



PONTIFICIA  
UNIVERSIDAD  
CATÓLICA  
DE CHILE

# IMT3850: Fundamentos Matemáticos para Inteligencia Artificial

## Clase 7

Profesor: Manuel A. Sánchez

Abril 2023

# Contenidos Clase 7:

Probabilidades:

- Matriz de Covarianza

# Matriz de Covarianza

## Definición

Sean  $X$  e  $Y$  dos variables aleatorias. Entonces, la covarianza se define

$$\sigma_{XY} = E[(X - E[X])(Y - E[Y])]$$

## Ejemplo

Sea las variables edad, altura, y peso  $(a, h, w)$  de una población. Sean  $(\mu_a, \mu_h, \mu_w)$  las medias y  $(\sigma_a, \sigma_h, \sigma_w)$  varianzas separadas. Entonces, la covarianza de las variables edad y altura es

$$\sigma_{a,h} = E[(a - \mu_a)(h - \mu_h)]$$

para esta definición necesitamos la probabilidad conjunta de cada par.

$p_{a,h}$  : probabilidad de edad  $a$  y altura  $h$  al mismo tiempo.

## Definición

Definimos la probabilidad conjunta del experimento 1,  $X$ , produce  $x_i$  y el experimento 2,  $Y$ , produce  $y_j$  por  $p_{i,j}$ . Así, la covarianza es

$$\sigma_{12} = E[(X - \mu_1)(Y - \mu_2)] = \sum_{i,j} p_{i,j}(x_i - \mu_1)(y_j - \mu_2)$$

## Ejemplo

Lanzamiento de dos monedas (c,c), (c,s), (s,c), (s,s), cada evento con probabilidad de  $1/4$ , son claramente independientes los lanzamientos, entonces

$$p_{i,j} = (\text{Probabilidad de } i) \times (\text{Probabilidad de } j)$$

Matriz de probabilidad conjunta y matriz de varianza ( $\sigma_{12} = \sigma_{21} = 0$ )

$$P = \begin{bmatrix} 1/4 & 1/4 \\ 1/4 & 1/4 \end{bmatrix}, \quad V = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

## Ejemplo

Lanzamiento de dos monedas pegadas, esto es, dan el mismo resultado (c,c), (c,s), (s,c), (s,s), con probabilidad de  $1/2, 0, 0, 1/2$ , respectivamente. Son claramente no independientes los lanzamientos. Matriz de probabilidad conjunta y matriz de covarianza

$$P = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}, \quad V = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

donde

$$\sigma_{12} = \sum_{i,j} p_{i,j} (x_i - \mu_1)(y_j - \mu_2) = \frac{1}{2} \left(1 - \frac{1}{2}\right) \left(1 - \frac{1}{2}\right) + \frac{1}{2} \left(0 - \frac{1}{2}\right) \left(0 - \frac{1}{2}\right)$$

**Definición** Dadas la muestra estadística de valores  $(X_1, \dots, X_N)$  la media muestral es una matriz semidefinida positiva. La media muestral y la varianza muestral están definidas por

$$\bar{X} = \frac{X_1 + \dots + X_N}{N}, \quad S = \frac{(X_1 - \bar{X})(X_1 - \bar{X})^\top + \dots + (X_N - \bar{X})(X_N - \bar{X})^\top}{N - 1}$$

**Observación:** La matriz de covarianza es semidefinida positiva

$$\begin{aligned} V &= \sum_{i,j} p_{i,j} \begin{bmatrix} (x_i - \mu_1)^2 & (x_i - \mu_1)(y_j - \mu_2) \\ (x_i - \mu_1)(y_j - \mu_2) & (y_j - \mu_2)^2 \end{bmatrix} \\ &= \sum_{i,j} p_{i,j} \begin{bmatrix} x_i - \mu_1 \\ y_j - \mu_2 \end{bmatrix} [x_i - \mu_1, y_j - \mu_2] \\ &= \sum_{i,j} p_{i,j} U U^\top \end{aligned}$$

## Observación

$$V = E[(X - \bar{X})(X - \bar{X})^\top]$$

$$E[c^\top X] = c^\top E[X]$$

$$\text{Var}(c^\top X) = E[(c^\top X - c^\top \bar{X})(c^\top X - c^\top \bar{X})^\top] = c^\top E[(X - \bar{X})(X - \bar{X})^\top]c = c^\top Vc$$

**Definición** Sean  $X$  e  $Y$  dos variables aleatorias con media  $\mu_X, \mu_Y$  y varianzas  $\sigma_X^2, \sigma_Y^2$ , respectively. Entonces, se define la correlación

