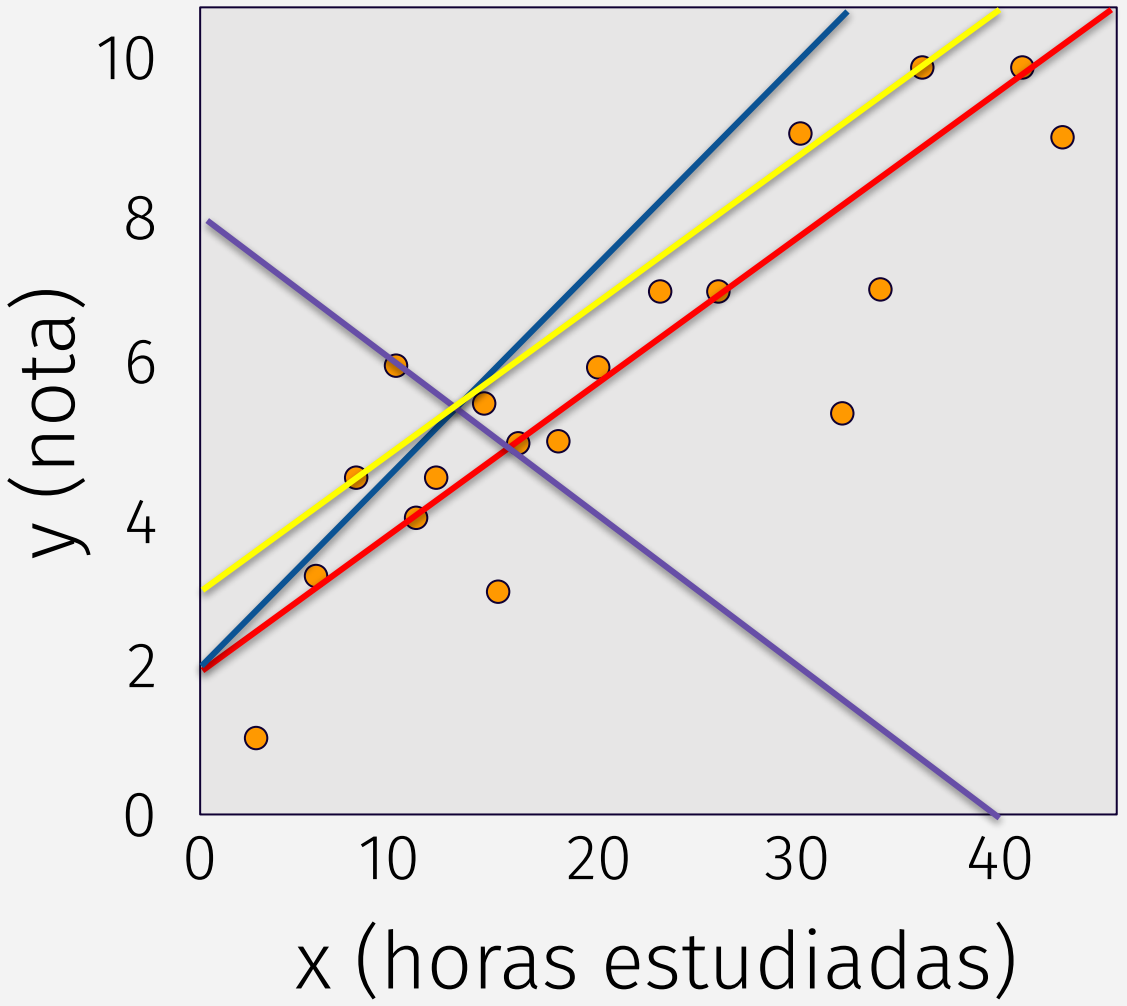
Regresión Lineal
Descenso de Gradiente

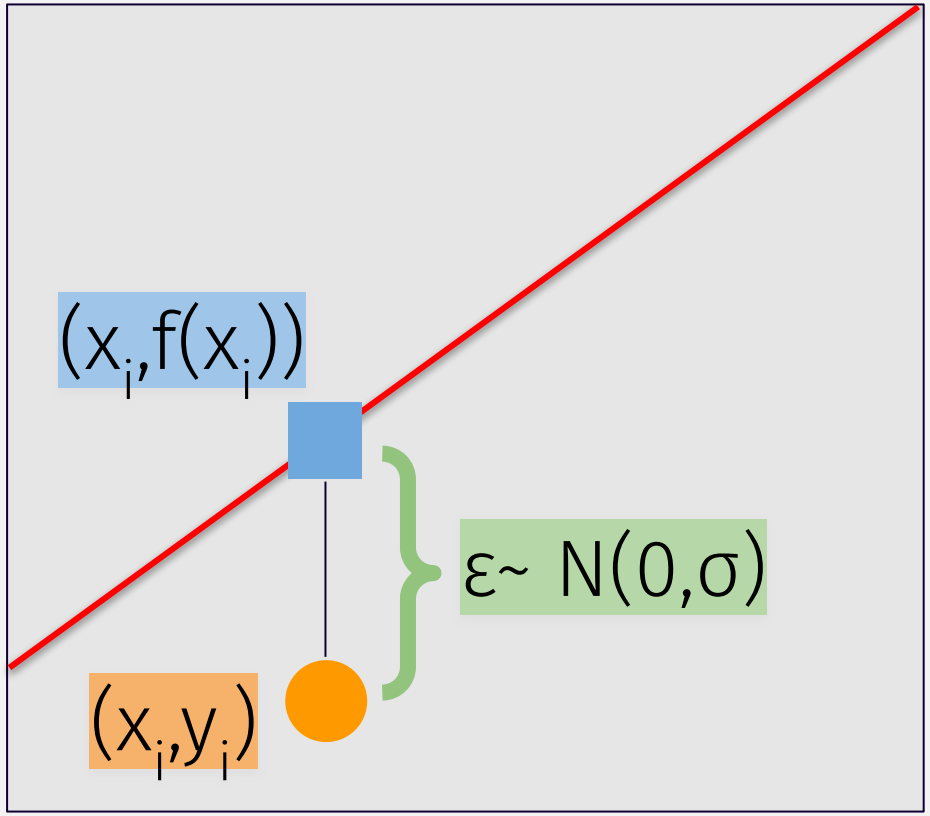
Optimización del Error **E**

- Problema:
  - Datos: ejemplos ($\mathbf{x_i}$, $\mathbf{y_i}$)
  - Parámetros: **m** y **b**
  - Función de error: **E**
    - Encontrar **m** y **b** que minimicen **E**(**m**,**b**,**x**,**y**)
    - **E** es derivable
      - Respecto de **m** y **b**



Regresioń Lineal con MLE

- MLE: Maximum Likelihood Estimation
  - (Método de Máxima Verosimilitud)
  - Método estadístico
- Asumo $y = f(x) + \varepsilon = m\,x + b + \varepsilon$
  - $\varepsilon \sim N(0,\sigma)$
  - m, b y σ son **parámetros**
    - Busco estimadores m', b', σ'
  - $\varepsilon = f(x) - y$
    - $(f(x) - y) \sim N(0,\sigma)$
    - Función de verosimilitud L
      - $L(m,b,\sigma) = \Pi_i^n \text{Gaussiana}[\mu=0,\sigma,](f(x_i) - y_i)$
    - MLE → fórmula para m', b' y σ'

