



Métodos de optimización basados en gradientes

Problema General de Optimización

$\min f(x); x$ vector de \mathbf{R}^n

Sujeto a: $Ax \leq B$

$$Cx = D$$

Ax y Cx son funciones lineales

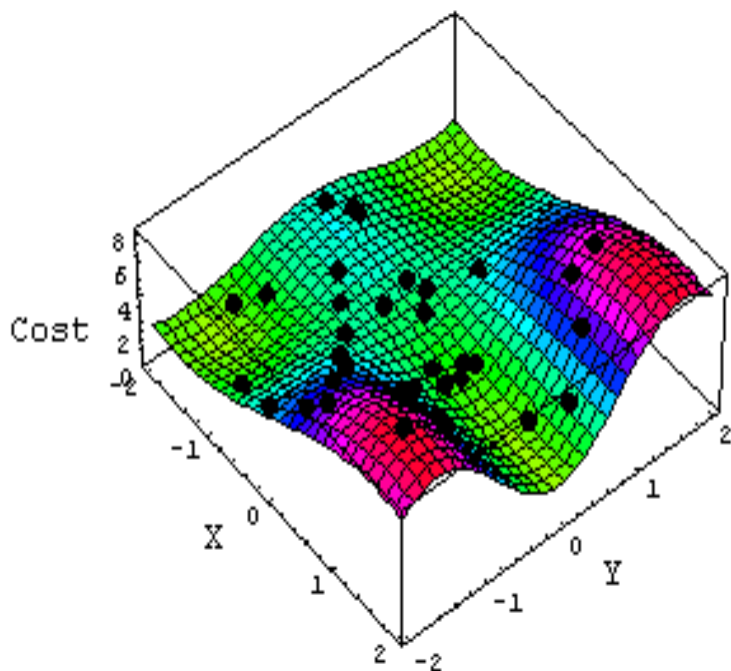
$$LB \leq x \leq UB$$

LB y UB son los límites superior e inferior del vector x

$$G(x) \leq 0$$

$$H(x) = 0$$

$G(x)$ y $H(x)$ son funciones vectoriales



Recuerde que en entrenamiento supervisado el problema general es (ej. usando MSE)

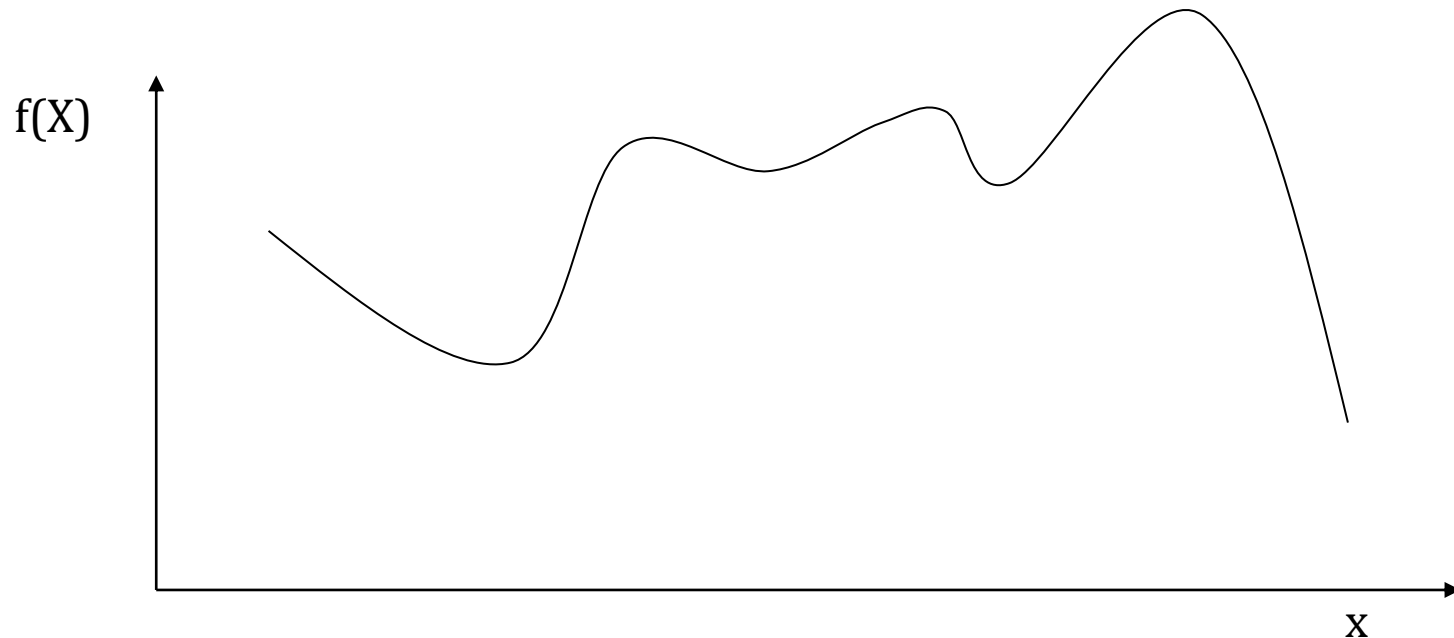
$$\text{Minimizar } \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2$$

$$\text{Donde } f(x) = \begin{cases} a+bX \\ 1/(1+e^{-(a+bX)}) \\ \text{red neuronal} \\ \text{etc.} \end{cases}$$

Optimización no restringida

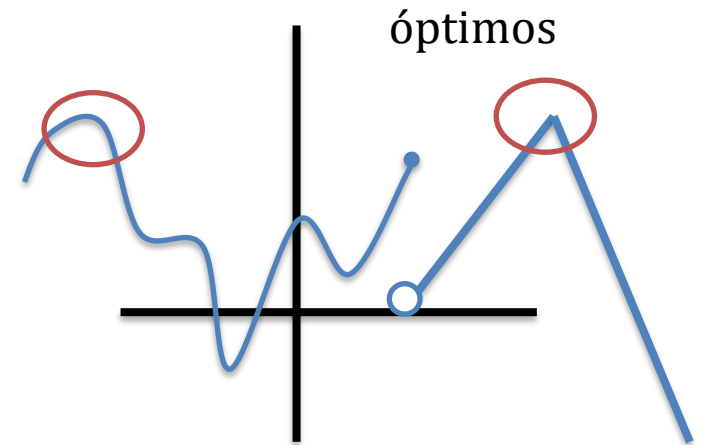
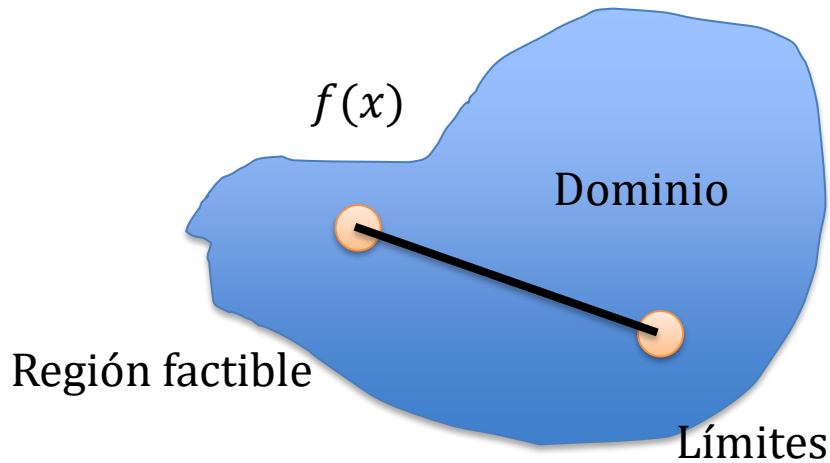
Corresponde al problema de:

