó
  - Với một ma trận vuông bậc $n > 1$:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \Rightarrow \det(\mathbf{A}) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(\mathbf{A}_{ij})$$

Với A_{ij} là ma trận thu được bằng cách xoá hang thứ i và cột thứ j của ma trận A, hay còn gọi là phần bù đại số của A ứng với phần tử ở hang i, cột j.

Định thức

❑ Tính chất:

- $detA = detA^T$
- $detI_n = 1$
- $\det(AB) = detA.detB$
- $detA^{-1} = \frac{1}{detA}$
- Nếu một ma trận có một hang hoặc một cột là một vecto 0 thì định thức của nó bằng 0
- Một ma trận là khả nghịch khi và chỉ khi định thức của nó khác 0
- Định thức của một ma trận tam giác (vuông) bằng tích các phần tử trên đường chéo chính

Tổ hợp tuyến tính-Không gian sinh

❑ Tổ hợp tuyến tính

Cho các vecto khác không $a_1, \dots, a_n \in R^m$ và các số thực $x_1, x_2, \dots, x_n$. Khi đó vecto:

$$b = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n$$

được gọi là một tổ hợp tuyến tính của $a_1, \dots, a_n \in R^m$.

Xét ma trận $A = [a_1, a_2, \dots, a_n] \in R^{m \times n}$ và $x = [x_1, x_2, \dots, x_n]^T$, ta có thể viết lại:

$$b = Ax$$

và b là một tổ hợp tuyến tính các cột của A

Tổ hợp tuyến tính-Không gian sinh

- Tập hợp tất cả các vecto có thể biểu diễn được như là một tổ hợp tuyến tính của các vecto khác không $a_1, \dots, a_n \in R^m$ được gọi là không gian sinh (span space) của hệ các vecto đó, và được ký hiệu là $\text{span}(a_1, \dots, a_n)$

- Nếu phương trình:

$$x_1 a_1 + x_2 a_2 + \cdots + x_n a_n = 0$$

Có nghiệm duy nhất $x_1 = x_2 = \cdots = x_n = 0$ thì ta nói hệ $\{a_1, a_2, \dots, a_n\}$ là độc lập tuyến tính. Ngược lại ta nói hệ đó là phụ thuộc tuyến tính.

Cơ sở của một không gian

- Một hệ các vecto $\{a_1, a_2, \dots, a_n\}$ trong không gian vecto m chiều $V = R^m$ được gọi là một cơ sở nếu hai điều kiện sau được thoả mãn:
  - $V \equiv span(a_1, a_2, \dots, a_n)$
  - $\{a_1, a_2, \dots, a_n\}$ là một hệ độc lập tuyến tính

→ Nhận thấy: n=m

Khi đó, mọi vecto $b \in V$ đều có thể biểu diễn duy nhất dưới dạng một tổ hợp tuyến tính của các a_i

Hạng của ma trận

- Xét một ma trận $A \in R^{m \times n}$. Hạng (rank) của ma trận này, ký hiệu là rank(A), được định nghĩa là số lượng lớn nhất các cột của nó tạo thành một hệ độc lập tuyến tính

- **Tính chất:**

- $rank(A) = rank(A^T)$
- Nếu $A \in R^{m \times n}$ thì $rank(A) \leq \min(m, n)$
- $rank(AB) \leq \min(rank(A), rank(B))$
- $rank(A + B) \leq rank(A) + rank(B)$
- Nếu $A \in R^{m \times n}, B \in R^{n \times k}$ thì: $rank(A) + rank(B) - n \leq rank(AB)$
- Nếu A là một ma trận vuông khả nghịch thì $rank(A) = n$

Hệ trực chuẩn, ma trận trực giao

- Một hệ cơ sở $\{u_1, u_2, \dots, u_m\} \in R^m$ được gọi là trực giao nếu:

$$u_i \neq 0 \text{ và } u_i^T u_j = 0 \ \forall 1 \leq i \neq j \leq m$$

- Một hệ cơ sở $\{u_1, u_2, \dots, u_m\} \in R^m$ được gọi là trực chuẩn nếu:

$$\|u_i\|_2^2 = u_i^T u_i = 1 \text{ và } u_i^T u_j = 0 \ \forall 1 \leq i \neq j \leq m$$

- Gọi $U = [u_1, u_2, \dots, u_m]$ với $\{u_1, u_2, \dots, u_m\} \in R^m$ là một hệ trực chuẩn thì $UU^T = U^T U = I_{\mathrm{m}}$.

Ngược lại nếu một ma trận U thoả mãn: $UU^T = U^T U = I_{\mathrm{m}}$ thì U được gọi là ma trận trực giao.