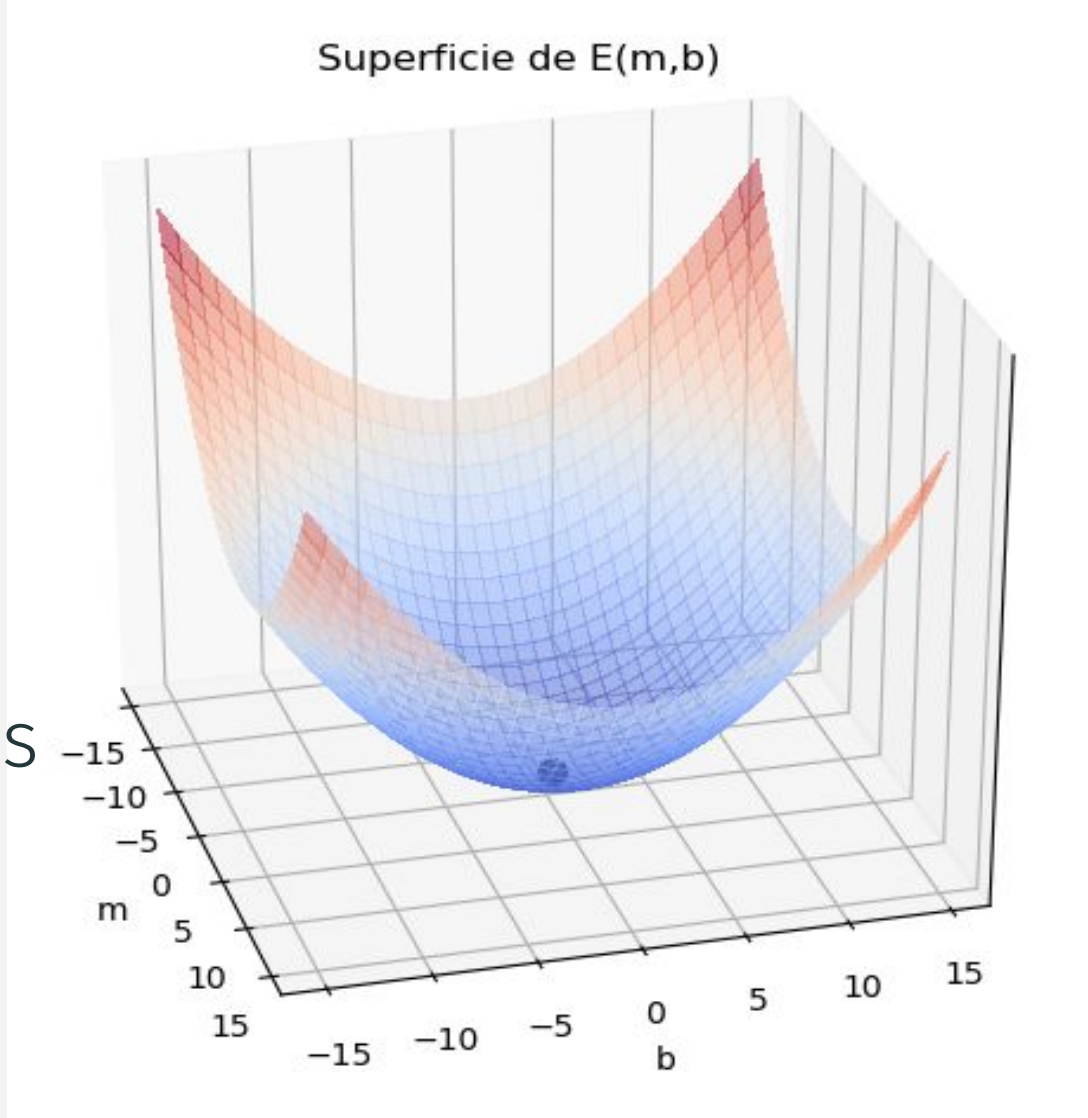


Métodos clásicos de Optimización para RL

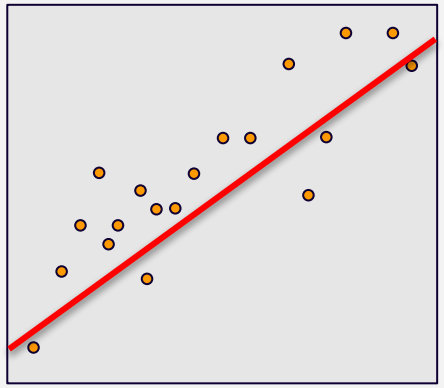
Métodos clásicos (analíticos)

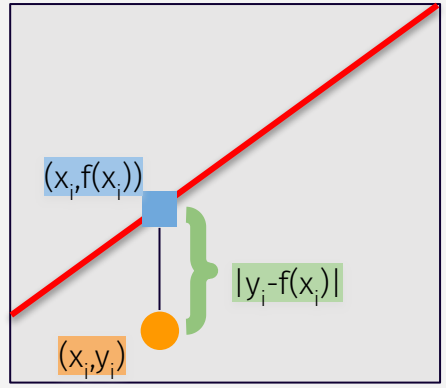
- Cálculo: $\frac{\partial E}{\partial b} = 0$ y $\frac{\partial E}{\partial m} = 0$, despejo m y b.
- Álgebra lineal: $Y = mX + b = \begin{pmatrix} 1 & x \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix}$, proyecto.
- Probabilidades: $y = mx + b + e$ con $e \sim \mathcal{N}(0, \sigma)$, estimo m y b con MLE

- Ventajas
  - Solución analítica, simple
- Desventajas
  - Poco eficiente con muchos datos
    - n >= 10^6
  - Problemas numéricos
- **Alternativa**
  - Descenso de gradiente

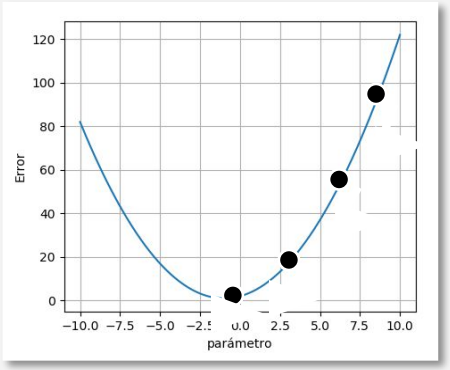


Descenso de gradiente para Regresión Lineal





- Modelo
  - $f(x) = m \times x + b$
  - $E(m,b) = (1/n) \sum_i^n E_i(m,b)$
  - $E_i(m,b) = (y_i - m\, x_i + b)^2$
- Ecuaciones de Descenso de Gradiente
  - $b = b - \alpha \ \delta E(m,b)/\delta b$
  - $m = m - \alpha \ \delta E(m,b)/\delta m$
    - **Derivadas parciales?**
      - $\delta E(m,b)/\delta b = 1/n \sum_i^n 2\,(y_i - f(x_i))$
      - $\delta E(m,b)/\delta m = 1/n \sum_i^n 2\,(y_i - f(x_i))\, x_i$
        - ¿Cómo las obtengo?



