



UTEC Posgrado



MAESTRÍA

Linear Regression

OLS, Bayesian regression. Learning theory.



Modelado

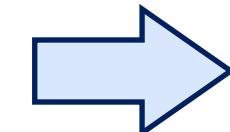
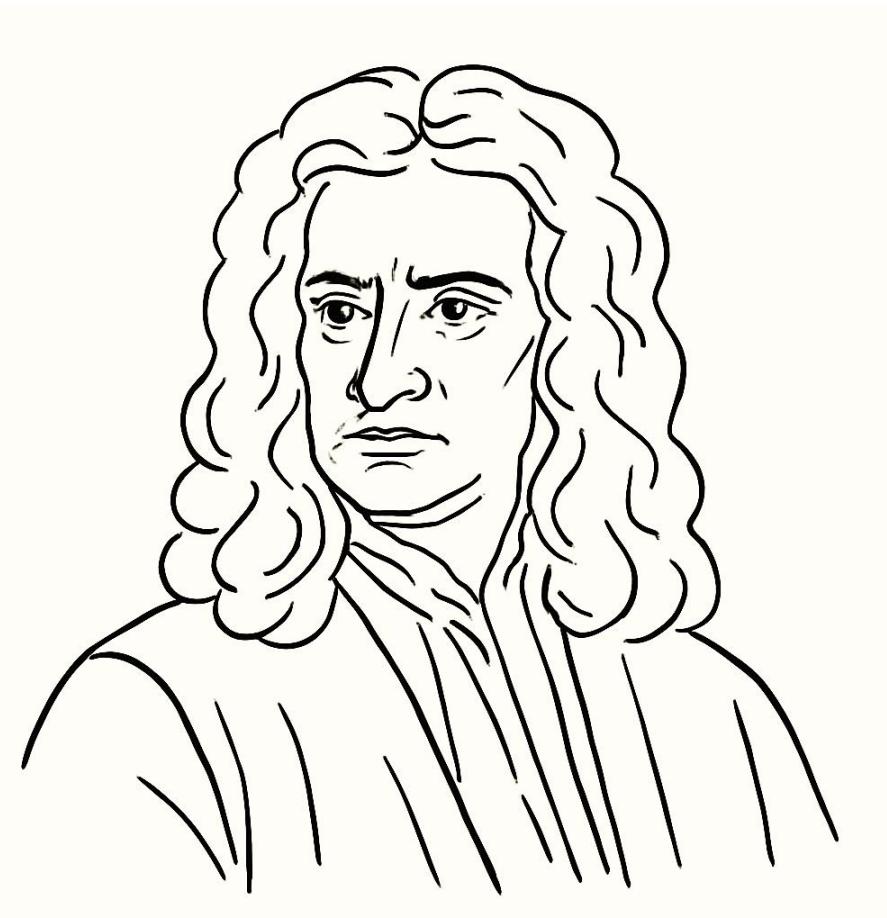
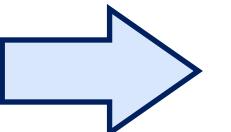
Modelo lineal



Modeling



Mundo real

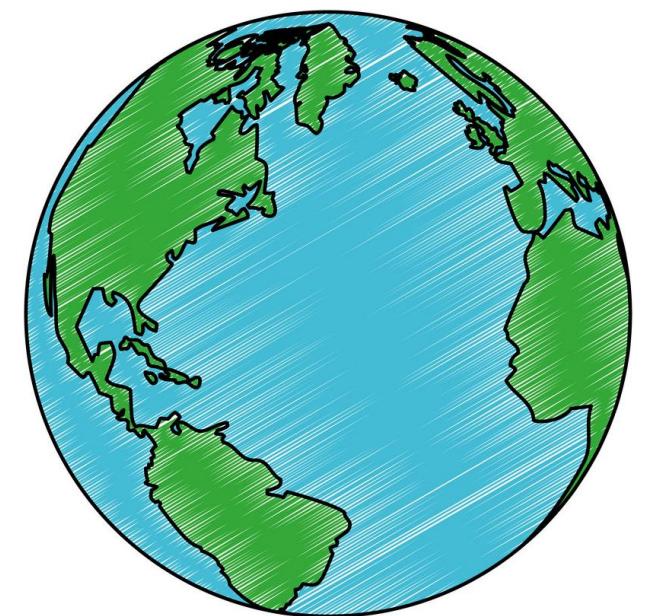


$$\vec{F} = \frac{\partial \vec{p}}{\partial t} = m \frac{\partial \vec{v}}{\partial t}$$
$$F = G \frac{m_1 m_2}{r^2}$$