Trị riêng và vector riêng

- Cho một ma trận vuông $A \in R^{n \times n}$, một vecto khác không $x \in R^n$ và một số vô hướng (có thể thực hoặc phức) λ. Nếu $Ax = \lambda x$ thì ta nói λ và x là một cặp trị riêng, vector riêng của ma trận A

- **Tính chất:**

- Nếu x là một vecto riêng của A ứng với λ thì kx với $k \neq 0$ cũng là vecto riêng ứng với λ.
- Tích tất cả các giá trị riêng của một ma trận bằng định thức của ma trận đó. Tổng tất cả các giá trị riêng của một ma trận bằng tổng các phần tử trên đường chéo của ma trận đó
- Mọi ma trận vuông bậc n đều có n trị riêng (thực hoặc phức, kể cả lặp)

Chéo hoá ma trận

- Giả sử $x_1, \dots, x_n \neq 0$ là các vecto riêng của một ma trận vuông A ứng với các giá trị riêng $\lambda_1, \dots, \lambda_n$

Đặt $\Lambda = diag(\lambda_1, \dots, \lambda_n)$ và $X = [x_1, \dots, x_n]$ ta sẽ có: $AX = X\Lambda$

Hơn nữa nếu các giá trị riêng x_i là độc lập tuyến tính thì X là một ma trận khả nghịch, do đó:

$$A = X\Lambda X^{-1}$$

Do Λ là một ma trận đường chéo nên biểu diễn trên được gọi là chéo héo ma trận

Chéo hoá ma trận

❑ Tính chất:

- Chéo hoá ma trận chỉ áp dụng với ma trận vuông
- Một ma trận vuông bậc n là chéo hoá được iff nó có đủ n trị riêng độc lập tuyến tính
- Chéo hoá ma trận giúp tính toán dễ dang các A^k

$$A^2 = (X\Lambda X^{-1})(X\Lambda X^{-1}) = X\Lambda^2 X^{-1}$$

$$A^k = X\Lambda^k X^{-1}$$

Nếu A khả nghịch: $A^{-1} = (X\Lambda X^{-1})^{-1} = X\Lambda^{-1}X^{-1}$

Ma trận xác định dương

- **Chỉ xét trên họ các ma trận đối xứng**

- Một ma trận đối xứng A bậc n được gọi là xác định dương nếu: $x^T Ax > 0 \; \forall x \neq 0$
- Một ma trận đối xứng A bậc n được gọi là bán xác định dương nếu: $x^T Ax \geq 0 \; \forall x \neq 0$

- **Tính chất:**

- Mọi giá trị riêng của một ma trận đối xứng xác định dương đều là một số thực dương
- $A = B^T B$ là ma trận bán xác định dương mới mọi ma trận B bất kỳ
- …

Chuẩn của ma trận

- Với một ma trận $A \in R^{m \times n}$, chuẩn thường dung nhất là chuẩn Frobenius, ký hiệu là $\|A\|_F$ là căn bậc hai của tổng bình phương tất cả các phần tử của ma trận đó

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

Vết của ma trận

- **Định nghĩa:** Vết của một ma trận vuông là tổng tất cả các phần tử trên đường chéo chính của nó, được ký hiệu là trace(A).
- **Tính chất**:

- $trace(A) = trace(A^T)$
- $trace(\sum_{i=1}^{k} A_i) = \sum_{i=1}^{k} trace(A_i)$
- $trace(A) = \sum_{i=1}^{n} \lambda_i$ với λ_i là các giá trị riếng của A
- $trace(AB) = trace(BA)$
- Nếu X là một ma trận khả nghịch cùng chiều với ma trận vuông A thì: $trace(XAX^{-1}) = trace(X^{-1}XA) = trace(A)$
- $trace(A^T A) = trace(AA^T) = \|A\|_F^2 \geq 0$ với A là ma trận bất kỳ

Phần 2
Giải tích

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một số vô hướng

Đạo hàm (gradient) của một hàm số: $f(x)$: $R^n \rightarrow R$ theo vecto x được định nghĩa như sau:

$$\nabla_{\mathbf{x}} f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^n$$

Trong đó $\frac{\partial f(x)}{\partial x_i}$ là đạo hàm của hàm số theo thành phần thứ I của vecto x. Đạo hàm này được lấy khi giả sử tất cả các biến còn lại là hằng số

Đạo hàm của hàm nhiều biến

- **Hàm cho giá trị là một số vô hướng**

Đạo hàm bậc hai (second-order gradient) của hàm số trên còn được gọi là Hessian và được định nghĩa như sau:

$$\nabla^2 f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{S}^n$$

Với $S^n \in R^{n \times n}$ là tập các ma trận vuông đối xứng $n \times n$

Đạo hàm của hàm nhiều biến

- **Hàm cho giá trị là một số vô hướng**

Đạo hàm của một hàm số $f(X)$: $R^{n\times m} \to R$ theo ma trận X được định nghĩa là:

$$\nabla^2 f(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix} \in \mathbb{S}^n$$

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một vecto

Giả sử một hàm số với đầu vào là một số thực $v(x)$: $R \rightarrow R^n$:

$$v(x) = \begin{bmatrix} v_1(x) \\ v_2(x) \\ \vdots \\ v_n(x) \end{bmatrix}$$

Đạo hàm bậc nhất và bậc hai của nó là một vecto hàng như sau:

$$\nabla v(x) \triangleq \begin{bmatrix} \frac{\partial v_1(x)}{\partial x} & \frac{\partial v_2(x)}{\partial x} & \cdots & \frac{\partial v_n(x)}{\partial x} \end{bmatrix}$$

$$\nabla^2 v(x) \triangleq \begin{bmatrix} \frac{\partial^2 v_1(x)}{\partial x^2} & \frac{\partial^2 v_2(x)}{\partial x^2} & \cdots & \frac{\partial^2 v_n(x)}{\partial x^2} \end{bmatrix}$$