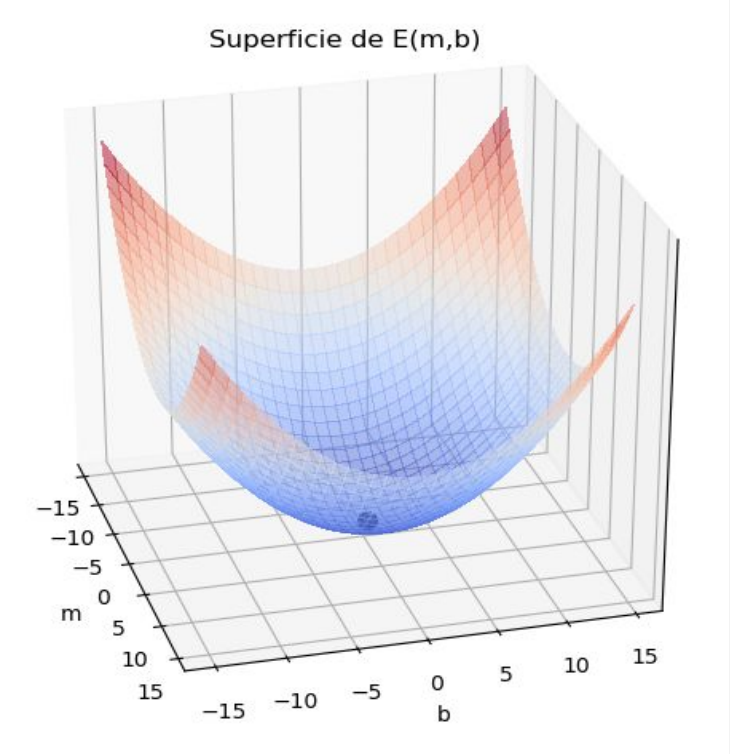
Repaso de derivadas

1. $\delta \mathbf{f(x)+g(f)}/\delta x = \delta \mathbf{f(x)}/\delta x + \delta \mathbf{g(x)}/\delta x$
2. $\delta \mathbf{c}/\delta x = 0$
   a. $\delta \mathbf{f(x)-g(f)}/\delta x = \delta \mathbf{f(x)}/\delta x - \delta \mathbf{g(x)}/\delta x$
   b. $\delta \mathbf{f(x)+c}/\delta x = \delta f(x)/\delta x$
   c. $\delta \left[\mathbf{\Sigma_i^n\, f(x_i)}\right] / \delta x = \mathbf{\Sigma_i^n} \left[\delta \mathbf{f(x_i)} / \delta x\right]$
3. $\delta \mathbf{f(g(x))}/\delta x = \delta \mathbf{f(g(x))}/\delta \mathbf{g(x)} \;\; \delta \mathbf{g(x)}/\delta x$
   a. $\delta \mathbf{(3x-4)^2}/\delta x = \delta \mathbf{(3x-4)^2}\, /\delta \mathbf{(3x-4)} \;\; \delta \mathbf{(3x-4)}/\delta x$

$= 2(3x-4)\, 3 = 6\,(3x-4)$

Derivada del error respecto de b $\mathbf{\delta E(m,b)/\delta b}$

$$\delta \mathbf{E(m,b)/\delta b} = \delta(\mathbf{1/n\ \Sigma_i^n\ E_i(m,b)})/\delta b$$
$$= \mathbf{1/n\ \Sigma_i^n}\ \delta E_i(m,b)/\delta b$$
$$= 1/n\ \Sigma_i^n\ \delta(\mathbf{(y_i - f(x_i))^2})/\delta b$$
$$= 1/n\ \Sigma_i^n\ \mathbf{2\ (y_i - f(x_i))}\ \delta(\mathbf{y_i - f(x_i)})/\delta b$$
$$= 1/n\ \Sigma_i^n\ 2\ (y_i - f(x_i))\ \delta\mathbf{(y_i - m\ x_i + b)}/\delta b$$
$$= 1/n\ \Sigma_i^n\ 2\ (y_i - f(x_i))\ \mathbf{\delta b/\delta b}$$
$$= 1/n\ \Sigma_i^n\ 2\ (y_i - f(x_i))$$



Derivada del error respecto de m **δE(m,b)/δm**

$$\delta \mathbf{E(m,b)/\delta m} = \delta(\mathbf{1/n\ \Sigma_i^n\ E_i(m,b)})/\delta m$$

$$= \mathbf{1/n\ \Sigma_i^n}\ \delta E_i(m,b)/\delta m$$

$$= 1/n\ \Sigma_i^n\ \delta(\mathbf{(y_i - f(x_i))^2})/\delta m$$

$$= 1/n\ \Sigma_i^n\ \mathbf{2\ (y_i - f(x_i))}\ \delta(\mathbf{y_i - f(x_i)})/\delta m$$

$$= 1/n\ \Sigma_i^n\ 2\ (y_i - f(x_i))\ \delta\mathbf{(y_i - m\ x_i + b)}/\delta m$$