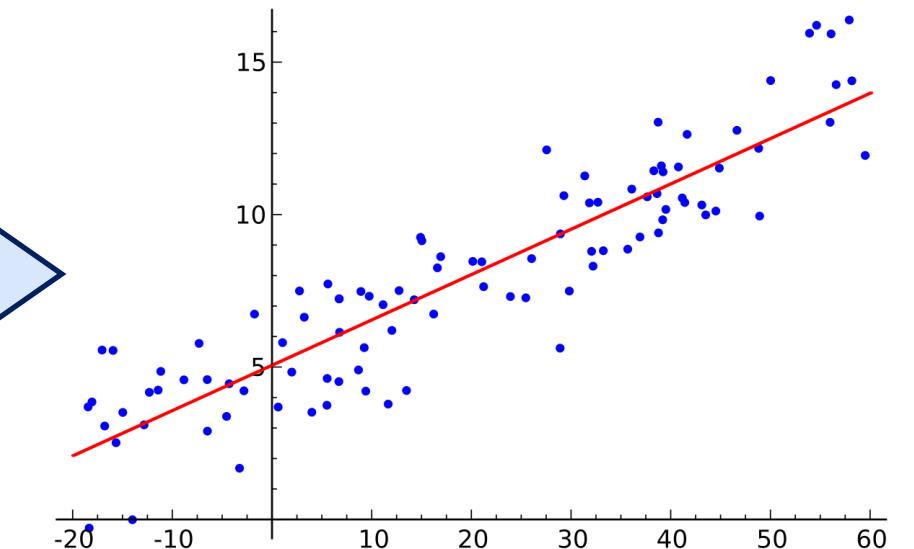
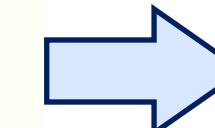
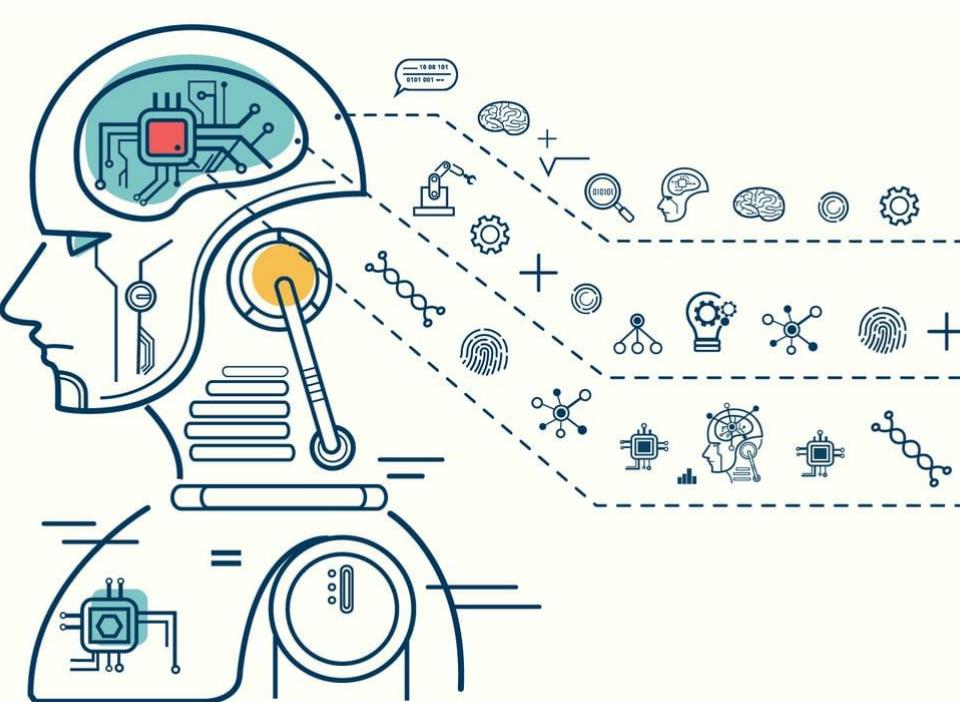
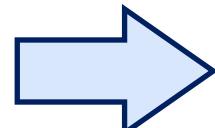
Leyes físicas