Đạo hàm của hàm nhiều biến

❑ Hàm cho giá trị là một vecto

Nếu đầu vào cũng là một vecto, tức có hàm số $h(x)$: $R^k \rightarrow R^n$ thì đạo hàm của nó là một ma trận kxn

$$\nabla h(\mathbf{x}) \triangleq \begin{bmatrix} \frac{\partial h_1(\mathbf{x})}{\partial x_1} & \frac{\partial h_2(\mathbf{x})}{\partial x_1} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_1} \\ \frac{\partial h_1(\mathbf{x})}{\partial x_2} & \frac{\partial h_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial h_1(\mathbf{x})}{\partial x_k} & \frac{\partial h_2(\mathbf{x})}{\partial x_k} & \cdots & \frac{\partial h_n(\mathbf{x})}{\partial x_k} \end{bmatrix}$$
$$= \left[\nabla h_1(\mathbf{x})\ \nabla h_2(\mathbf{x}) \ldots\ \nabla h_n(\mathbf{x})\right] \in \mathbf{R}^{k \times n}$$

Đạo hàm của hàm nhiều biến

❑ Tính chất quan trọng

Để cho tổng quát, ta giả sử biến đầu vào là một ma trận và các hàm số có chiều phù hợp để các phép nhân thực hiện được

- Product rules:

$$\nabla\big(f(X)^T g(X)\big) = \big(\nabla f(X)\big)g(X) + \big(\nabla g(X)\big)f(X)$$

- Chain rules:

$$\nabla_X g\big(f(X)\big) = \nabla_X f^T \nabla_f g$$

Đạo hàm của hàm nhiều biến

- **Bảng các đạo hàm thường gặp:**

$f(\mathbf{x})$	$\nabla f(\mathbf{x})$
$\mathbf{a}^T\mathbf{x}$	$\mathbf{a}$
$\mathbf{x}^T\mathbf{A}\mathbf{x}$	$\mathbf{A}+\mathbf{A}^T)\mathbf{x}$
$\mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2$	$2\mathbf{x}$
$\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{x}-\mathbf{b})$
$\mathbf{a}^T\mathbf{x}^T\mathbf{x}\mathbf{b}$	$2\mathbf{a}^T\mathbf{b}\mathbf{x}$
$\mathbf{a}^T\mathbf{x}\mathbf{x}^T\mathbf{b}$	$(\mathbf{a}\mathbf{b}^T+\mathbf{b}\mathbf{a}^T)\mathbf{x}$

Đạo hàm theo vector

$f(\mathbf{X})$	$\nabla f(\mathbf{X})$
$\|\mathbf{X}\|_F^2$	$2\mathbf{X}$
$\mathbf{A}\mathbf{X}$	$\mathbf{A}^T$
$\|\mathbf{A}\mathbf{X}-\mathbf{B}\|_F^2$	$2\mathbf{A}^T(\mathbf{A}\mathbf{X}-\mathbf{B})$
$\|\mathbf{X}\mathbf{A}-\mathbf{B}\|_F^2$	$2(\mathbf{X}\mathbf{A}-\mathbf{B})\mathbf{A}^T$
$\mathbf{a}^T\mathbf{X}^T\mathbf{X}\mathbf{b}$	$\mathbf{X}(\mathbf{a}\mathbf{b}^T+\mathbf{b}\mathbf{a}^T)$
$\mathbf{a}^T\mathbf{X}\mathbf{X}^T\mathbf{b}$	$(\mathbf{a}\mathbf{b}^T+\mathbf{b}\mathbf{a}^T)\mathbf{X},$
$\mathbf{a}^T\mathbf{Y}\mathbf{X}^T\mathbf{b}$	$\mathbf{b}\mathbf{a}^T\mathbf{Y}$
$\mathbf{a}^T\mathbf{Y}^T\mathbf{X}\mathbf{b}$	$\mathbf{Y}\mathbf{a}\mathbf{b}^T$
$\mathbf{a}^T\mathbf{X}\mathbf{Y}^T\mathbf{b}$	$\mathbf{a}\mathbf{b}^T\mathbf{Y}$
$\mathbf{a}^T\mathbf{X}^T\mathbf{Y}\mathbf{b}$	$\mathbf{Y}\mathbf{b}\mathbf{a}^T$

Đạo hàm theo ma trận

Khai triển Taylor

- **Khai triển Taylor cho hàm một biến:**

Taylor's theorem : Gọi $f : R \to R$ là một hàm liên tục có đạo hàm tới bậc k tại x với k là một số nguyên và ϵ là một số thực đủ nhỏ hơn 1. Ta có:

$$f(x+\epsilon) = f(x) + f'(x)\epsilon + \frac{f''(x)}{2!}\epsilon^2 + \ldots + \frac{f^{(k)}(x)}{k!}\epsilon^k + o(\epsilon^k) \qquad (8)$$

Ở đây, $f^{(k)}(x)$ là đạo hàm cấp k của f tại x.