$$= 1/n\ \Sigma_i^n\ 2\ (y_i - f(x_i))\ \ m\ x_i/\delta m$$

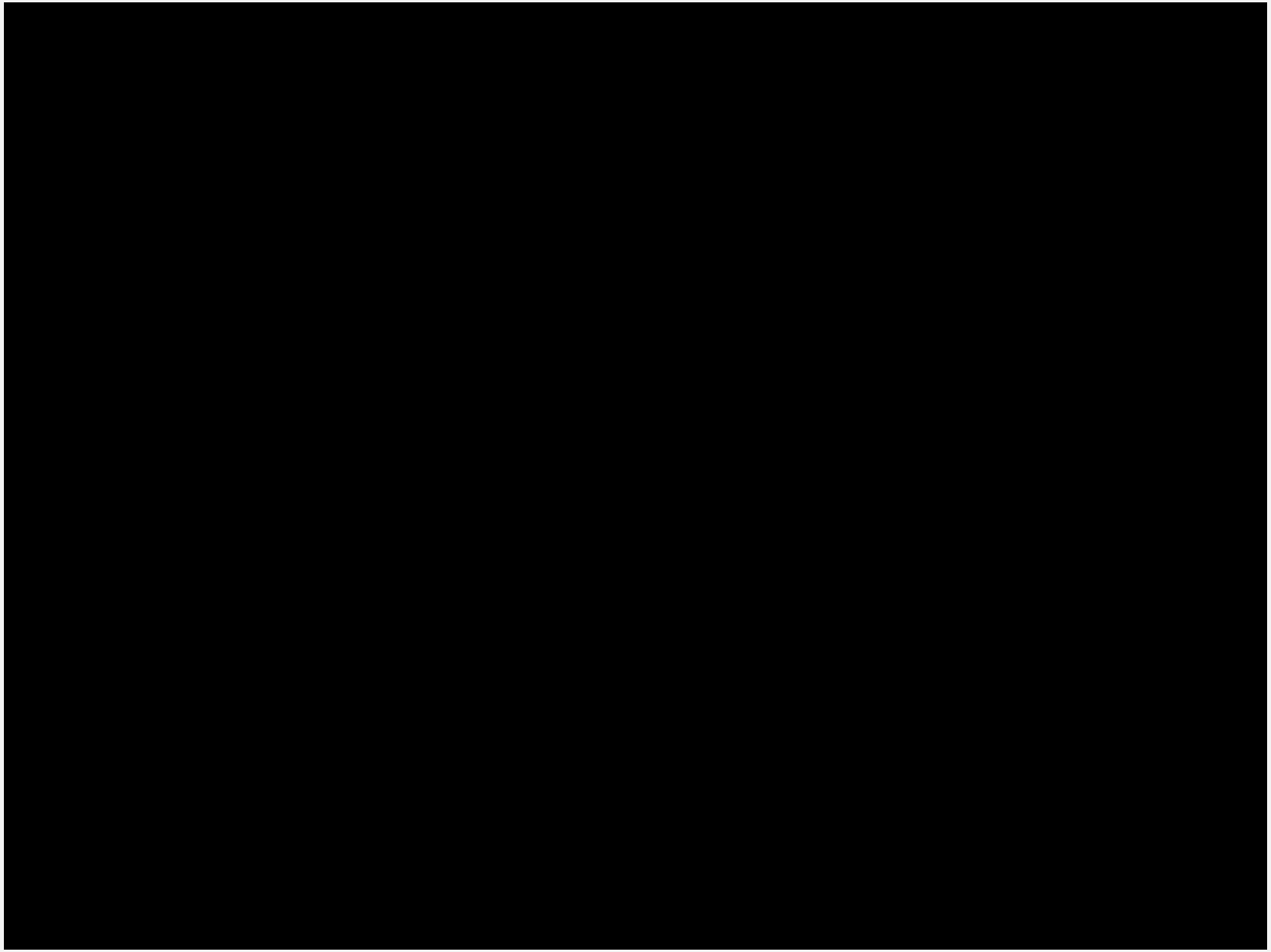
$$= 1/n\ \Sigma_i^n\ 2\ (y_i - f(x_i))\ x_i$$

Igual a δb pero con δm

Pseudocódigo

- **descenso_gradiente_rl**(x,y,α,iteraciones)
  - Inicializar m y b de forma aleatoria
  - for i in iteraciones:
    - Calcular derivadas
      - $\delta E(m,b)/\delta b = 1/n \sum_i^n 2\ (y_i - f(x_i))$
      - $\delta E(m,b)/\delta m = 1/n \sum_i^n 2\ (y_i - f(x_i))\ x_i$
    - $b = b - \alpha\ \delta E(m,b)/\delta b$
    - $m = m - \alpha\ \delta E(m,b)/\delta m$
    - $E = (1/n) \sum_i^n (y_i - m\ x_i + b)^2$
  - retornar m,b