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

$$-1 \leq \rho_{X,Y} \leq 1$$



# Gaussiana multivariada

**1 d:** La función de densidad de una distribución Normal (Gaussiana) con Media  $\mu$ , Varianza  $\sigma^2$  es  $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$ .  
Sabemos que

$$\int_{-\infty}^{\infty} p(x)dx = 1; \quad \int_{\mu-\sigma}^{\mu+\sigma} p(x)dx = \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-x^2/2} dx \approx \frac{2}{3}$$

## 2 d, independientes

Sean  $X$  e  $Y$  variables aleatorias independientes, dist. normal, con medias  $\mu_1, \mu_2$  y varianzas  $\sigma_1^2, \sigma_2^2$ , entonces la función de densidad

$$\begin{aligned} p(x, y) &= \frac{1}{\sqrt{2\pi}\sigma_1} e^{-(x-\mu_1)^2/(2\sigma_1^2)} \frac{1}{\sqrt{2\pi}\sigma_2} e^{-(y-\mu_2)^2/(2\sigma_2^2)} \\ &= \frac{1}{2\pi\sigma_1\sigma_2} e^{-(x-\mu_1)^2/(2\sigma_1^2) - (y-\mu_2)^2/(2\sigma_2^2)} \end{aligned}$$

## Observación:

$$-(x - \mu_1)^2 / (2\sigma_1^2) - (y - \mu_2)^2 / (2\sigma_2^2) = -\frac{1}{2} [x - \mu_1, y - \mu_2] \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} x - \mu_1 \\ y - \mu_2 \end{bmatrix}$$

## Gaussiana multivariada

La función de densidad de Gaussiana multivariada para  $x = (x_1, \dots, x_M)$ , con media  $\mu = (\mu_1, \dots, \mu_M)$  y matriz de covarianza  $V$  es

$$p(x) = \frac{1}{(2\pi)^{M/2} \sqrt{\det(V)}} e^{-\frac{1}{2}(x-\mu)^\top V^{-1}(x-\mu)}$$

## ■ Observación:

La matriz de covarianza es simétrica y semidefinida positiva, entonces

$$V = Q\Lambda Q^\top \implies V^{-1} = Q\Lambda^{-1}Q^\top$$

$$X = (x - \mu); \quad X^\top V^{-1} X = X^\top Q\Lambda^{-1}Q^\top X = Y^\top \Lambda^{-1} Y$$

La covarianza de las nuevas variables aleatorias  $Y$  es cero!

$$\begin{aligned} \int p(x) dx &= \frac{1}{(2\pi)^{M/2} \sqrt{\det(V)}} \int e^{-\frac{1}{2} Y^\top \Lambda^{-1} Y} dY \\ &= \frac{1}{(2\pi)^{M/2} \sqrt{\det(V)}} \left( \int e^{-y_1^2 / (2\lambda_1)} dy_1 \right) \cdots \left( \int e^{-y_M^2 / (2\lambda_M)} dy_M \right) \\ &= \frac{1}{(2\pi)^{M/2} \sqrt{\det(V)}} \sqrt{2\pi\lambda_1} \cdots \sqrt{2\pi\lambda_M} = 1 \end{aligned}$$

# Mínimos cuadrados ponderados

## Problema.

Minimizar  $\|b - Ax\|_2^2$ ,  $A \in \mathbb{R}^{m \times n}$ ,  $b \in \mathbb{R}^m$ .

Si los errores de medición en  $b$  son variables aleatorias independientes con media  $\mu = 0$ , con varianza  $\sigma^2 = 1$  y distribuidos normal. Entonces debemos minimizar  $\|b - Ax\|_2^2$ .

Asumimos que los errores de medición en  $b$  no son independientes o sus varianzas no son iguales entonces tenemos un problema de mínimos cuadrados ponderados WLS,

$$A^\top V^{-1} Ax = A^\top V^{-1} b$$

Ejemplos comunes ocurren con  $M$  errores independientes en  $b$ . Estos errores en  $b$  tienen varianzas  $\sigma_1, \dots, \sigma_M^2$ ,  $V$  es diagonal.

$$\text{WLS : } \text{minimizar } \sum_{i=1}^M \frac{(b - Ax)_i^2}{\sigma_i^2}$$

## Resumen:

- Resolver:  $Ax = b$ ,  $A \in \mathbb{R}^{M \times N}$ ,  $M > N$
- Cada  $b_i$  tiene media cero y varianza  $\sigma_i^2$  son independientes
- Dividimos la ecuación  $i$ -ésima por  $\sigma_i$  para tener varianza 1 para cada  $b_i/\sigma_i$
- Así, el problema queda  $V^{-1/2}Ax = V^{-1/2}b$