Minimizar $f(x)$

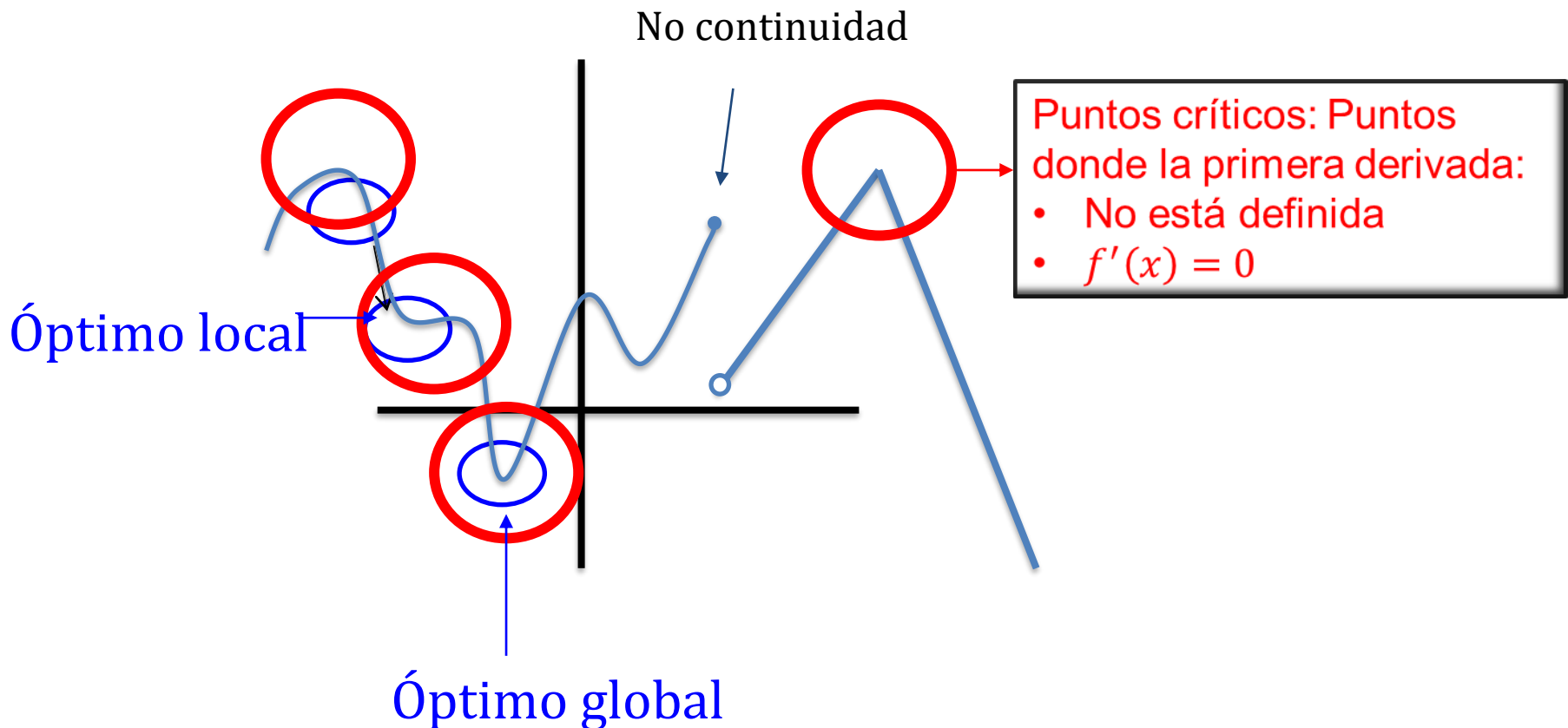


Por ahora, vamos a olvidarnos de las restricciones y vamos a concentrarnos solamente en la función objetivo.

Conceptos básicos sobre funciones

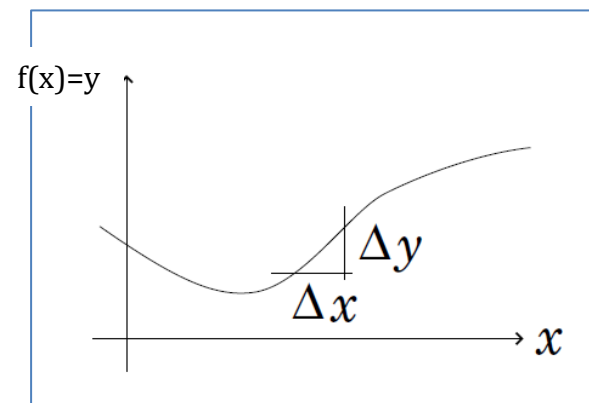
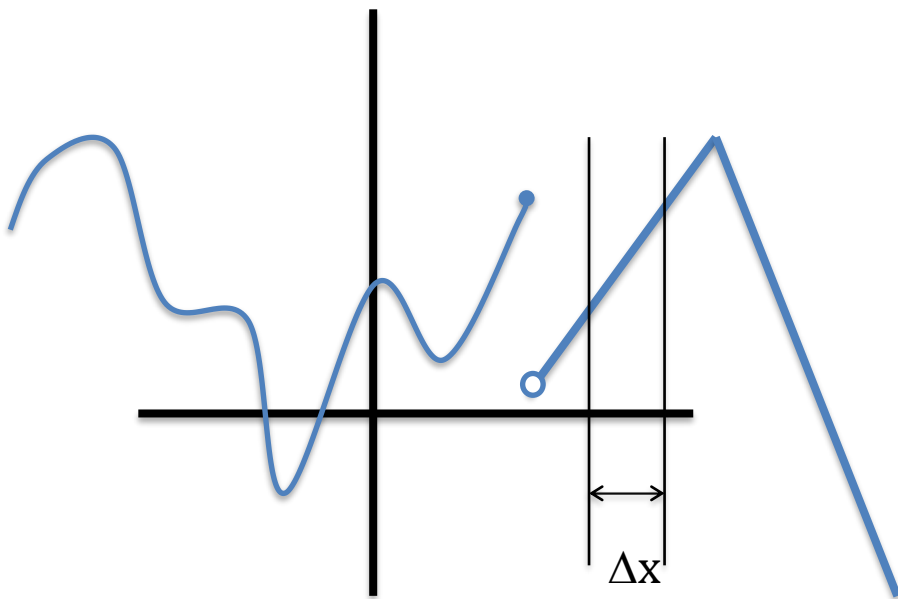