Khai triển Taylor

- **Khai triển Taylor cho hàm nhiều biến:**

Taylor's theorem : Gọi $f : \mathbb{R}^d \to R$ là một hàm liên tục có đạo hàm tới cấp 2 tại $\mathbf{x}$ và ϵ là một số thực đủ nhỏ hơn 1. Với mọi vector cố định cho trước $\mathbf{y}$, ta có:

$$\begin{aligned} f(\mathbf{x}+\epsilon\mathbf{y}) &= f(\mathbf{x}) + \epsilon\langle\nabla f(\mathbf{x}), \mathbf{y}\rangle + o(\epsilon) \\ f(\mathbf{x}+\epsilon\mathbf{y}) &= f(\mathbf{x}) + \epsilon\langle\nabla f(\mathbf{x}), \mathbf{y}\rangle + \epsilon^2 \tfrac{1}{2}\mathbf{y}^T\nabla^2 f(\mathbf{x})\mathbf{y} + o(\epsilon^2) \end{aligned} \tag{14}$$

→ Khai triển Taylor là cơ sở lý thuyết cho rất nhiều thuật toán tối ưu bằng cách xấp xỉ, trong đó điển hình là Gradient descent và Newton step

Phần 3
Xác suất cơ bản

Sự kiện và xác suất

- **Định nghĩa 1:** Một không gian xác suất bao gồm 3 thành phần:
  - Một không gian mẫu Q: là một tập các kết quả có thể của một quá trình ngẫu nhiên được mô hình hoá bởi không gian xác suất đó.
  - Sự kiện: mỗi sự kiện có thể được coi là một tập con của Q. Tập các sự kiện được kí hiệu là F.
  - Một hàm xác suất: Pr: F → R thoả mãn những điều kiện sau:
    - Với mỗi sự kiện E: $0 \leq \Pr[E] \leq 1$
    - $\Pr[Q] = 1$
    - Với một tập hữu hạn hoặc đếm được các sự kiện $E_1, E_2, \ldots,$ đôi một không giao nhau: $\Pr[\cup_{i \geq 1} E_i] = \sum_{i \geq 1} \Pr[E_i]$

Sự kiện và xác suất

- **Bổ đề 1:** Cho hai sự kiện E_1, E_2 bất kỳ:

$$\Pr[E_1 \cup E_2] = \Pr[E_1] + \Pr[E_2] - \Pr[E_1 \cap E_2]$$

- **Bổ đề 2:** Cho một tập hữu hạn hoặc đếm được các sự kiện E_1, E_2 bất kỳ:

$$\Pr[\cup_{i\geq 1} E_i] \leq \sum_{i\geq 1} \Pr[E_i]$$

- **Bổ đề 3:** Nguyễn lý bù trừ

Cho một tập n sự kiện $E_1, E_2, \ldots, E_n$ bất kỳ:

$$\begin{aligned} \Pr[\cup_{i\geq 1} E_i] &= \sum_{i\geq 1} \Pr[E_i] - \sum_{i<j} \Pr[E_i \cap E_j] \\ &+ \sum_{i<j<k} \Pr[E_1 \cap E_j \cap E_k] \\ &- \ldots + (-1)^{\ell+1} \sum_{i_1<i_2<\ldots<i_\ell} \Pr[\cap_{r=1}^{\ell} E_{i_r}] \end{aligned}$$

Sự kiện và xác suất

- **Định nghĩa 2:** Hai sự kiện E_1, E_2 được gọi là độc lập nếu:

$$\Pr[E_1 \cap E_2] = \Pr[E_1] \,.\, \Pr[E_2]$$

Tương tự như vậy, các sự kiện $E_1, E_2, \dots, E_n$ được gọi là độc lập nếu: $\Pr[\bigcap_{i=1}^{n} E_i] = \prod_{i=1}^{n} \Pr[E_i]$

- **Định nghĩa 3:** Xác suất có điều kiện của một sự kiện E khi biết sự kiện F xảy ra là:

$$\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}$$

Sự kiện và xác suất

Một luật rất quan trọng để tính xác suất là luật tổng xác suất:

- **Định lý 1** (**Law of total probability**): Gọi $E_1, E_2, \dots, E_n$ là các sự kiện đôi một không giao nhau trong một không gian mẫu Q thoả mãn $\bigcup_{i=1}^{n} E_i = Q$, ta có:

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B \cap E_i] = \sum_{i=1}^{n} \Pr[B|E_i] \Pr[E_i]$$

- **Định lý 2** (**Bayes' Law**): Gọi $E_1, E_2, \dots, E_n$ là các sự kiện đôi một không giao nhau trong một không gian mẫu Q thoả mãn $\bigcup_{i=1}^{n} E_i = Q$, ta có:

$$\Pr[E_j|B] = \frac{\Pr[E_j \cap B]}{\Pr[B]} = \frac{\Pr[B|E_j] \Pr[E_j]}{\sum_{i=1}^{n} \Pr[B|E_i] \Pr[E_i]}$$

Biến ngẫu nhiên

- **Định nghĩa 4:** Biến ngẫu nhiên (đại lượng ngẫu nhiên) là một đại lượng mà giá trị của nó là ngẫu nhiên, phụ thuộc vào kết quả phép thử.
  - Biến ngẫu nhiên được gọi là ***rời rạc***, nếu tập giá trị của nó là một tập hữu hạn hoặc vô hạn đếm được các phần tử
  - Biến ngẫu nhiên được gọi là ***liên tục***, nếu tập giá trị của nó lấp kín một khoảng hoặc một số khoảng của trục số hoặc cũng có thể là cả trục số.

- **Định nghĩa 5:** Hàm phân phối xác suất của biến ngẫu nhiên X, kí hiệu là F(x) và được xác định như sau:

$$F(x) = P(X < x)$$

Biến ngẫu nhiên

- **Định nghĩa 6:** Hàm mật độ xác suất f(x) của biến ngẫu nhiên liên tục X thể hiện mức độ tập trung xác suất của X xung quanh điểm x.