## La varianza de la solución de WLS

A veces la pregunta importante no es el mejor  $\hat{x}$  para un  $b$  en particular, este sólo es una muestra. El objetivo real es saber la confiabilidad del experimento. Esto lo medimos con la varianza del estimado  $\hat{x}$ . Primero, media cero en  $b$  nos da media cero en  $\hat{x}$ . La fórmula que conecta la varianza  $v$  de  $b$  con la varianza  $W$  de  $\hat{x}$  es

$$\text{Varianza-Covarianza para } \hat{x} \quad W = E[(\hat{x} - x)(\hat{x} - x)^\top] = (A^\top V^{-1} A)^{-1}$$

## En efecto

Si  $b$  tiene covarianza  $V$  y  $\hat{x} = Lb$  entonces  $\hat{x}$  tiene covarianza  $LV L^\top$ .

$$(A^\top V^{-1} A) \hat{x} = A^\top V^{-1} b \quad \rightarrow \quad L = (A^\top V^{-1} A)^{-1} A^\top V^{-1}$$

$$\text{y así } LV L^\top = (A^\top V^{-1} A)^{-1}$$

## Ejemplo.

Suponga que un doctor mide el corazón 3 veces,  $M = 3$ ,  $N = 1$ .

$$\begin{array}{rcl} x & = & b_1 \\ x & = & b_2 \\ x & = & b_3 \end{array} \quad \Longleftrightarrow \quad Ax = b; \quad A = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad V = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

Suponga  $\sigma_1^2 = 1/9$ ,  $\sigma_2^2 = 1/4$ ,  $\sigma_3^2 = 1$ . Entonces

$$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \hat{x} = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

Por lo tanto  $\hat{x} = (9b_1 + 4b_2 + b_3)/14$ ,  $W = 1/14$



# Curse of dimensionality

## Curse of Dimensionality

Un fenómeno que complica el trabajo de un probabilista en dimensiones altas ( es decir cuando tenemos un número elevado de variables aleatorias) es que la mayor parte de la probabilidad en la distribución conjunta está lejos de la región central del espacio de variables. Como consecuencia, observaciones muestreadas de la distribución en dimensión alta tienden a salir de la región central. Por lo tanto, se vuelve difícil aprender acerca de que está haciendo la distribución en la región central. Este fenómeno se conoce como la maldición de la dimensionalidad.

**Ejemplo.** Considere  $X_1, \dots, X_n$  variables aleatorias independientes distribuídas uniforme en  $[-1, 1]$ . Cual es la probabilidad que  $X = (X_1, \dots, X_n)$  esté en la esfera unitaria  $B_n = \{X : X_1^2 + \dots + X_n^2 \leq 1\}$

Densidad de la prob. conjunta:  $f(x_1, \dots, x_n) = \left(\frac{1}{2}\right)^n, \quad -1 \leq x_i \leq 1.$

$$\text{Así} \quad P(X \in B_n) = \int_{B_n} \frac{1}{2^n} dx = \frac{\text{Vol}(B_n)}{2^n} = \frac{\pi^{n/2}}{2^n \Gamma(n/2 + 1)}$$

# Contenidos Clase 7:

Optimizacion:

- Calculo Vectorial

# Función de una variable

Considere una función  $f : \mathbb{R} \rightarrow \mathbb{R}$ , entonces

- La pendiente de la recta secante que pasa por los  $(x, f(x))$  y  $(x + h, f(x + h))$  está dada por  $m = \frac{1}{h}(f(x + h) - f(x))$ ; la razón de cambio promedio.
- La pendiente de la recta tangente que pasa por el punto  $(x, f(x))$  está dada por  $f'(x) = \lim_{h \rightarrow 0} \frac{1}{h}(f(x + h) - f(x))$ ; la razón de cambio instantáneo.