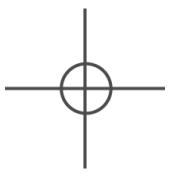
Modeling



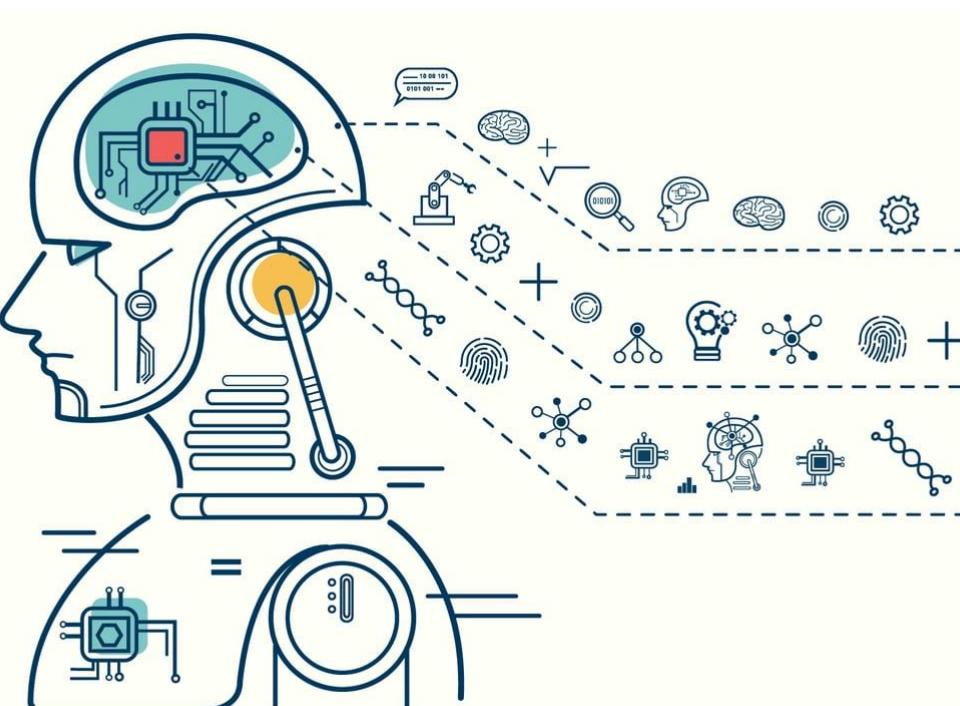
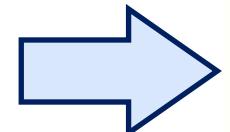
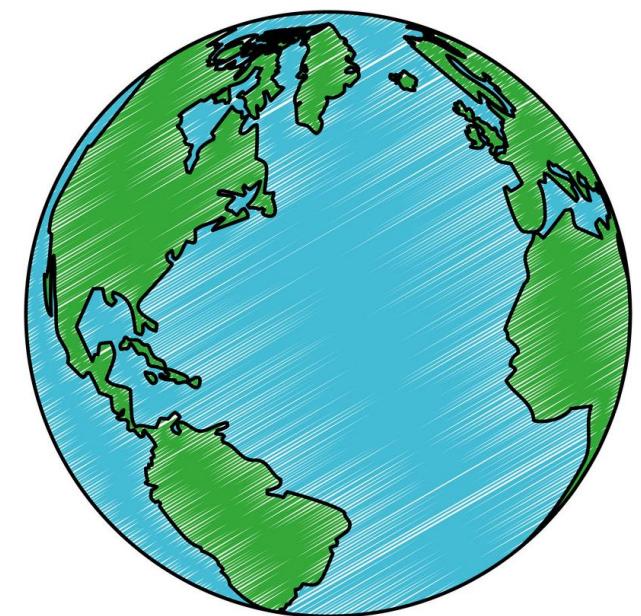
Mundo real



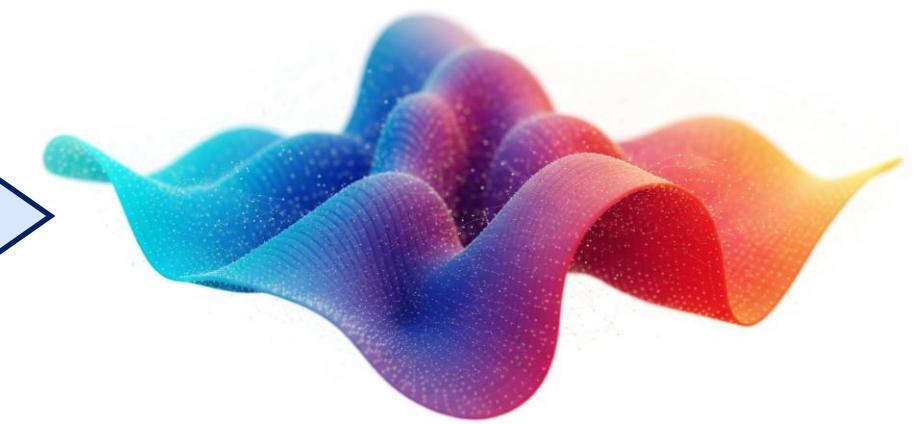
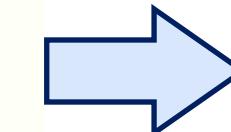
Predicciones