Conceptos básicos sobre funciones



Derivada:

$$\frac{df(x)}{dx} = f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$



Recuerde: Algunas reglas de derivación

$$1. \quad \frac{d}{dx} c = 0$$

$$2. \quad \frac{d}{dx} x = 1$$

$$3. \quad \frac{d}{dx} cv = c \frac{dv}{dx}$$

$$4. \quad \frac{d(u + v + w)}{dx} = \frac{du}{dx} + \frac{dv}{dx} + \frac{dw}{dx}$$

$$5. \quad \frac{d(x^n)}{dx} = nx^{n-1}$$

$$6. \quad \frac{d}{dx} v^n = nv^{n-1} \frac{dv}{dx}$$

$$7. \quad \frac{d}{dx} (uv) = u \frac{dv}{dx} + v \frac{du}{dx}$$

$$8. \quad \frac{d}{dx} \left(\frac{u}{v} \right) = \frac{v \frac{du}{dx} - u \frac{dv}{dx}}{v^2}$$

$$f(x) = ax + bx \rightarrow \frac{\partial f(x)}{\partial x} = a$$

$$f(x) = \ln(x) \rightarrow \frac{\partial f(x)}{\partial x} = \frac{1}{x}$$

$$f(x) = e^{ax} \rightarrow \frac{\partial f(x)}{\partial x} = a e^{ax}$$

Ejemplo: Derivada de la función sigmoidea

$$f(x) = \frac{1}{1 + e^{-x}} = (1 + e^{-x})^{-1} \quad \rightarrow$$

$$\begin{aligned} \frac{\partial f(x)}{\partial x} &= -(1 + e^{-x})^{-2}(-e^{-x}) = -\frac{1}{(1 + e^{-x})^2}(-e^{-x}) \\ &= \left(\frac{1}{1 + e^{-x}}\right) \left(\frac{1 + e^{-x} - 1}{1 + e^{-x}}\right) \end{aligned}$$

$$\boxed{\frac{\partial f(x)}{\partial x} = f(x)(1 - f(x))}$$

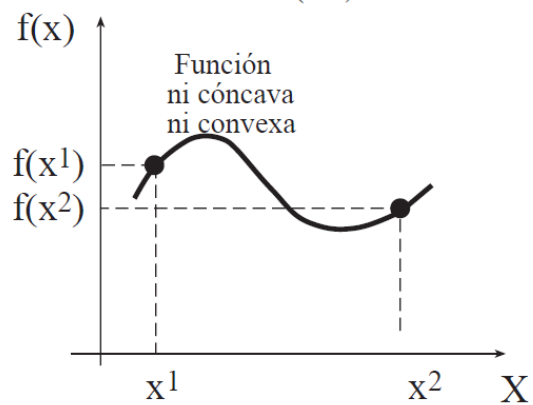
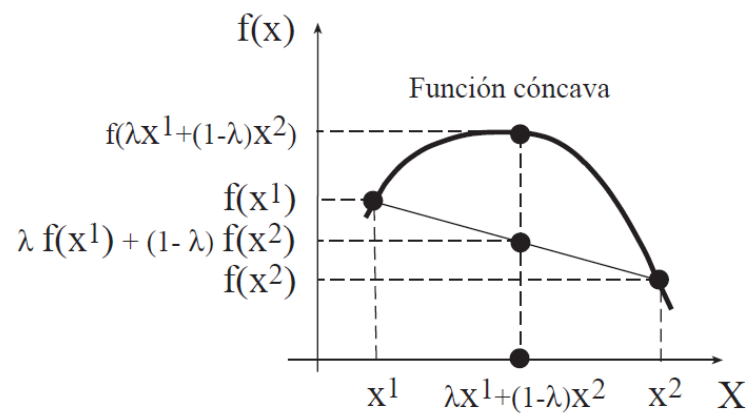
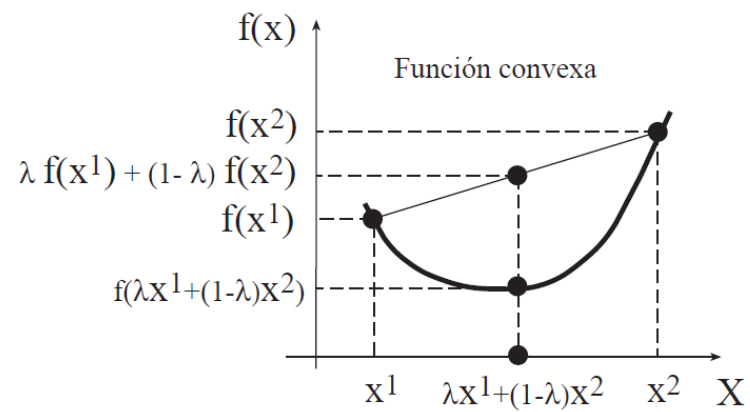
Lo anterior hace que sea fácilmente derivable.

Condiciones necesarias para la optimalidad de problemas no restringidos

Función continuamente diferenciable *en \mathbf{x} si todas sus derivadas parciales son continuas en \mathbf{x} .*

La condición necesaria para que una solución sea un mínimo local es que el gradiente sea cero

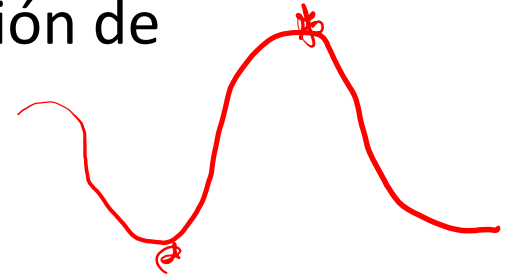
Convexidad: permite garantizar que todo óptimo local del PNL también lo es global.



Optimización unidimensional no restringida

La solución óptima x^* debe cumplir la condición de estacionaridad:

$$\nabla f(x^*) = 0$$



que es una condición necesaria pero no suficiente.

La otra condición es:

$$\nabla^2 f(x^*) < 0$$

máximo local en x^*

$$\nabla^2 f(x^*) > 0$$

Mínimo local en x^*

$$\frac{df(x)}{dx} = f'(x) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}$$

si $f'(x^*)=0$, entonces el signo del cambio en la función al pasar de x^* a $(x^*+\Delta x)$ dependerá del signo de $f''(x^*)=0$, por lo tanto:

- i) si $f''(x^*) < 0 \Rightarrow f(x^* + \Delta x) < f(x^*) \Rightarrow$ máximo local en x^*
- ii) si $f''(x^*) > 0 \Rightarrow f(x^* + \Delta x) > f(x^*) \Rightarrow$ mínimo local en x^*

En una función multidimensional

- Sea $\mathbf{x} = [x_1, \dots, x_n]^T$ y sea $f(\mathbf{x})$ del vector \mathbf{x} .