Tính chất:

- $f(x) \geq 0 \; \forall x$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$
- $P(a \leq X \leq b) = \int_{a}^{b} f(x)dx$ (có thể bỏ các dấu "=")
- Hàm phân phối xác suất:

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)dt$$

Từ đó suy ra: $f(x) = F'(x)$

Các tham số đặc trưng

- **Kỳ vọng:**
  - Là đại lượng đặc trưng có giá trị trung bình của một biến ngẫu nhiên, ***kí hiệu là E(X) hoặc EX.***
  - Tính chất:
    - E(c)=c với c là hằng số
    - E(aX)=aEX với a là hằng số
    - E(X+Y)=EX+EY với X, Y là hai biến ngẫu nhiên bất kỳ
    - E(XY)=EX.EY nếu X, Y là hai biến ngẫu nhiên độc lập

Các tham số đặc trưng

Phương sai:

- Là đại lượng đặc trưng cho trung bình của bình phương sai số, phản ánh mức độ phân tán của các giá trị của biến ngẫu nhiên xung quanh giá trị trung bình của nó là kỳ vọng, ***ký hiệu là V(X) hoặc VX***
- Công thức tính:

$$VX = E(X - EX)^2 = E(X^2) - (EX)^2$$

- Tính chất:
  - Vc=0 với c là hằng số
  - V(aX)=a^2VX với a là hằng số
  - V(X+b)=VX
  - V(X+Y)=VX+VY nếu X, Y là hai biến ngẫu nhiên độc lập

Các tham số đặc trưng

❑Hiệp phương sai

Giả sử X, Y là các biến ngẫu nhiên, hiệp phương sai của X và Y được ký hiệu là μ_{XY}, và được xác định bởi:

$$\mu_{XY} = E[(X - EX)(Y - EY)] = E(XY) - EX.EY$$

Trong đó $E(XY)$ được xác định theo công thức:

$$E(XY) = \begin{cases} \sum_i \sum_j x_i y_j p_{ij}, & \text{đối với biến ngẫu nhiên rời rạc} \\ \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} xy.f(x,y), & \text{đối với biến ngẫu nhiên liên tục} \end{cases}$$

Các tham số đặc trưng

❑Hệ số tương quan

Giả sử X, Y là các biến ngẫu nhiên, hệ số tương quan của X và Y được ký hiệu là ρ_{XY}, xác định bởi:

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y} = \frac{\mu_{XY}}{\sqrt{VX.VY}}$$

- Có thể chứng minh được $|\rho_{XY}| \leq 1$
- Nếu $\rho_{XY} = \pm 1$ ta nói X và Y có tương quan tuyến tính
- Nếu $\rho_{XY} = 0$ ta nói X và Y là không tương quan

Các tham số đặc trưng

Ta có các tham số đặc trưng cho bộ dữ liệu gồm N điểm $x_1, x_2, \dots, x_N$, như sau:

- Vecto kỳ vọng: $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$
- Ma trận hiệp phương sai: $S = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})(x_n - \bar{x})^T = \frac{1}{N}\hat{X}\hat{X}^T$

Trong đó $\hat{X}$ được tạo bằng cách trừ mỗi cột của

$X = [x_1, x_2, \dots, x_N]$ đi $\bar{x}$

- Mọi phần tử trên đường chéo của ma trận S là phương sai của từng chiều dữ liệu
- Các phần tử nằm ngoài đường chéo thể hiện sự tương quan giữa các thành phần của dữ liệu, chính là hiệp phương sai

Một số phân phối xác suất thường gặp

❑ Phân phối Bernoulli:

Phân phối Bernoulli là một phân phối rời rạc mô tả các biến ngẫu nhiên nhị phân: trường hợp đầu ra chỉ nhận một trong hai giá trị 0, 1.

Phân phối Bernoulli được mô tả bằng một tham số $\lambda \in [0,1]$ và là xác suất để bnn x=1:

$$p(x = 1) = \lambda, p(x = 0) = 1 - \lambda$$

→ $p(x) = \lambda^x (1 - \lambda)^{1-x}$

Một số phân phối xác suất thường gặp

- **Phân phối Categorical:**

Trong nhiều trường hợp, đầu ra của bnn rời rạc có thể là K đầu ra, phân phối Categorical sẽ được mô tả bởi K tham số, viết dưới dạng vecto: $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_k]$ với λ_k là các số không âm và có tổng bằng 1

$$p(x = k) = \lambda_k$$

Một số phân phối xác suất thường gặp

❑ Phân phối Chuẩn:

- Tổng quát với biến ngẫu nhiên D chiều. Có hai tham số mô tả phân phối này là: vecto kỳ vọng $\mu \in R^D$ và ma trận hiệp phương sai $\Sigma \in S^D$ là một ma trận đối xứng xác định dương:

- Hàm mật độ xác suất có dạng:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Một số phân phối xác suất thường gặp

❑ Phân phối Beta:

- Phân phối Beta là một phần phối liên tục được định nghĩa trên một biến ngẫu nhiên $\lambda \in [0,1]$, được dung để mô tả sự biến động của tham số λ trong phân phối Bernoulli.
- Phân phối Beta được mô tả bởi hai tham số dương: α, β.
- Hàm mật độ xác suất là:

$$p(\lambda) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\lambda^{\alpha-1}(1-\lambda)^{\beta-1}$$