Modeling



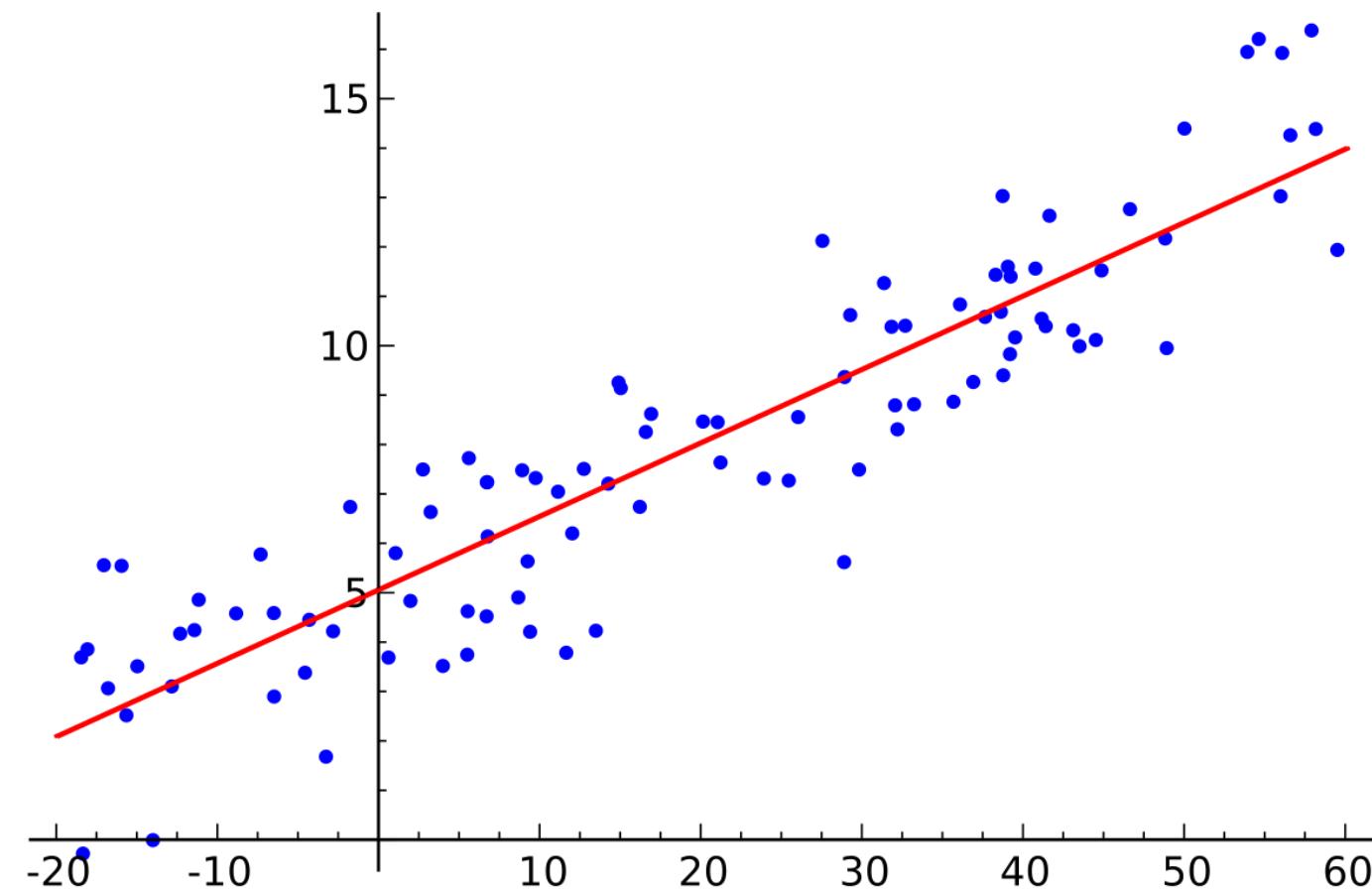
Mundo real



Representaciones



Regresión lineal



$$y_i = \beta_0 + \sum_{j=1}^p x_{ij} \beta_j + \varepsilon_i$$

number of features → p

response → y_i

global intercept → β_0

feature j of observation i → x_{ij}

coefficient for feature j → β_j

noise term → $\varepsilon_i \stackrel{iid}{\sim} \mathbf{N}(0, \sigma^2)$

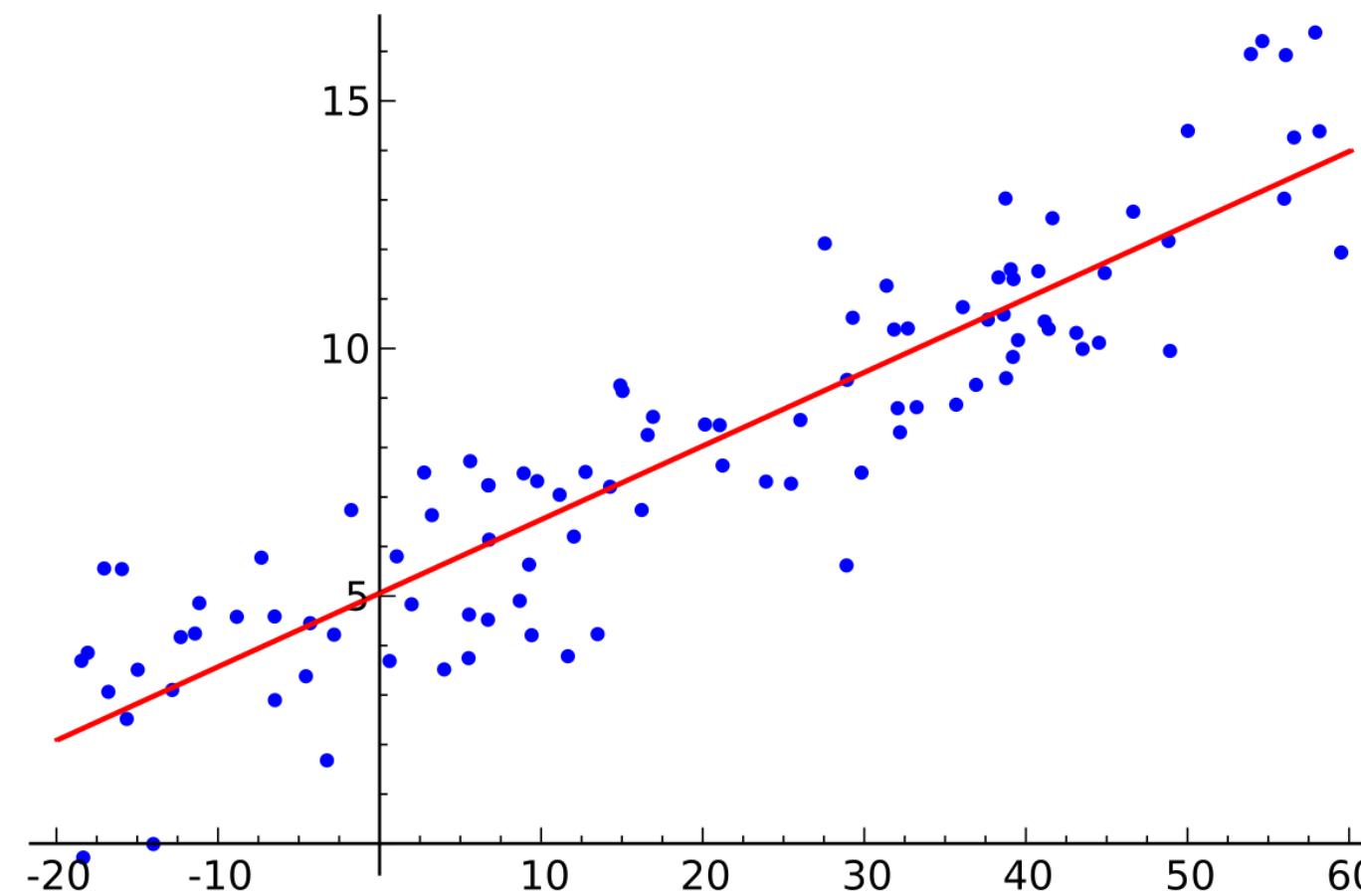