- $$g(\mathbf{x}) = \nabla f(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \end{bmatrix}$$

Condiciones de optimalidad

Sea el problema a Optimizar $f(\mathbf{x})=f(x_1,x_2,\dots,x_n)$

La solución óptima \mathbf{x}^* habrá de cumplir la condición de estacionaridad:

$$\nabla f(\mathbf{x}^*) = 0$$

que es una condición necesaria pero no suficiente.

Y que la matriz Hessiana sea definida positiva (para mínimo) o negativa (para máximo)

$$H \quad f(x_1, x_2, \dots, x_n) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \dots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

Ejemplo

$$\text{Max } f(x_1, x_2) = -(x_1 - 3)^2 - (x_2 - 2)^2$$

El gradiente de $f(x_1, x_2)$:

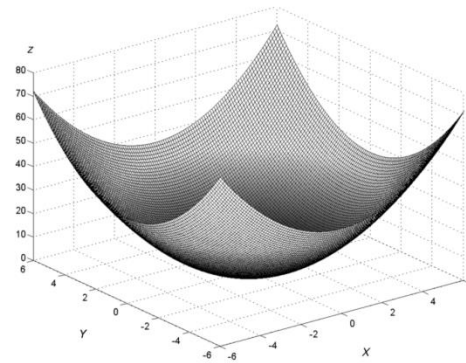
$$\nabla f(x_1, x_2) = \left[\frac{\partial f(x_1, x_2)}{\partial x_1}, \frac{\partial f(x_1, x_2)}{\partial x_2} \right]$$

$$\nabla f(x_1, x_2) = [-2(x_1 - 3), -2(x_2 - 2)] = [0, 0]$$

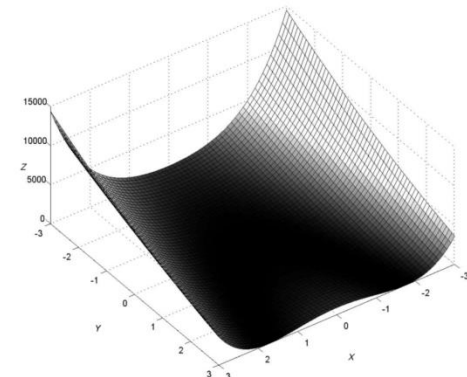
$$x_1 = 3, x_2 = 2$$

Pero casi nunca es tan fácil...otra opción Ensayo y error

Ejemplos de funciones complejas para derivar

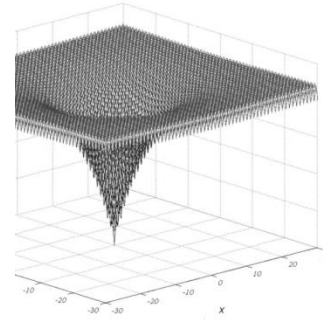


(a) f_1 : Función esférica

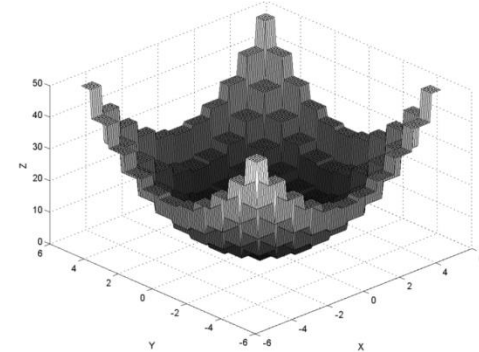


(b) f_2 : Función Rosenbrock

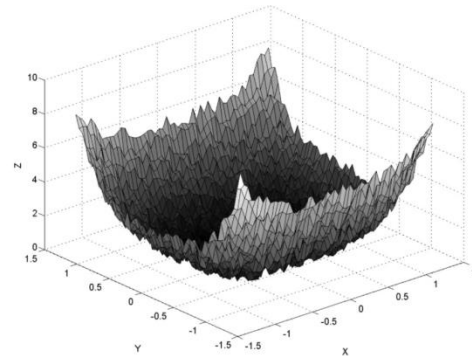
Función de prueba	Dim(D)	Región	mín
$f_1(x) = \sum_{i=1}^D x_i^2$	2	$[-5, 12, 5, 12]^D$	0
$f_2(x) = \sum_{i=1}^n 100(x_i^2 - x_{i+1})^2 + (1 - x_i^2)^2$	2	$[-5, 12, 5, 12]^D$	0
$f_3(x) = -20 \cdot \exp(-0.2 \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n x_i^2}) - \exp(\frac{1}{n} \cdot \sum_{i=1}^n \cos(2\pi x_i)) + 20 + \exp(1)$	2	$[-30, 30]^D$	0
$f_4(x) = \sum_{i=1}^D ((x_i + 0.5))^2$	2	$[-5, 12, 5, 12]^D$	0
$f_5(x) = \sum_{i=1}^D i(x_i^4) + \text{random}[0, 1)$	2	$[-1, 28, 1, 28]^D$	0
$\frac{1}{f_6(x)} = \frac{1}{500} \sum_{j=1}^{25} \frac{1}{j + \sum_{i=1}^2 (x_i - a_{ij})^6}$	2	$[-65, 536, 65, 356]^D$	≈ 1