Ejemplo con simulador

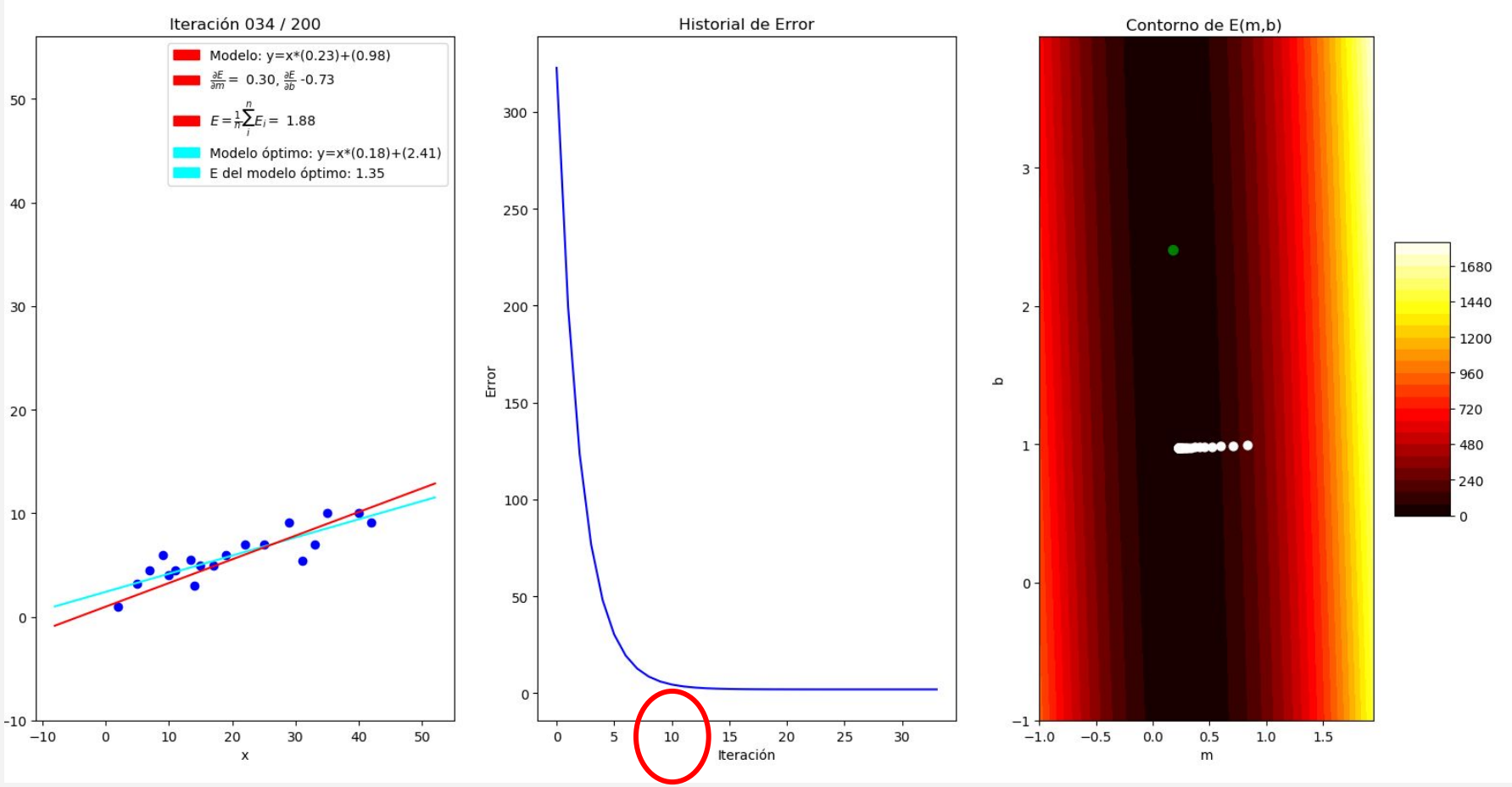


Cuestiones prácticas

- m_0 y b_0 → Valores iniciales de m y b
  - Afectan a la optimización.
  - Aprovechar la experticia del dominio.
    - Ejemplo de las notas,
      - ¿valores sensatos?
      - $y_{min} = 2$ → Nadie tiene una nota menor a 2
        - $b_0 = y_{min} = 2$
      - $x_{max} = 40$ → Nadie estudia más de 40 horas
      - $y_{max} = 10$ → Nota máxima 10
        - $m_0 = (y_{max} - y_{min})/x_{max} = (10-2)/40 = 8/40 = 0.2$
- Normalización de variables
  - Escala de notas de 0 al 10 vs 0 al 100
    - ¿Afecta al descenso?

Descenso de gradiente con **variables sin normalizar**

- Variables sin normalizar
  - Diferentes escalas
  - **Idem parámetros**
  - Dificulta encontrar el mínimo
  - Tarda más tiempo.