independence assumption → $\varepsilon_i \stackrel{iid}{\sim} \mathbf{N}(0, \sigma^2)$

noise level → σ^2



Ordinary Least-Squares Estimator

El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{y} = Xw$ y los valores reales y .



$$J(w) = \frac{1}{2} \|Xw - y\|^2 \quad \text{Mean squared error (MSE)}$$

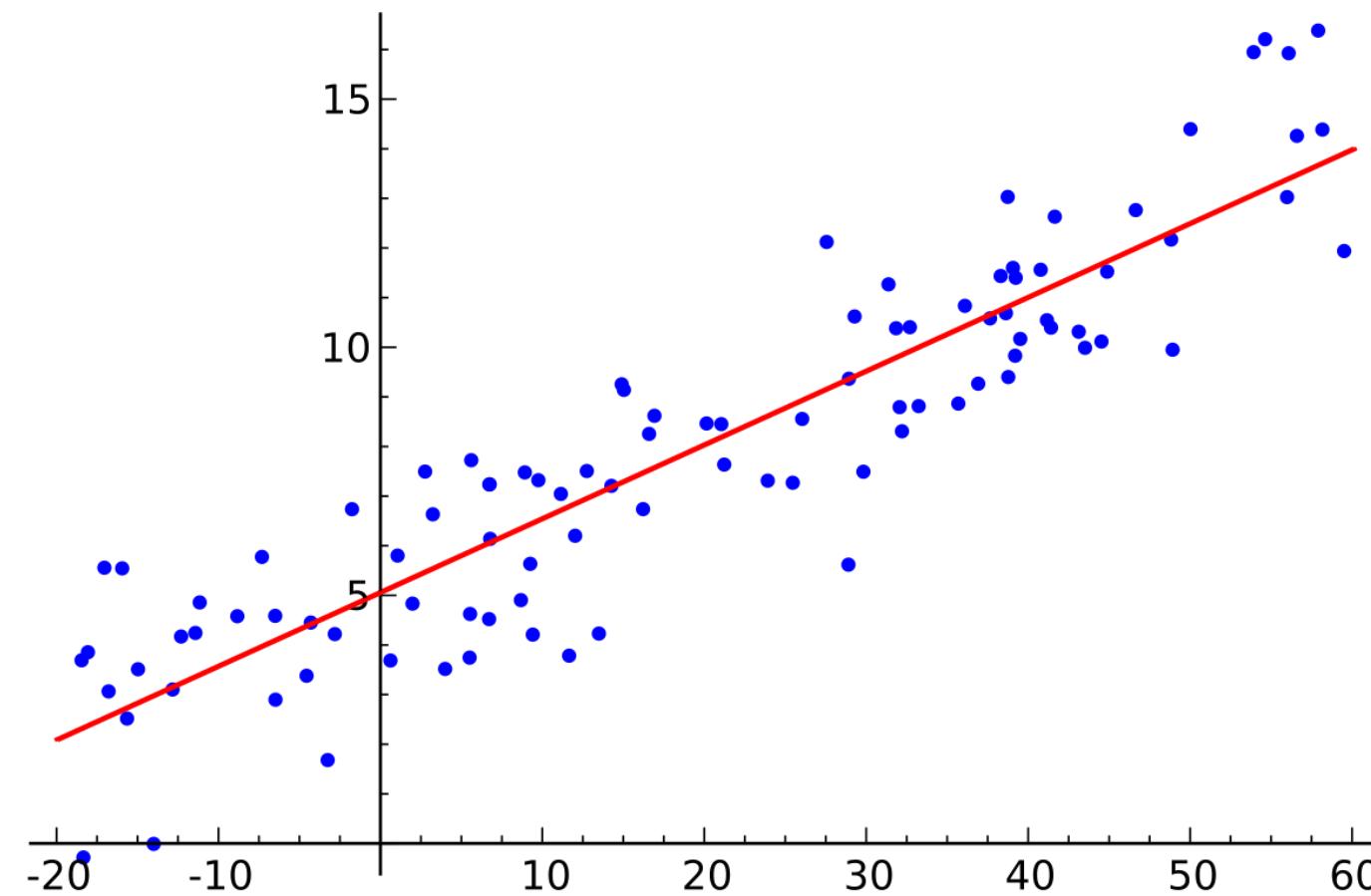
- m : número de muestras
- n : número de features