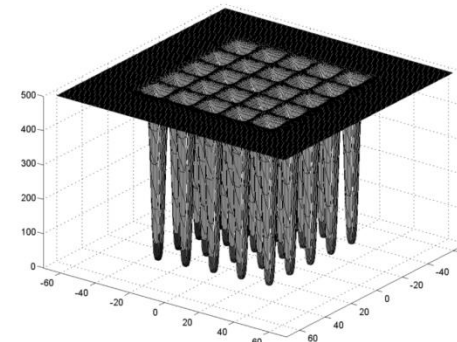
(c) f_3 : Función Ackley



(d) f_4 : Función Step



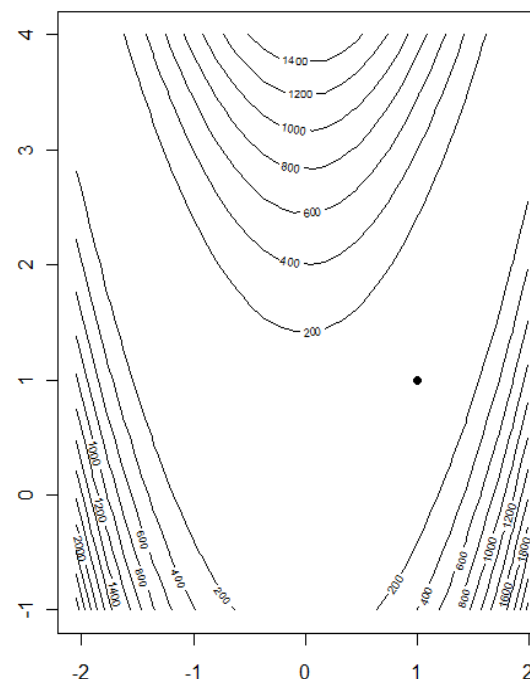
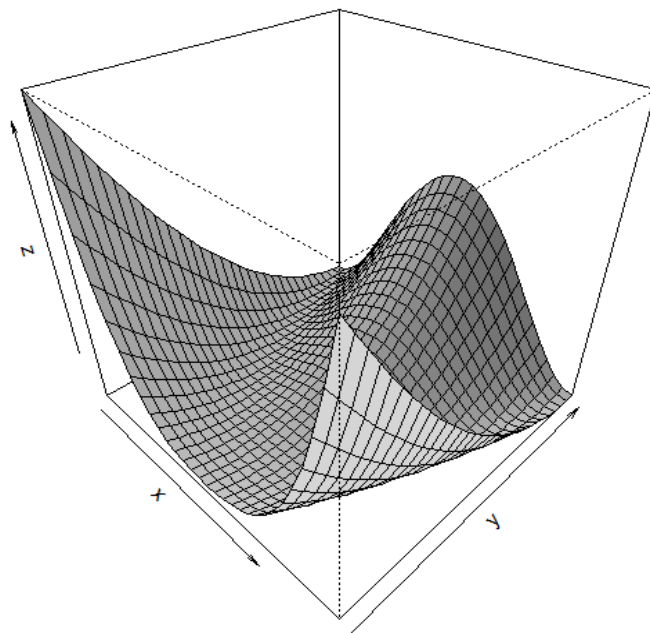
(e) f_5 : Función Cuadrática ruidosa



(f) f_6 : Función Foxholes

Ejemplo: función Rosenbrock

Minimizar $f(x, y) = 100(x^2 - y)^2 + (1 - x)^2$



$x \in [-2.048, 2.048], y \in [-1.000, 4.000]$.

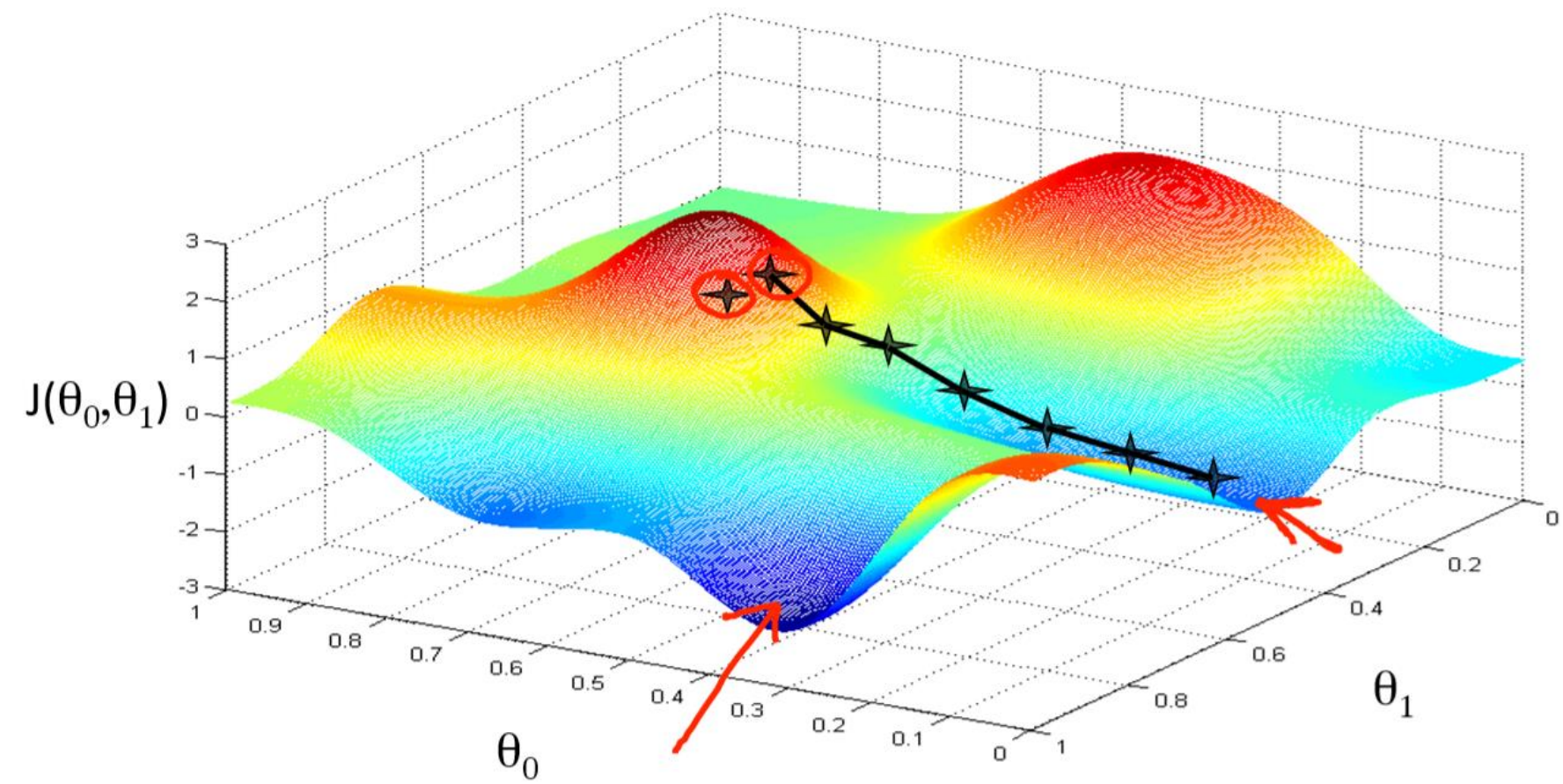
El mínimo esta en $x^* = (1.0, 1.0)$: $f(1.0, 1.0) = 0.0$