## Ilustración

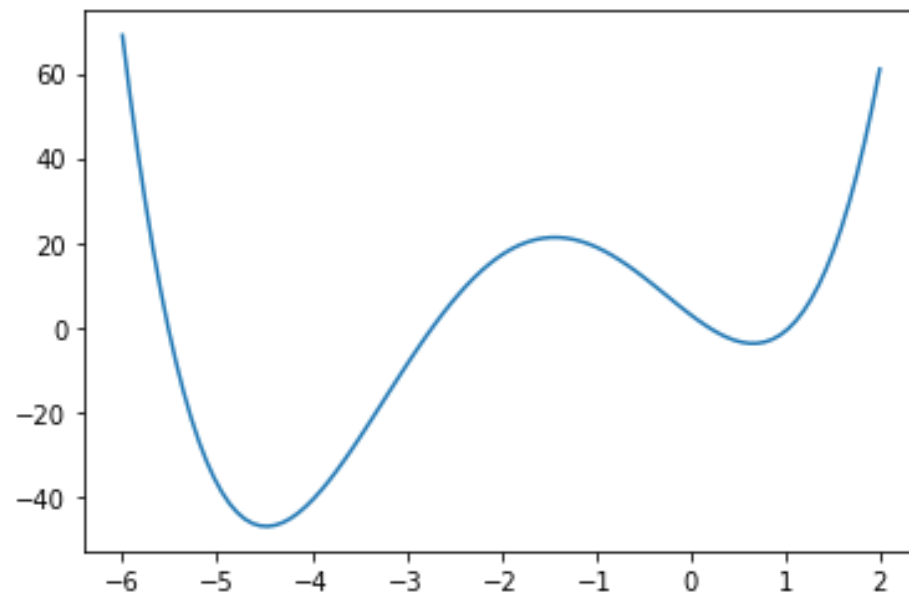
$$f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$$

$$f'(x) = 4x^3 + 21x^2 + 10x - 17$$

$$f''(x) = 12x^2 + 42x + 10$$

mínimo local/global

máximo local/global



**Definición** Una función  $f : \mathbb{R} \rightarrow \mathbb{R}$  es diferenciable en  $x = a$  si

$$f'(a) = \lim_{h \rightarrow 0} \frac{f(a+h) - f(a)}{h} \quad \text{existe.}$$

Esto significa que  $\lim_{h \rightarrow 0^-} \frac{f(a+h) - f(a)}{h} = \lim_{h \rightarrow 0^+} \frac{f(a+h) - f(a)}{h}$ .

**Ejemplo** Funciones no derivables

$$1. \quad f(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}, \text{ en } x = 0,$$

$$2. \quad f(x) = \frac{1}{x-1}, \text{ en } x = 1,$$

$$3. \quad f(x) = |x|, \text{ en } x = 0.$$

**Propiedades** Sean  $f, g$  funciones tales que sus derivadas existen.

$$1. (f \pm g)'(x) = f'(x) \pm g'(x)$$

$$2. (fg)'(x) = f'(x)g(x) + f(x)g'(x) \quad \text{regla del producto}$$

$$3. (f/g)'(x) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \quad \text{regla del cociente}$$

$$4. f(g(x))' = f'(g(x))g'(x) \quad \text{regla de la cadena}$$

## Propiedades

- $\frac{d}{dx}(c) = 0, c \in \mathbb{R}$
- $\frac{d}{dx}(\sin(x)) = \cos(x)$
- $\frac{d}{dx}(a^x) = a^x \ln(a), a > 0$
- $\frac{d}{dx}(x^n) = nx^{n-1}$
- $\frac{d}{dx}(\cos(x)) = -\sin(x)$
- $\frac{d}{dx}(\log_a(x)) = \frac{1}{x \ln(a)}, a > 0$

## Polinomios de Taylor

$$T_n(x) = \sum_{k=0}^n \frac{f^{(k)}(a)}{k!} (x-a)^k; \quad \text{para } f \text{ una función con } n\text{-ésimas derivadas}$$

$\Rightarrow$  Aproximación de  $f$  cerca de  $a$ .

## Ejemplos

$$\cos(x) \approx T_3(x) = \cos(0) + (\cos'(0))x + (\cos''(0))\frac{x^2}{2} + (\cos^{(3)}(0))\frac{x^3}{6}$$

$$e^x \approx T_n(x) = \sum_{k=0}^n \frac{(e^x)^{(k)}(0)}{k!} x^k = \sum_{k=0}^n \frac{x^k}{k!}$$

## Función multivariable

Considere una función  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $f(x) = f(x_1, \dots, x_n)$ .  
Definimos el gradiente de  $f$  como el vector

$$\nabla f(x) = \left[ \frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right]^\top \in \mathbb{R}^n$$

$$\frac{\partial f}{\partial x_i} = \lim_{h \rightarrow 0} \frac{1}{h} (f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)); \quad i = 1, \dots, n$$

Definimos la matrix Hessiana  $H \in \mathbb{R}^{n \times n}$  por

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}, \quad i, j = 1, \dots, n.$$



# Función multivariable

## Ejemplos

Calcule el gradiente de  $f(x, y, z) = (x + e^y)^2 + \ln(yz)$

$$\frac{\partial f}{\partial x} = 2(x + e^y); \quad \frac{\partial f}{\partial y} = 2(x + e^y)e^y + \frac{z}{yz}; \quad \frac{\partial f}{\partial z} = \frac{y}{yz}.$$

$$\nabla f(x, y, z) = \left[ 2(x + e^y), 2(x + e^y)e^y + \frac{z}{yz}, \frac{y}{yz} \right]^\top$$

## Observación:

$$\frac{\partial f}{\partial x_i} = \nabla f^\top e_i; \quad e_i = (0, \dots, 1, \dots, 0)$$

Podemos interpretar la derivada parcial como la razón de cambio en la dirección  $e_i$ .