Normalización de variables

Original

Horas	Nota
2	1
5	3.2
7	4.5
9	6
10	4
11	4.5
13.4	5.5
14	3
15	5

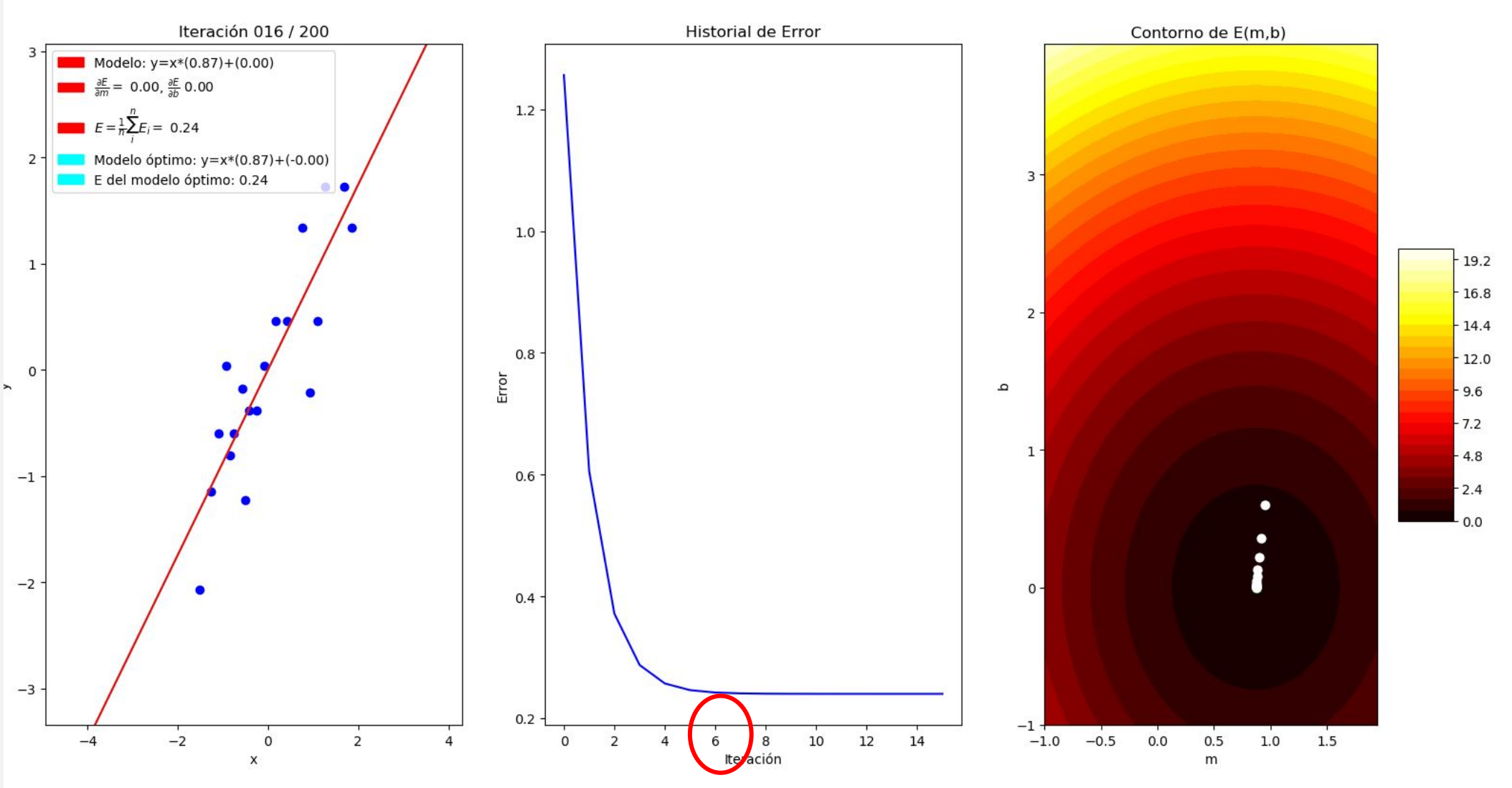
Normalización μ/σ

Horas	Nota
-1.75	-2.03
-1.06	-0.58
-0.60	0.28
-0.14	1.27
0.09	-0.05
0.32	0.28
0.87	0.94
1.01	-0.71
1.24	0.61

Normalización min/max

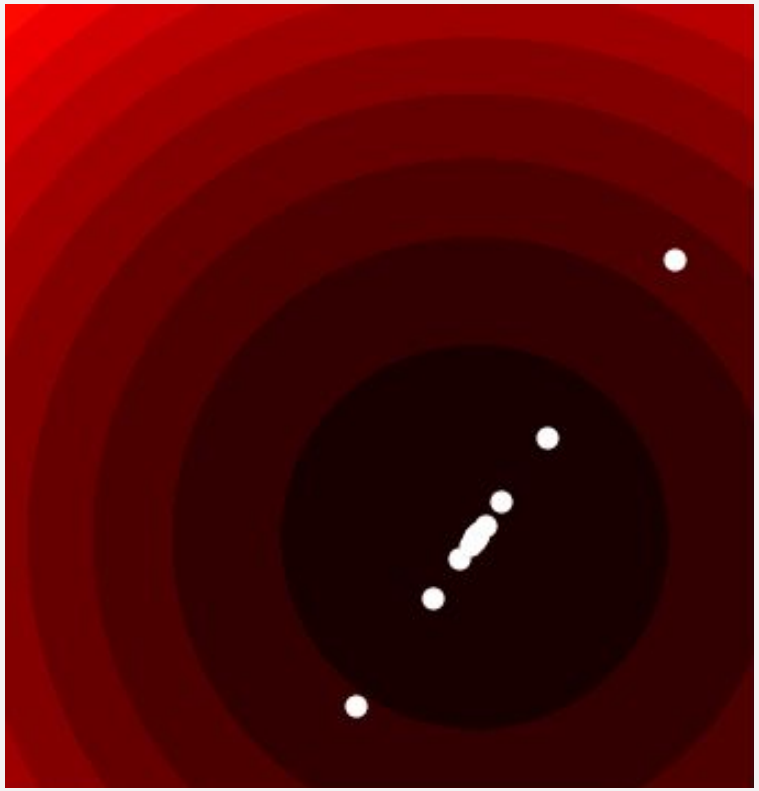
Horas	Nota
0	0
0.23	0.44
0.38	0.7
0.54	1
0.62	0.6
0.69	0.7
0.88	0.9
0.92	0.4
1.00	0.8