Métodos basados en gradiente

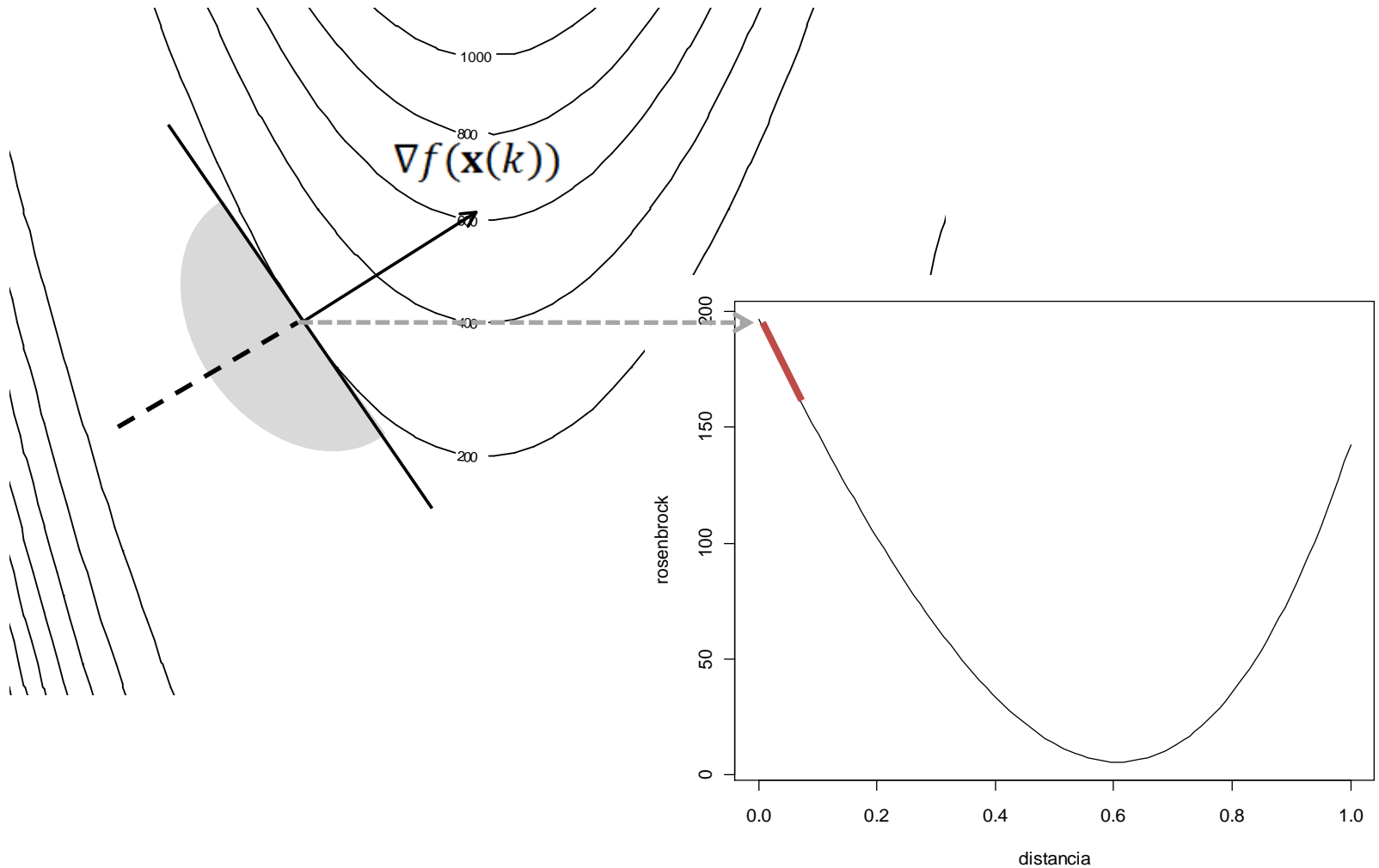
Se basan en:

- la pendiente en un punto de una función es su derivada respecto a la posición del punto.

Estrategia: el algoritmo comienza en un punto cualquiera y se mueve de forma iterativa en direcciones que estima conveniente para la minimización de $f(x)$, según el gradiente en ese punto, hasta que se llegue a una solución que cumpla las condiciones de optimalidad.



Interpretación geométrica del gradiente



Algunos métodos conocidos son:

- Método del gradiente
- Método de Newton-Raphson
- Método de Fletcher-Powell
- etc.

Algunos usan la primera derivada y otros la segunda.

Algoritmo básico:

Si $\mathbf{x}(k)$ es el punto de evaluación en la iteración k :

$$\begin{aligned}\mathbf{x}(k+1) &= \mathbf{x}(k) + \Delta\mathbf{x}(k) \\ &= \mathbf{x}(k) + \alpha(k) \mathbf{d}^*(k)\end{aligned}$$

Donde: $\Delta\mathbf{x}(k)$: es un vector que va de $\mathbf{x}(k)$ a $\mathbf{x}(k+1)$

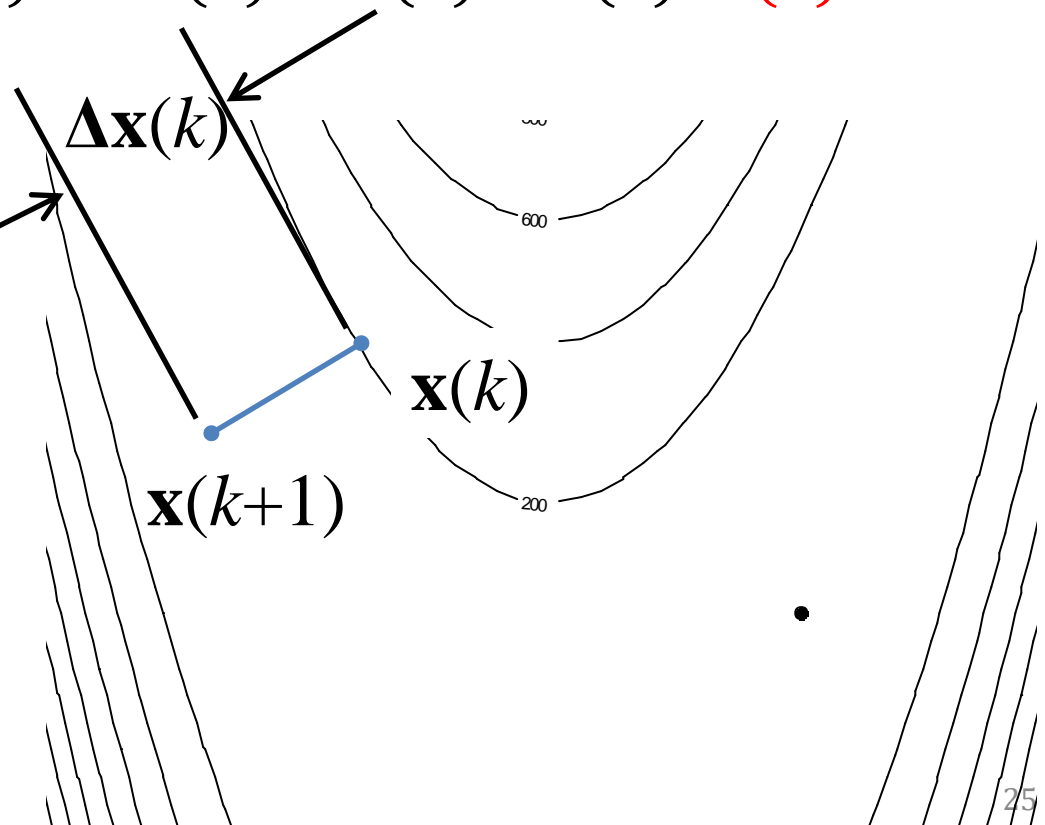
- $\mathbf{d}^*(k)$: es un vector unitario en la dirección de $\Delta\mathbf{x}(k)$
- $\alpha(k)$ escalares (tamaño de paso)