## Derivada direccional

Sea  $v \in \mathbb{R}^n$ , entonces la derivada direccional de  $f$  en la dirección de  $v$  es  $\nabla f(x)^\top v$ . La interpretamos como la razón de cambio de  $f$  en la dirección de  $v$ .

¿En que dirección la función tiene la mayor razón de cambio?

→ si  $v$  apunta en la misma dirección que  $\nabla f(x)$

## Función multivariable multivariada

Sea  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  definida por  $f(x) = [f_1(x), f_2(x), \dots, f_m(x)]^\top \in \mathbb{R}^m$ ,  $x \in \mathbb{R}^n$ .

Entonces  $\frac{\partial f}{\partial x_i} = \begin{bmatrix} \partial f_1 / \partial x_i \\ \vdots \\ \partial f_m / \partial x_i \end{bmatrix}$ , y la matriz Jacobiano  $J \in \mathbb{R}^{m \times n}$

se define

$$J_{ij} = \frac{\partial f_i}{\partial x_j}; \quad J(x) = \begin{bmatrix} \partial f_1 / \partial x_1 & \dots & \partial f_1 / \partial x_n \\ \vdots & \ddots & \vdots \\ \partial f_m / \partial x_1 & \dots & \partial f_m / \partial x_n \end{bmatrix} = \begin{bmatrix} \nabla f_1(x)^\top \\ \vdots \\ \nabla f_m(x)^\top \end{bmatrix}$$

## Ejemplos

$f(x) = Ax$ ,  $A \in \mathbb{R}^{m \times n}$ . Entonces  $J(x) = A$ .

## Ejemplos

$$f : \mathbb{R}^2 \rightarrow \mathbb{R}, \text{ y } g : \mathbb{R} \rightarrow \mathbb{R}^2.$$


$$f(x, y) = e^{xy^2}; \quad g(t) = \begin{bmatrix} t \cos(t) \\ t \sin(t) \end{bmatrix} = \begin{bmatrix} g_1(t) \\ g_2(t) \end{bmatrix}$$

$$h : \mathbb{R} \rightarrow \mathbb{R}, \quad h(t) := (f \circ g)(t)$$

$$\frac{dh}{dt} = J_f(g(t)) \cdot J_g(t) = \left[ \frac{\partial f}{\partial x}(g(t)), \frac{\partial f}{\partial y}(g(t)) \right] \cdot \begin{bmatrix} \partial g_1 / \partial t \\ \partial g_2 / \partial t \end{bmatrix}$$

$$\frac{\partial f}{\partial x} = e^{xy^2}; \quad \frac{\partial f}{\partial y} = 2xye^{xy^2}; \quad J_f(g(t)) = [1, 2t^2 \cos(t) \sin(t)] e^{t^3 \cos(t) \sin(t)}$$

$$\frac{\partial g_1}{\partial t} = \cos(t) - t \sin(t); \quad \frac{\partial g_2}{\partial t} = \sin(t) + t \cos(t);$$


$$\frac{dh}{dt} = e^{t^3 \cos(t) \sin(t)} (\cos(t) - t \sin(t) + 2t^2 \cos(t) \sin(t)(\sin(t) + t \cos(t)))$$

## Ejemplos

Gradiente de la función de pérdida de mínimos cuadrados  
Modelo Lineal

$$y = \Phi\theta; \quad \theta \in \mathbb{R}^D, \quad \Phi \in \mathbb{R}^{N \times D}$$

Error  $L(e) = \|e\|^2$ ;  $e = y - \Phi\theta$ ;  $y$  observaciones. Queremos encontrar

$$\frac{\partial L}{\partial \theta} \in \mathbb{R}^{1 \times D}, \quad \frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial e} \frac{\partial e}{\partial \theta}$$
$$\frac{\partial L}{\partial e} = 2e^\top \in \mathbb{R}^{1 \times N} \quad \text{y} \quad \frac{\partial e}{\partial \theta} = -\Phi \in \mathbb{R}^{N \times D}$$

Entonces,

$$\frac{\partial L}{\partial \theta} = -2e^\top \Phi = -2(y - \Phi\theta)^\top \Phi$$

## Argmin

Cuando buscamos minimizar una función  $f(x)$  definimos  
 $\operatorname{argmin} f(x) =$  valor(es) de  $f(x)$  donde alcanza su mínimo.