Descenso de gradiente con **variables normalizadas**

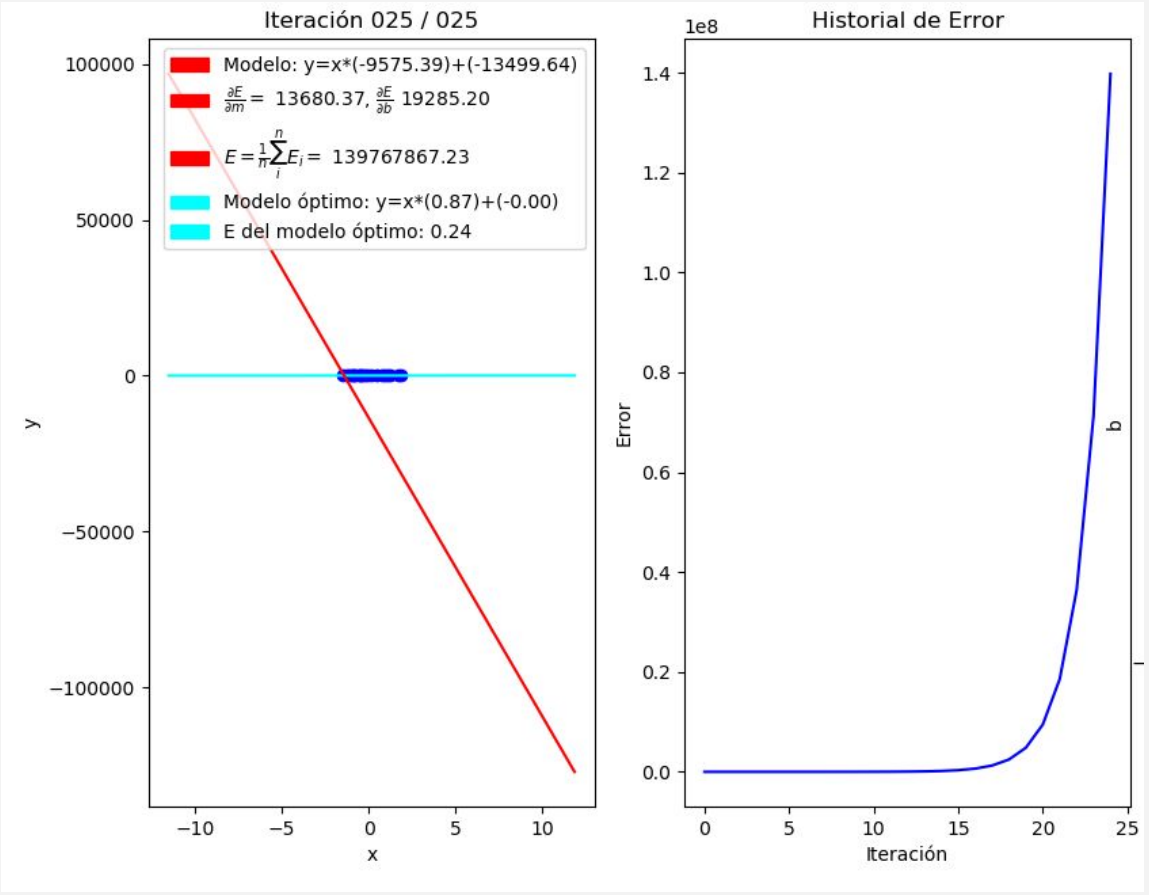


Divergencia con α muy grande

Saltos erráticos



El algoritmo diverge



Resumen

- Ecuaciones de Descenso de Gradiente
  - Especializadas para Regresión Lineal
  - $b = b - \alpha \ \delta E(m,b)/\delta b$
  - $m = m - \alpha \ \delta E(m,b)/\delta m$
    - Derivadas parciales
      - $\delta E(m,b)/\delta b = 1/n \ \Sigma_i^n \ 2 \ (y_i - f(x_i))$
      - $\delta E(m,b)/\delta m = 1/n \ \Sigma_i^n \ 2 \ (y_i - f(x_i)) \ x$
- Cuestiones prácticas
  - Valores iniciales de **m** y **b**
  - Valor de **α**
  - Normalización de las variables
    - Convergencia y velocidad

Ejercicio: Archivo **Regresión Lineal - Aprendizaje.ipynb**

- **Probar e interpretar**
  - Ejecutar el código y ver como se entrena el modelo
  - Cambiar los valores iniciales de **m** y **b**
    - ¿Cómo afecta esto al entrenamiento?
  - Cambiar el valor de **α**
    - ¿Cómo afecta esto al entrenamiento?
    - ¿Qué sucede si utilizo un α muy chico?
      - ¿y uno muy grande?
  - ¿Son comparables los valores $\delta E/\delta m$ y $\delta E/\delta b$?
    - ¿De qué depende su magnitud?
  - Cambiar el conjunto de datos utilizado por **anscombe1.csv**, **anscombe2.csv**, **anscombe3.csv** o **anscombe4.csv** y repetir.