Métodos que utilizan el criterio de la primera derivada

Se base en la aproximación: $f(\mathbf{x} + \boldsymbol{\delta}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T \times \boldsymbol{\delta} + H.O.T$
H.O.T son los términos de orden superior

$$\mathbf{x}(k+1) = \mathbf{x}(k) + \Delta \mathbf{x}(k) = \mathbf{x}(k) + \lambda(k) \mathbf{s}^*(k)$$

Con $\mathbf{s}^*(k) = -\frac{\nabla f(\mathbf{x}(k))}{\|\nabla f(\mathbf{x}(k))\|}$



Pseudocódigo para un $\lambda(k)$ fijo

```
01:    $k \leftarrow 1; \lambda; K$   
02:   loop while  $k \leq K$   
03:        $\mathbf{x}(k + 1) = \mathbf{x}(k) + \lambda(k) \mathbf{s}^*(k)$   
04:   end loop
```

Con $\mathbf{s}^*(k) = -\frac{\nabla f(\mathbf{x}(k))}{\|\nabla f(\mathbf{x}(k))\|}$

Ejemplo numérico

Minimizar función de Rosenbrock

$$f(x, y) = 100 (x^2 - y)^2 + (1 - x)^2$$

Gradiente:

$$\nabla f(x, y) = \begin{bmatrix} 400x(x^2 - y) - 2(1 - x) \\ -200(x^2 - y) \end{bmatrix}$$

Punto inicial cualquiera: $\mathbf{x}(1) = \begin{bmatrix} -0.78 \\ 2.00 \end{bmatrix}$

$$\nabla f(-0.78, 2.00) = \begin{bmatrix} 430.62 \\ 278.32 \end{bmatrix}$$

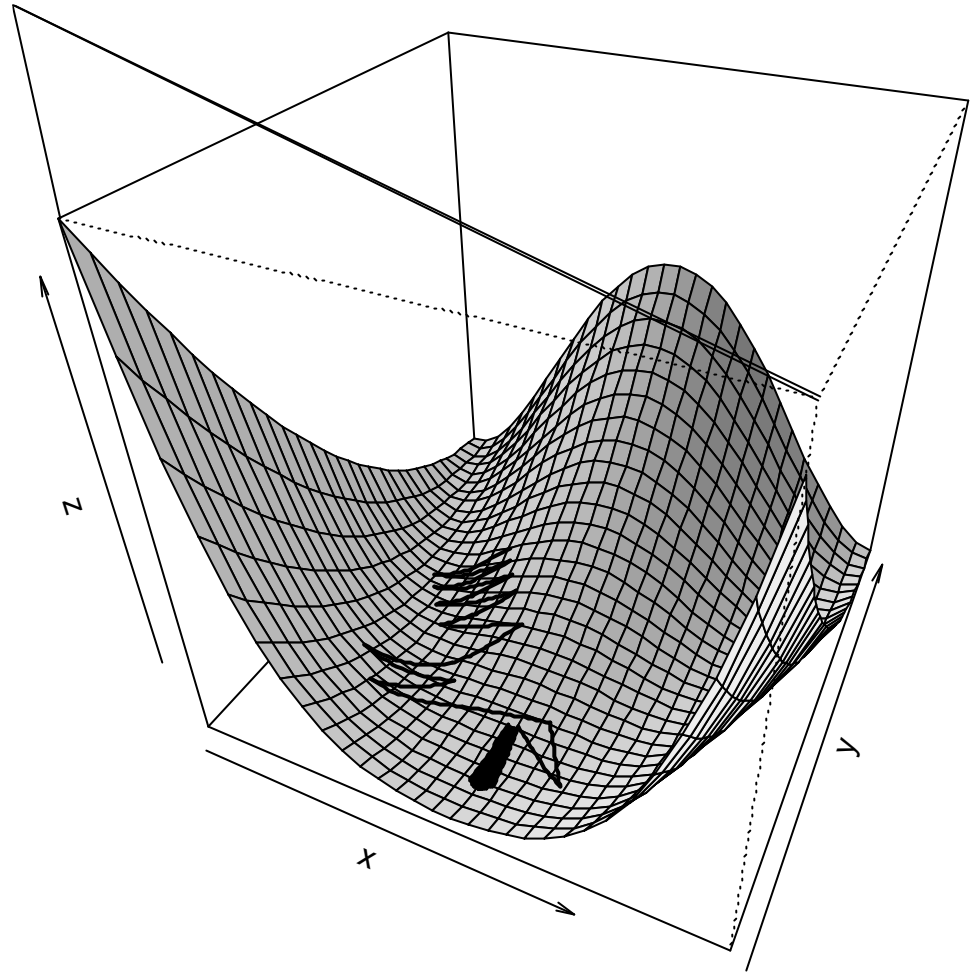
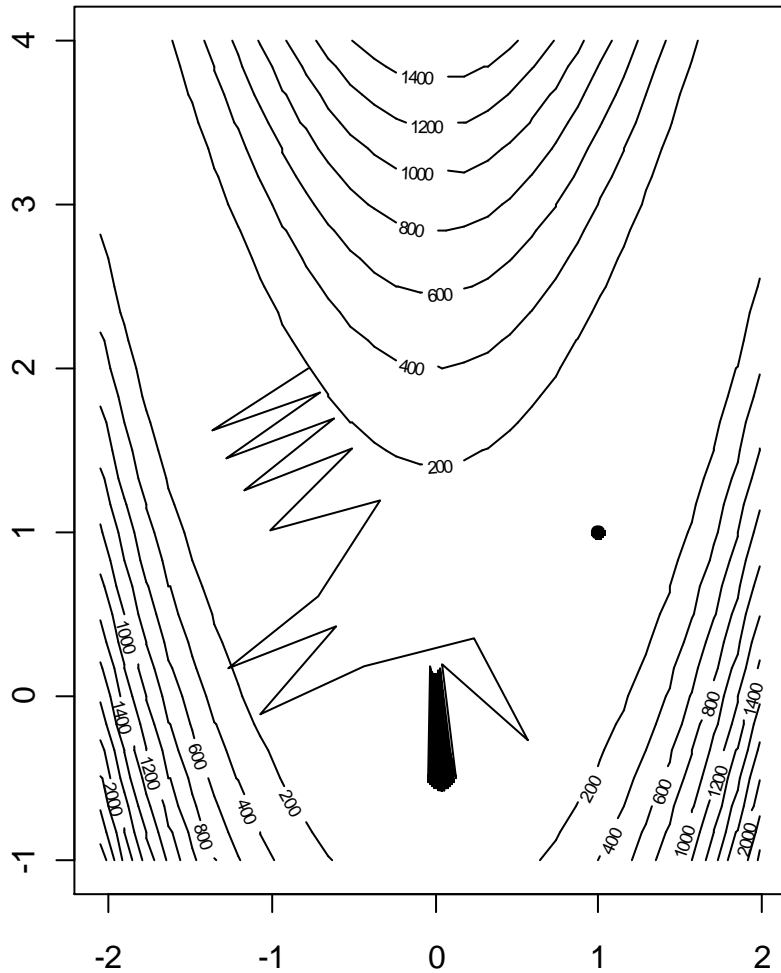
$$(430.62^2 + 278.32^2)^{0.5} = 512.73$$