Với hàm số Gama: $$\Gamma(z) = \int_0^\infty t^{z-1}\exp(-t)dt$$

Một số phân phối xác suất thường gặp

❑ Phân phối Dirichlet:

- Phân phối Dirichlet là trường hợp tổng quát của phân phối Beta khi được dung để mô tả tham số của phần phối Categorical.
- Phân phối Dirichlet được định nghĩa trên K biến liên tục $\lambda_1, \lambda_2, \dots, \lambda_k$ với λ_k là các số không âm và có tổng bằng 1.
- Có K tham số dương để mô tả phân phối Dirichlet là: $\alpha_1, \alpha_2, \dots, \alpha_k$
- Hàm mật độ xác suất có dạng:

$$p(\lambda_1, \dots, \lambda_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}$$

5 công cụ của xác suất

❑ Tính tuyến tính của kỳ vọng (1)

Gọi $X_1, X_2, \ldots, X_n$ là n biến ngẫu nhiên trong cùng một không gian xác suất. Gọi $X = \sum_{i=1}^{n} X_i$, ta có:

$$E(X) = \sum_{i=1}^{n} EX_i$$

❑ Union bound (2)

Union Bound: Gọi $E_1, E_2, \ldots E_n$ là tập n sự kiện, ta có:

$$\Pr[\text{ ít nhất một sự kiện } E_i \text{ xảy ra}] \leq \sum_{i=1}^{n} \Pr[E_i] \qquad (16)$$

5 công cụ của xác suất

❑ Bất đẳng thức Markov (3)

Cho một biến ngẫu nhiên X không âm, với mọi hằng số $c > 1$, ta có:

$$\Pr[X > cE(X)] \leq \frac{1}{c}$$

Ví dụ: trong bài giải thuật Quicksort ngẫu nhiên, ta tính được kỳ vọng thời gian của Quiksort khi chọn pivot ngẫu nhiên là $E[T(n)] = O(nlogn)$. Theo bất đẳng thức Markov, xác suất để giải thuật này chạy lâu hơn 10 lần $E[T(n)]$ là không quá 10%

5 công cụ của xác suất

❑ Bất đẳng thức Chebyshev (4)

Cho một biến ngẫu nhiên X, với mọi hằng số $c > 1$, ta có:

$$\Pr[|X - EX| \geq c.\sigma(X)] \leq \frac{1}{c^2}$$

❑ Bất đẳng thức Chernoff (5)

Chernoff Bounds: Giả sử $X = \sum_{i=1}^{n} X_i$ trong đó $X_1, X_2, \ldots, X_n$ là n biến ngẫu nhiên độc lập có giá trị trong đoạn $[0, 1]$. Gọi $\mu = E[X]$. Ta có:

- Với mọi $\delta > 0$,

$$\Pr[X \geq (1+\delta)\mu] \leq \left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \qquad (10)$$

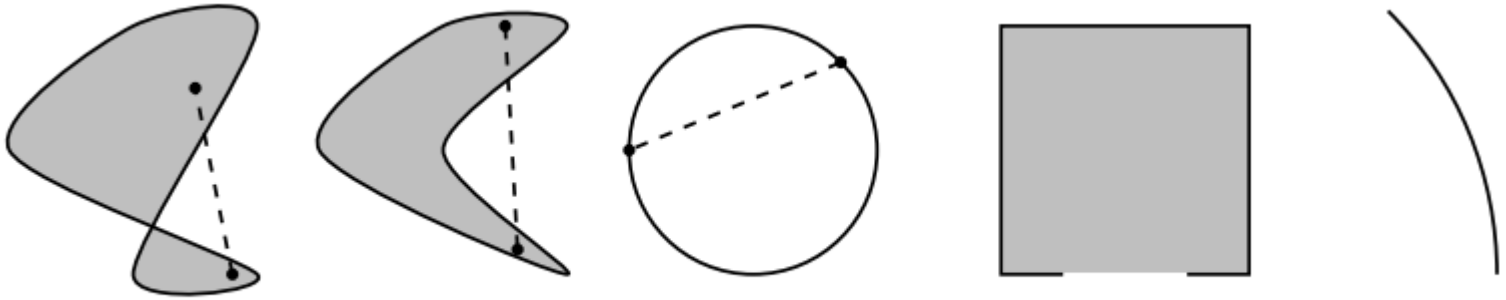
- Với mọi $\delta \in (0, 1)$,

$$\Pr[X \geq (1-\delta)\mu] \leq e^{-\delta^2\mu/2} \qquad (11)$$

Phần 4
Một số vấn đề về tối ưu hoá

Convex set

- **Định nghĩa:** Một tập hợp C được gọi là một tập lồi nếu với hai điểm $x_1, x_2 \in C$ thì điểm $x_\theta = \theta x_1 + (1-\theta)x_2$ cũng nằm trong C với mọi $\theta \in [0,1]$



Examples of nonconvex sets

Convex set

❑ Một số ví dụ về tập lồi:

- Siêu phẳng (Hyerplane): $a_1x_1 + a_2x_2 + \cdots + a_nx_n = a^Tx = b$
- Nửa không gian (Halfspace):