- $X \in \mathbb{R}^{m \times n}$
- $w \in \mathbb{R}^{n \times 1}$
- $y \in \mathbb{R}^{m \times 1}$



Ordinary Least-Squares Estimator

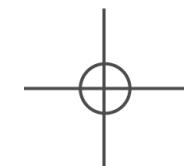
El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ y los valores reales \mathbf{y} .



$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \frac{1}{2} [\mathbf{w}^T \mathbf{X}^T \mathbf{X}\mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}]$$

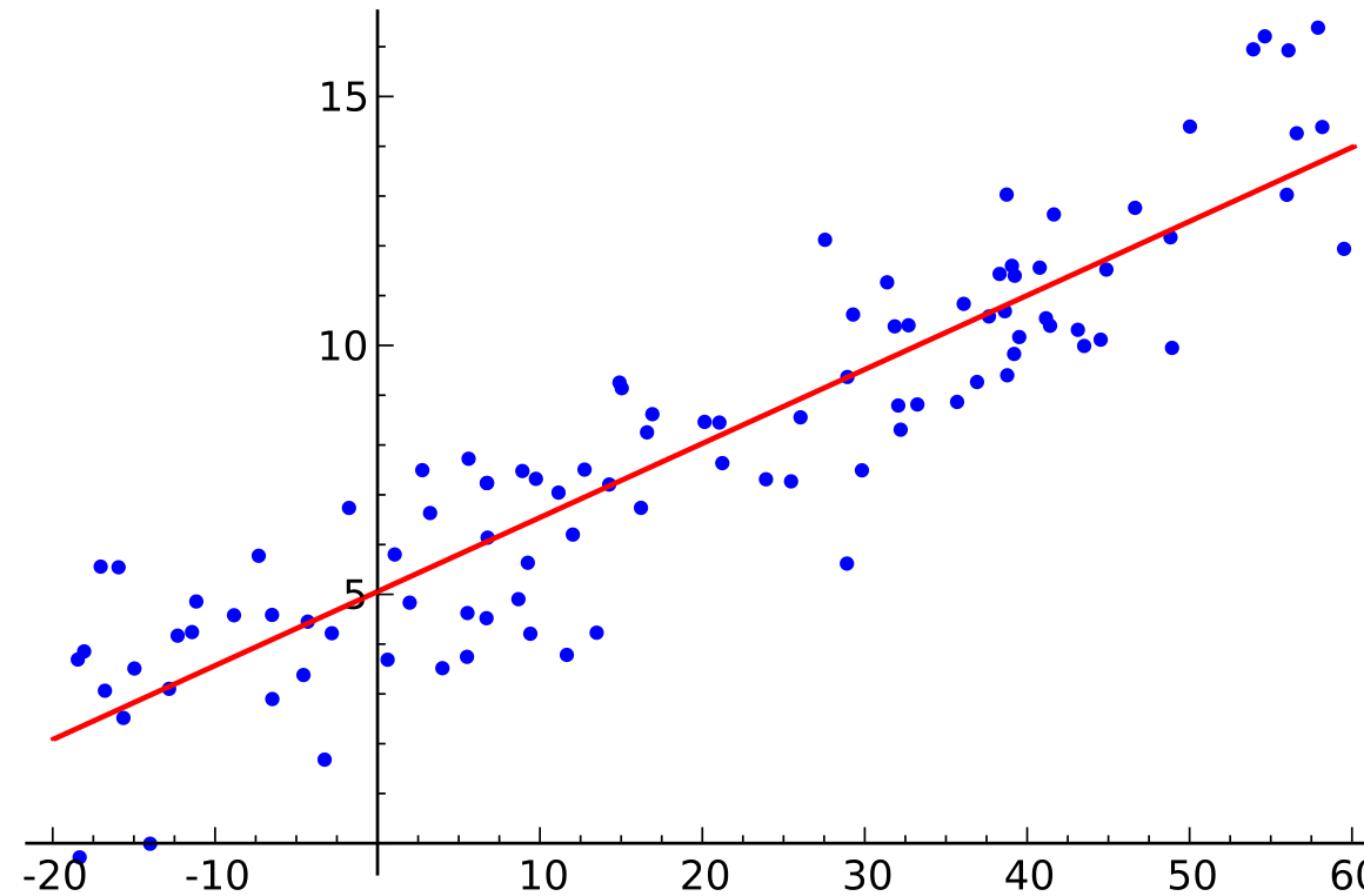
- m : número de muestras
- n : número de features