suponga $\lambda=0.7$

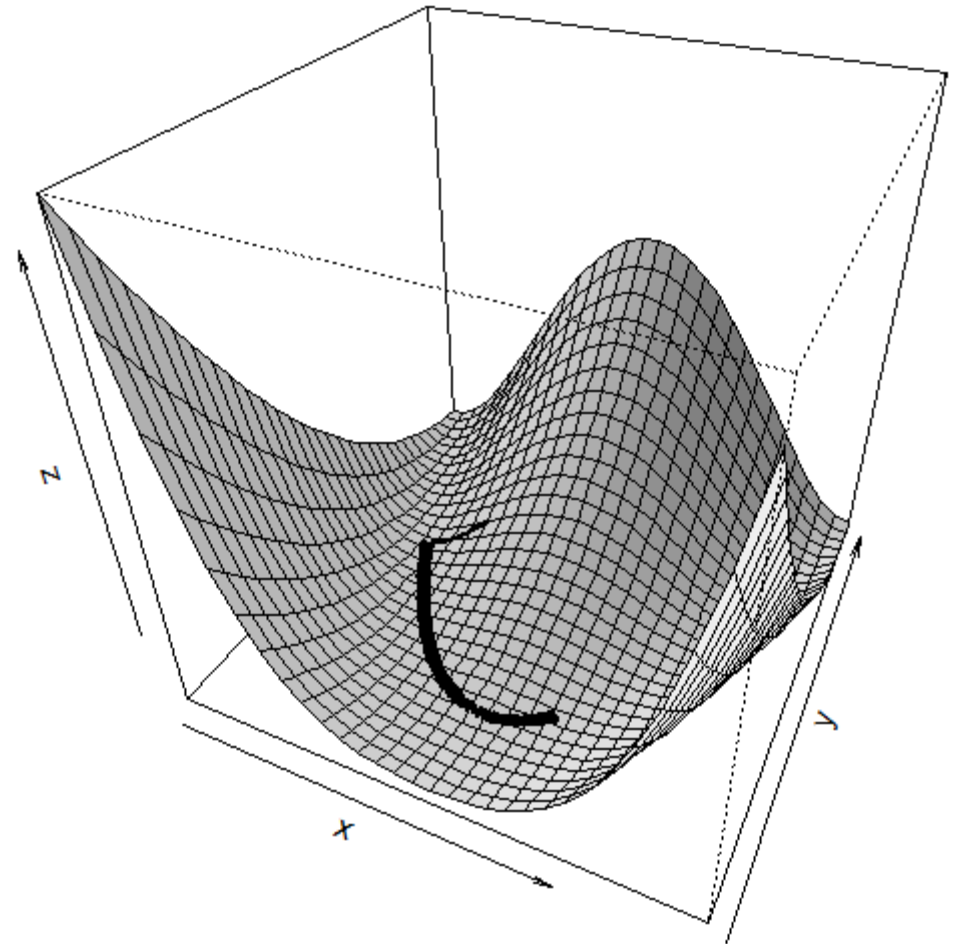
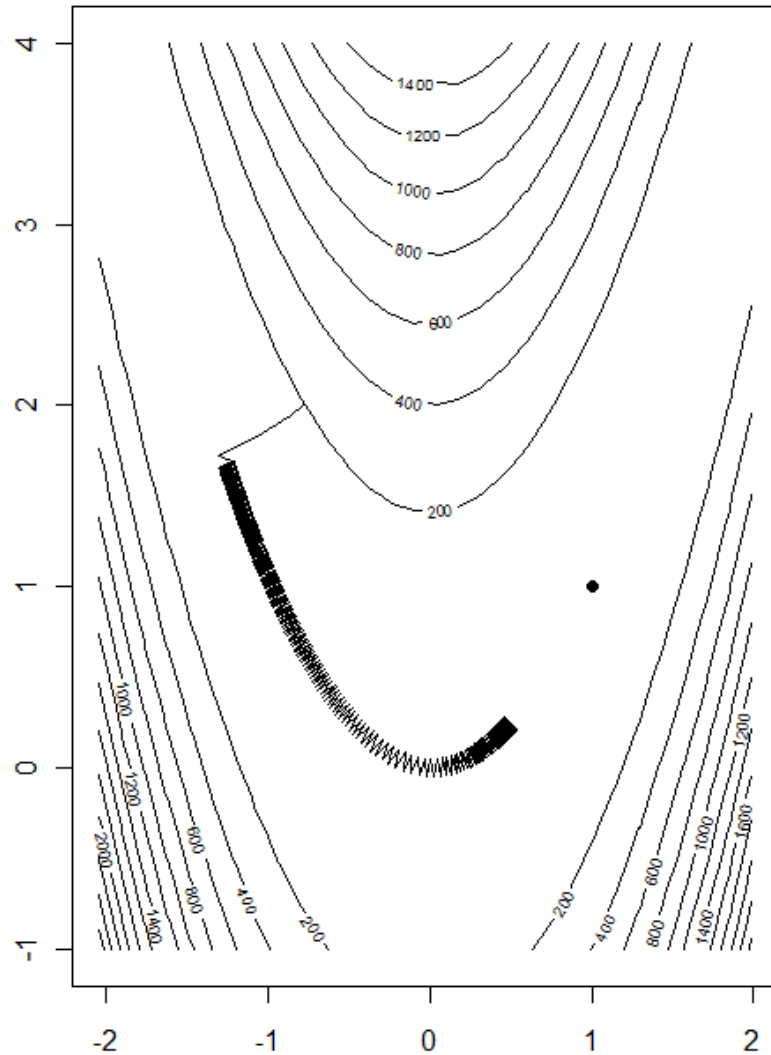
$$\mathbf{x}(2) = \mathbf{x}(1) - \lambda \mathbf{g}(1) = \begin{bmatrix} -0.78 \\ 2.00 \end{bmatrix} - 0.7 \begin{bmatrix} 430.62 \\ 512.73 \\ 278.32 \\ 512.73 \end{bmatrix} = \begin{bmatrix} -1.3679 \\ 1.62003 \end{bmatrix}$$

Sigue
iterando

Resultados para $\lambda(k)=0.7$, $k=1000$



Resultados para $\lambda(k)=0.1$, $k=1000$



¿Cuándo para?

Cuando se cumpla uno de los criterios:

- C_1 : el número máximo de iteraciones.
- C_2 es la tolerancia para la función objetivo. Hay convergencia cuando el cambio en la función objetivo es menor que este valor.
- C_3 es la tolerancia para \mathbf{x} . Hay convergencia cuando el cambio en \mathbf{x} es menor que este valor.
- C_5 es la tolerancia para el gradiente