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = a^Tx \leq b$$

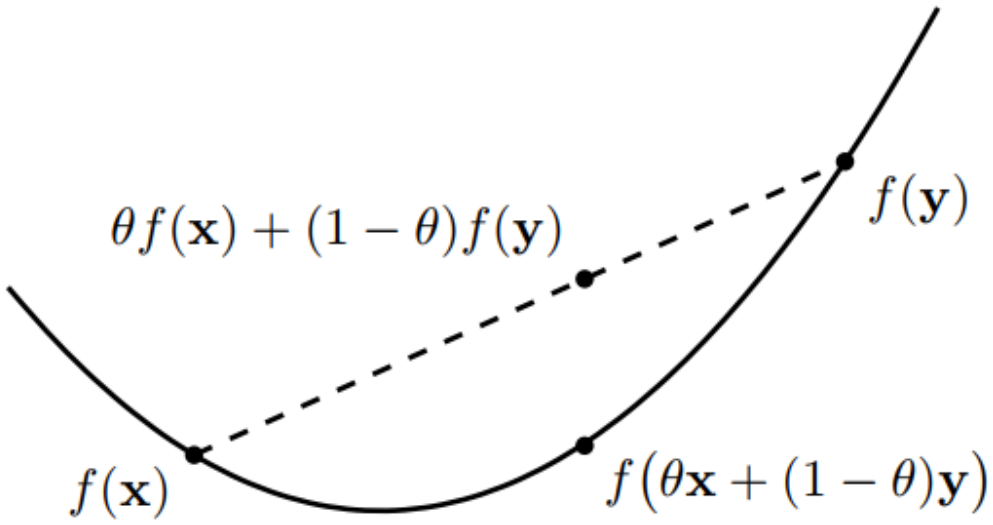
- Norm ball: $B(x_c, r) = \{x | \|x - x_c\|_2 \leq r\} = \{x_c + ru | \|u\|_2 \leq 1\}$

Convex function

- **Định nghĩa:** Một hàm số $f: R^n \to R$ được gọi là một hàm lồi nếu $domf$ là một hàm lồi và:

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$

Với mọi $x, y \in domf$ và $0 \leq \theta \leq 1$.



$$\theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y}) \geq f(\theta \mathbf{x} + (1-\theta)\mathbf{y})$$

Convex function

Các tính chất cơ bản:

- Nếu f(x) là convex thì af(x) là convex với a>0, và là concave nếu a<0
- Tổng của hai hàm lồi là một hàm lồi với tập xác định là giao của hai tập xác định của hai hàm đã cho
- **Pointwise maximum và supremum:** Nếu các hàm số $f_1, f_2, \dots, f_m$ là convex thì

$$f(x) = max\{f_1(x), f_2(x), \dots, f_m(x)\}$$

cũng là hàm lồi trên tập xác định là giao của các tập xác định của các hàm số trên

- Mọi hàm số bất kỳ thoả mãn 3 điều kiện của norm đều là convex

Convex function

❑Kiểm tra tính convex

- **Cách 1**: sử dụng First-order condition

Giả sử hàm số f(x) có tập xác định $domf$ convex và f(x) khả vi trên $domf$. Khi đó f(x) là convex iff:

$$f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0)\ \forall x, x_0 \in domf$$

- **Cách 2:** sử dụng Second-order condition

Một hàm số có đạo hàm bậc hai là convex nếu domf là convex và Hessian của nó là một ma trận ***bán xác định dương*** với mọi $x \in domf$

$$\nabla^2 f(x) \succcurlyeq 0$$

Convex optimization problem

❑ **Định nghĩa:**

Một bài toán tối ưu lồi là một bài toán tối ưu có dạng:

$$x^* = argmin_x f_0(x)$$

thoả mãn:

$$f_i(x) \leq 0 \; i = 1,2, \dots, m \text{ và}$$
$$h_j(x) = a_j^T x - b_j = 0, \mathrm{j} = 1, \dots$$

trong đó $f_0, f_1, \dots, f_m$ là các hàm lồi.

Convex optimization problem

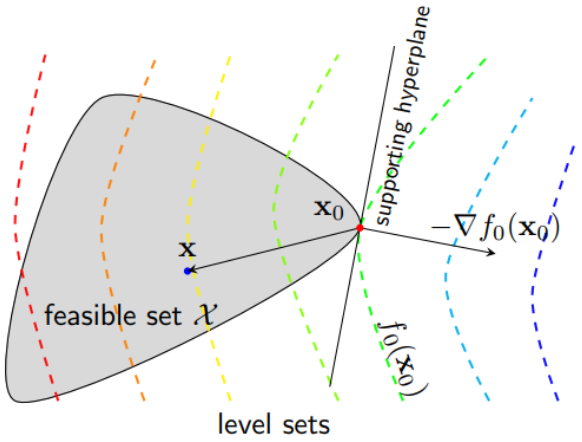
❑ Tính chất:

- Với bài toán tối ưu lồi, local optimum cũng chính là global optimum của nó
- Nếu f_0 là hàm khả vi, theo first-order condition:
  $f_0(x) \geq f_0(x_0) + \nabla f_0(x_0)^T(x - x_0) \;\; \forall x, x_0 \in dom f_0$

Đặt X là tập các điểm thoả mãn các điều kiện của bài toán.