- $\mathbf{X} \in \mathbb{R}^{m \times n}$
- $\mathbf{w} \in \mathbb{R}^{n \times 1}$
- $\mathbf{y} \in \mathbb{R}^{m \times 1}$



Ordinary Least-Squares Estimator

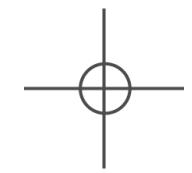
El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ y los valores reales \mathbf{y} .



$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \frac{1}{2} [\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}]$$

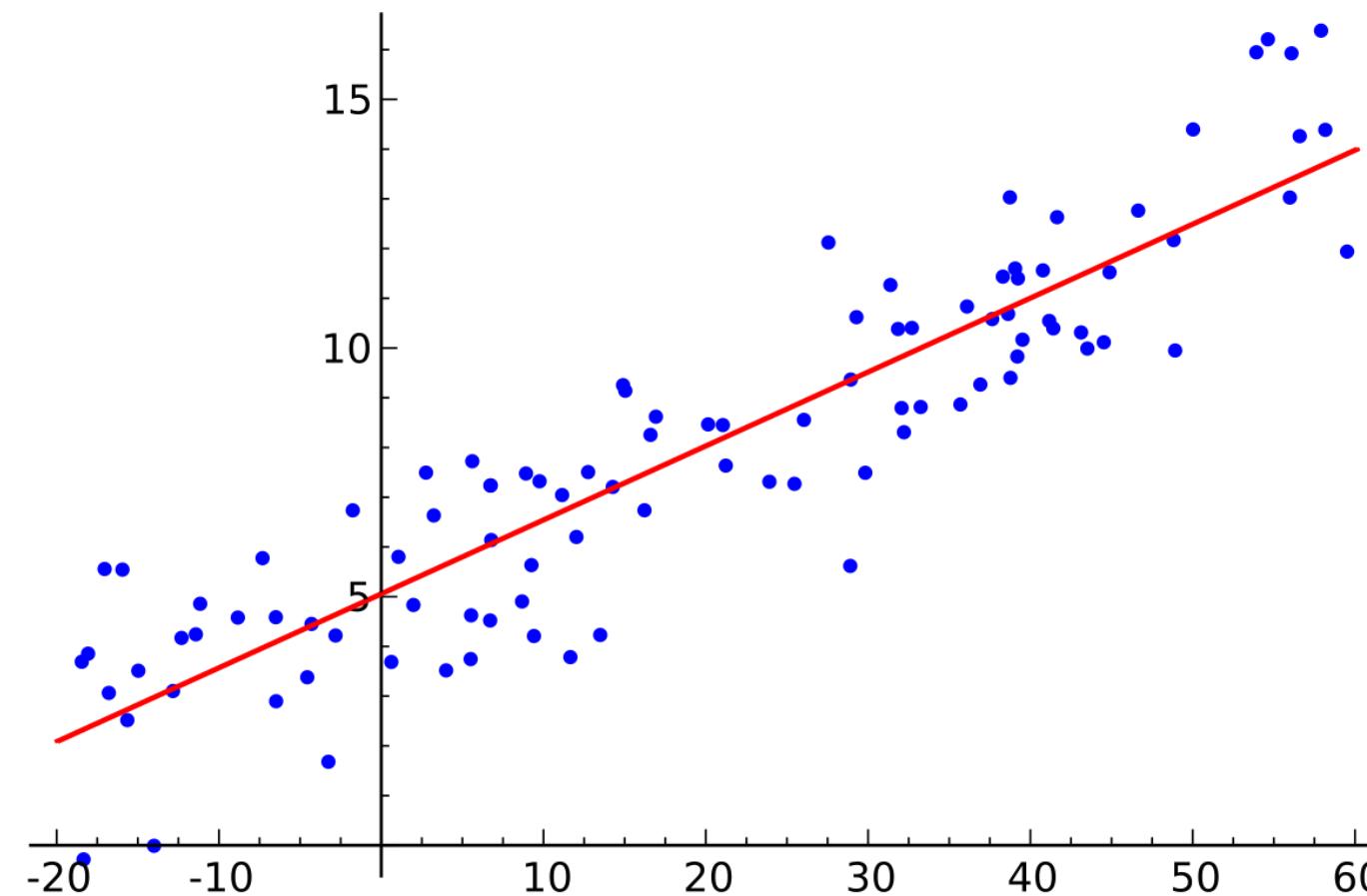
Derivando respecto a \mathbf{w} :

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right) + \frac{\partial}{\partial \mathbf{w}} (-\mathbf{w}^T \mathbf{X}^T \mathbf{y}) + \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{y}^T \mathbf{y} \right)$$



Ordinary Least-Squares Estimator

El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ y los valores reales \mathbf{y} .