## Aproximacion usando Taylor

$$f(x + \Delta x) \approx f(x) + (\Delta x)^\top \nabla f + \frac{1}{2}(\Delta x)^\top H(\Delta x)$$

# Convexidad

## Definición

Un conjunto  $C$  es convexo si para cada  $x, y \in C$  y para cada escalar  $\theta$  con  $0 \leq \theta \leq 1$ , tenemos que

$$\theta x + (1 - \theta)y \in C$$

## Definición

Sea  $f : \mathbb{R}^D \rightarrow \mathbb{R}$  una función cuyo dominio es un conjunto convexo. La función  $f$  es una función convexa si para todo  $x, y$  en el dominio de  $f$ , y para cada escalar  $\theta$  con  $0 \leq \theta \leq 1$ , tenemos que

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

## Observación

Si una función  $f : \mathbb{R}^D \rightarrow \mathbb{R}$  es diferenciable entonces podemos decir que es convexa si y sólo si para los puntos  $x, y$  en el dominio de  $f$  se tiene que.

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) \quad \text{desigualdad de Jensen}$$



## Observación

Si tenemos que  $f$  es 2 veces diferenciable, entonces el Hessiano existe. Así la función es convexa si y sólo si la matriz Hessiana es semidefinida positiva.

$$x^\top Ax \geq 0, \quad \forall x \in \mathbb{R}^d$$

## Proposición

Si  $f : \mathbb{R}^D \rightarrow \mathbb{R}$  es una función convexa, entonces todo mínimo local es un mínimo global.

## Lema

Si  $f : \mathbb{R}^D \rightarrow \mathbb{R}$  está definida por

$$f(x) = g(y^\top x + b), \quad y \in \mathbb{R}^D, b \in \mathbb{R},$$

donde  $g : \mathbb{R} \rightarrow \mathbb{R}$  es una función convexa. Entonces  $f$  es convexa.

## Ejemplos

- Sea  $f(x) = (y^\top x - b)^2$ , es una función convexa?
- Sea  $f(x) = \ln(1 + e^{-y^\top x + b})$ , es una función convexa?

## Lema

Si las funciones  $f_i : \mathbb{R}^D \rightarrow \mathbb{R}$ ,  $1 \leq i \leq n$ , son convexas.  
Entonces,

$$\max_{1 \leq i \leq n} f_i(x) \quad \text{y} \quad \sum_{i=1}^n a_i f_i(x), \quad (a_i \geq 0)$$

son también funciones convexas.

## Ejemplos

- Sea  $f(x) = |x|$ , es una función convexa?

**Proposición** Sea  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Entonces

- $f$  es convexa si y sólo si su matriz Hessiana  $H(x)$  es semidefinida positiva para todo  $x$ .
- La matriz Hessiana es simétrica.
- La función  $f$  es estrictamente convexa si  $H(x)$  es definida positiva para todo  $x$ .

**Ejemplos** Sea  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Entonces

- la función lineal  $f(x) = c^\top x$  es convexa pero no estrictamente convexa.
- La función cuadrática  $f(x) = \frac{1}{2}x^\top Sx$ , con  $S$  simétrica y definida positiva es estrictamente convexa.

## Normas

Las normas como función  $f(x) = \|x\|$  son funciones convexas de  $x$ . Además la bola unitaria  $\|x\| \leq 1$  es un conjunto convexo de vectores  $x$ , esto se observa por la propiedad de desigualdad triangular

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

Veamos tres ejemplos de normas para  $x = (x_1, x_2) \in \mathbb{R}^2$

- $\|x\|_1 = |x_1| + |x_2|$
- $\|x\|_2 = \sqrt{x_1^2 + x_2^2}$
- $\|x\|_\infty = \max\{|x_1|, |x_2|\}$

# Método de Newton

## Problema

Buscamos minimizar una función  $F(x)$ , esto es buscamos  $x^* \operatorname{argmin} F(x)$  tal que  $\nabla F(x^*) = 0$ .

Supongamos que estamos en un punto cercano  $x_k$  y queremos movernos a un nuevo punto  $x_{k+1}$  mas cerca de  $x^*$ . Como escogemos  $x_{k+1}$ ?