Điều kiện cần và đủ để một điểm $x_0 \in X$ là optimal point là:

$\nabla f_0(x_0)^T(x - x_0) \geq 0 \; \forall x, x_0 \in X$



Convex optimization problem

- **Tính chất:**
  - Với bài toán mà $f_0(x)$ hoặc tập các điều kiện có dạng phức tạp, thường không có các phương pháp chung hiệu quả để giải.
  - Một số phương pháp kinh điển để giải:
    - Phương pháp nhân tử Lagrange và bài toán đối ngẫu: ***sử dụng hiệu quả khi các hàm $f_i(x), i = 0, 1, ..$ ở một số dạng đặc biệt, và thường nghiệm của bài toán "closed form"***
    - Phương pháp xấp xỉ: ***sử dụng khi tập điều kiện thoả mãn K có dạng đơn giản, nghiệm của bài toán không tính trực tiếp được dưới các điều kiện tối ưu***

      - Thuật toán Gradient descent
      - Thuật toán Newton
      - Thuật toán Frank-Wolfe

Convex optimization problem

❑ Phương pháp nhân tử Lagrange

Để giải bài toán $x^* = argmin_x f_0(x)$

thoả mãn:

$$f_i(x) \leq 0 \; i = 1,2, \dots, m \text{ và}$$
$$h_j(x) = a_j^T x - b_j = 0, j = 1, \dots p$$

Xét hàm Lagrangian sau:

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} v_j h_j(x)$$

Đi giải bài toán tương đương: $\boldsymbol{\min_{x,\lambda,v}} \boldsymbol{L(x, \lambda, v)}$ với $\lambda_i \geq 0$

Convex optimization problem

❑ Phương pháp nhân tử Lagrange

Để giải bài toán $\min_{x,\lambda,v} L(x,\lambda,v)$ với $\lambda_i \geq 0$, ta giải hệ phương trình các đạo hàm riêng bằng 0:

$$\begin{cases} \dfrac{\partial L(x,\lambda,v)}{\partial x} = 0 \\ \dfrac{\partial L(x,\lambda,v)}{\partial \lambda} = 0 \\ \dfrac{\partial L(x,\lambda,v)}{\partial v} = 0 \end{cases}$$

Nghiệm của hệ phương trình này là các điểm stationary của bài toán tối ưu ban đầu

Convex optimization problem

❑ Phương pháp nhân tử Lagrange

Trong nhiều trường hợp, thay vì giải bài toán Lagrange gốc, chúng ta sẽ đi giải bài toán đối ngẫu của nó:

$$g(\lambda, v) = \inf_{x \in D} L(x, \lambda, v) = \inf_{x \in D} (f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} v_j h_j(x))$$

- Tính chất của bài toán đối ngẫu:
  - Với mọi (λ, v): $g(\lambda, v) \leq p^*$ với p^* là giá trị tối ưu của bài toán ban đầu.
  - $g(\lambda, v)$ luôn là convex

Convex optimization problem

❑ Thuật toán Gradient descent

Cho hàm lồi $f(x)$ với tập xác định lồi K, xét bài toàn tìm:

$$\min_{x \in K} f(x)$$

- Nếu $K = R^n$ ta có bài toán tối ưu không ràng buộc, và ta có thuật toán như sau:

GradientDescent(f, ϵ):
  pick a point $\mathbf{x}_0 \in \mathbb{R}^d$
  choose a large integer T
  for $t \leftarrow 0$ to $T-1$
    choose a positive step size η_t appropriately
    $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$
  return $\mathbf{x}_T$

  - η_t là tốc độ học, và thường là các con số dương, nhỏ.

Convex optimization problem

❑ Thuật toán Gradient descent

Cho hàm lồi $f(x)$ với tập xác định lồi K, xét bài toàn tìm:

$$\min_{x \in K} f(x)$$

- Nếu $K \neq R^n$ ta có bài toán tối ưu ràng buộc

ProjectedGradientDescent(f, ϵ):

pick a point $\mathbf{x}_0 \in \mathbb{R}^d$
choose a large integer T
for $t \leftarrow 0$ to $T-1$
  choose a positive step size η_t appropriately
  $\mathbf{x}'_{t+1} \leftarrow \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$
  $\mathbf{x}_{t+1} \leftarrow \Pi_K(\mathbf{x}'_{t+1})$
return $\mathbf{x}_T$

- $\Pi_K(x)$ là phép chiếu điểm x vào tập K.

Convex optimization problem

❑ Thuật toán Frank-Wolfe

- Trong nhiều trường hợp, phép chiếu có thể tính toán trong thời gian đa thức. Tuy nhiên phần lớn các trường hợp thì việc tìm hình chiếu tương đương với một bài toán tối ưu bậc 2, chi phí tính toán rất tốn kém nếu bài toán đầu vào có số chiều lớn.
- Thuật toán Frank-Wolf thay phép chiếu bằng một bài toán tuyến tính → giảm độ phức tạp tính toán tại mỗi vòng lặp.

Algorithm Frank-Wolfe (1956)

Let $\boldsymbol{x}^{(0)} \in \mathcal{D}$

for $k = 0 \dots K$ **do**

Compute $\boldsymbol{s} := \arg\min_{\boldsymbol{s} \in \mathcal{D}} \left\langle \boldsymbol{s}, \nabla f(\boldsymbol{x}^{(k)}) \right\rangle$

Update $\boldsymbol{x}^{(k+1)} := (1-\gamma)\boldsymbol{x}^{(k)} + \gamma \boldsymbol{s}, \quad \text{for } \gamma := \frac{2}{k+2}$

end for

Tài liệu tham khảo

- Boyd, Stephen, and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.