Cerca del punto  $x_k$  el gradiente  $\nabla F$  es bien aproximado por su primera derivada, la matriz Hessiana, esto es

$$\nabla F(x_{k+1}) \approx \nabla F(x_k) + H(x_k)(x_{k+1} - x_k)$$

Como queremos que el lado izquierdo se vaya a cero, escogemos  $x_{k+1}$  satisfaga la ecuación, esto es

$$\text{Método de Newton: } H(x_k)\Delta x_k = H(x_k)(x_{k+1} - x_k) = -\nabla F(x_k)$$

$$x_{k+1} = \operatorname{argmin} T_2(x); \quad T_2(x) = F(x_k) + \nabla F(x_k)^\top (x - x_k) + \frac{1}{2} (x - x_k)^\top H(x_k) (x - x_k)$$

$$\text{Convergencia cuadrática: } \|x_{k+1} - x^*\| \leq C \|x_k - x^*\|^2$$

**Ejemplo** Minimizar  $F(x) = \frac{1}{3}x^3 - 4x$ , con  $\nabla F(x) = x^2 - 4$ ,  $H(x) = 2x$

$$\text{Iteración } x_{k+1} = x_k - x_k/2 + 2/x_k$$

## Método de Newton. Que tan bien funciona en la practica?

Primero,  $x_0$  puede no estar cerca de  $x^*$ . Entonces no podemos confiar en que la derivada de  $x_0$  sea útil. Así calculamos el paso  $\Delta x_0 = x_1 - x_0$  y permitimos **backtracking**

Escoger  $\alpha < 1/2$  y  $\beta < 1$  y reducir el paso  $\Delta x$  por un factor  $\beta$  hasta que sabemos que el nuevo  $x_{k+1} = x_k + t\Delta x$  es tal que:

$$F(x_k + t\Delta x) \leq F(x_k) + \alpha t \nabla F^\top \Delta x$$



# Levenberg-Marquardt para NLS

## Mínimos cuadrados lineales

Sea el conjunto de  $m$  puntos de datos  $(t_i, y_i)$ . Queremos ajustar una función  $\hat{y}(t, p)$  que depende, de forma no lineal, de  $n$  parámetros  $p = (p_1, \dots, p_n)$ . Suponga que  $p = (C, D)$ ,  $\hat{y} = C + Dt$ . Entonces

Square loss

$$E(p) = E(C, D) = \sum_{i=1}^m (y_i - \hat{y}_i)^2 = (y_1 - C - Dt_1)^2 + \dots + (y_m - C - Dt_m)^2$$

Queremos minimizar  $E(p)$ . Calculamos  $\nabla E$  e igualamos a cero. Tanto  $\partial E / \partial C$   $\partial E / \partial D$  son lineales, entonces nos queda el problema de encontrar  $C, D$

$$J \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} = y; \quad J^\top J \hat{p} = J^\top J \begin{bmatrix} \hat{C} \\ \hat{D} \end{bmatrix} = J^\top y$$

Ec. normales

## Mínimos cuadrados no lineales

Sea el conjunto de  $m$  puntos de datos  $(t_i, y_i)$ . Queremos ajustar una función  $\hat{y}(p)$  que depende, de forma no lineal, de  $n$  parámetros  $p = (p_1, \dots, p_n)$ . Al minimizar el error total, suma de cuadrados, esperamos  $n$  ecuaciones no lineales:

Square loss

$$E(p) = \sum_{i=1}^m (y_i - \hat{y}_i)^2 = (y - \hat{y}(p))^T (y - \hat{y}(p)) = y^T y - 2y^T \hat{y}(p) + \hat{y}(p)^T \hat{y}(p)$$

Queremos minimizar  $E(p)$ .

$$\nabla E = 2J^T (y - \hat{y}(p)) = 0, \quad \text{con } J = \frac{\partial y}{\partial p}(\hat{p})$$

## Método de Levenberg-Marquardt

Descenso del Gradiente  $p_{k+1} - p_k = -s J^\top (y - \hat{y}(p_k))$

Newton (aproximado)  $J^\top J(p_{k+1} - p_k) = J^\top (y - \hat{y}(p_k))$

$$J^\top J \approx \frac{1}{2} H$$

$$\begin{aligned} E(p + \Delta p) &\approx (y - \hat{y}(p) - J\Delta p)^\top (y - \hat{y}(p) - J\Delta p) \\ &= E(p) - 2(y - \hat{y}(p))^\top (J\Delta p) + \Delta p^\top J^\top J \Delta p \end{aligned}$$

Levenberg-Marquardt:  $(J^\top J + \lambda I)(p_{k+1} - p_k) = J^\top (y - \hat{y}(p_k))$

para un parámetro  $\lambda$ .

# Multiplicadores de Lagrange

**Problema** Considere el problema de minimizar  $F(x) = x_1^2 + x_2^2$  sobre la línea  $K : a_1x_1 + a_2x_2 = b$ .

## Multiplicadores de Lagrange

- Multiplicamos la restricción  $a_1x_1 + a_2x_2 - b$  por un multiplicador (desconocido)  $\lambda$  y lo agregamos a  $F(x)$
- Lagrangeano:  $L(x, \lambda) = F(x) + \lambda(a_1x_1 + a_2x_2 - b)$
- Igualamos las derivadas de  $L$  a cero.
- Resolvemos las ecuaciones para  $x_1, x_2, \lambda$

**Solución.**

$$\frac{\partial L}{\partial x_1} = 2x_1 + \lambda a_1 = 0 \quad \rightarrow \quad x_1 = -\lambda a_1 / 2$$

$$\frac{\partial L}{\partial x_2} = 2x_2 + \lambda a_2 = 0 \quad \rightarrow \quad x_2 = -\lambda a_2 / 2$$

$$\frac{\partial L}{\partial \lambda} = a_1 x_1 + a_2 x_2 - b = 0 \quad \rightarrow \quad \lambda = -2b / (a_1^2 + a_2^2)$$

El punto mas cercano  $(x_1^*, x_2^*)$  y el costo mínimo es

$$x_1^* = \frac{a_1 b}{a_1^2 + a_2^2}, \quad x_2^* = \frac{a_2 b}{a_1^2 + a_2^2}, \quad F(x_1^*, x_2^*) = \frac{b^2}{a_1^2 + a_2^2}$$

Además la derivada del costo mínimo respecto al nivel de la restricción  $b$  es

$$\frac{d}{db} \left( \frac{b^2}{a_1^2 + a_2^2} \right) = \frac{2b}{a_1^2 + a_2^2} = -\lambda.$$

## Minimización de función cuadrática con restricciones lineales

Sea  $S \in \mathbb{R}^{n \times n}$  simétrica y definida positiva,  $A \in \mathbb{R}^{n \times m}$ ,  $b \in \mathbb{R}^m$ . El problema es:

$$\text{Minimizar } F(x) = \frac{1}{2}x^\top Sx \text{ sujeto a } A^\top x = b$$

Ahora tenemos  $m$  mult. de Lagrange  $\lambda = (\lambda_1, \dots, \lambda_m)$  y el Lagrangeano es:

$$L(x, \lambda) = \frac{1}{2}x^\top Sx + \lambda^\top (A^\top x - b)$$

Calculando las derivas de  $L$  e igualando a cero obtenemos  $n + m$  ecuaciones:

$$\frac{\partial L}{\partial x} = Sx + A\lambda = 0, \quad \frac{\partial L}{\partial \lambda} = A^\top x - b = 0$$



## Solución

Complemento  
de Schur

$$(x^*, \lambda^*) \text{ satisfacen: } \begin{bmatrix} S & A \\ A^\top & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix} \rightarrow \begin{bmatrix} S & A \\ 0 & -A^\top S^{-1} A \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ b \end{bmatrix}$$

$$\lambda^* = -(A^\top S^{-1} A)^{-1} b, \quad x^* = S^{-1} A (A^\top S^{-1} A)^{-1} b$$

$$\text{Costo mínimo: } F(x^*, \lambda^*) = \frac{1}{2} b^\top (A^\top S^{-1} A)^{-1} b$$

$$\text{Gradiente del costo: } \frac{\partial F}{\partial b}(x^*, \lambda^*) = (A^\top S^{-1} A)^{-1} b = -\lambda^*$$

## Observación

Punto de silla

La función  $L = \frac{1}{2} x^\top S x + \lambda^\top (A^\top x - b)$  es convexa en  $x$  y cóncava en  $\lambda$ .

## Propiedad minimax/maxmin

Suponga que separamos dos problemas:

- Minimizar  $L(x, \lambda)$  para  $\lambda$  fijo, y luego maximizar sobre  $\lambda$

$$\max_{\lambda} \min_x L = \frac{1}{2} b^{\top} (A^{\top} S^{-1} A)^{-1} b$$

- Maximizar  $L(x, \lambda)$  para  $x$  fijo y luego minimizar sobre  $x$

$$\min_x \max_{\lambda} L = \frac{1}{2} (A^{\top} S^{-1} A)^{-1} b$$

En el punto de silla  $(x^*, \lambda^*)$  tenemos:

$$\frac{\partial L}{\partial x} = \frac{\partial l}{\partial \lambda} = 0 \quad \text{y} \quad \min_x \max_{\lambda} L = \max_{\lambda} \min_x L$$