$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \frac{1}{2} [\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}]$$

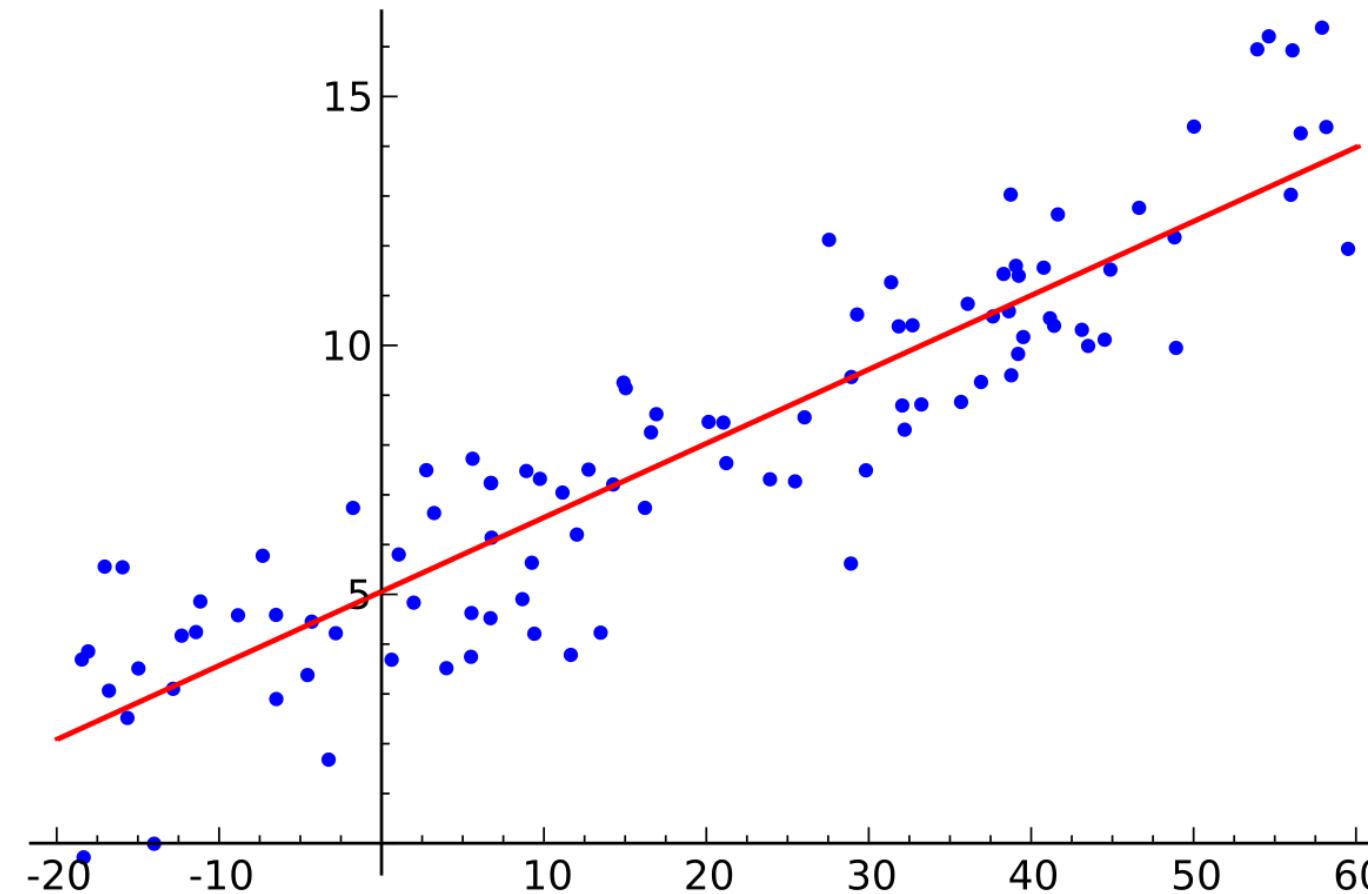
Derivando respecto a \mathbf{w} :

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right) + \frac{\partial}{\partial \mathbf{w}} (-\mathbf{w}^T \mathbf{X}^T \mathbf{y}) + \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{y}^T \mathbf{y} \right) = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y}$$



Ordinary Least-Squares Estimator

El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ y los valores reales \mathbf{y} .



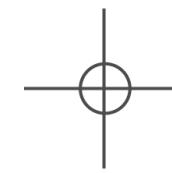
$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \frac{1}{2} [\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}]$$

Derivando respecto a \mathbf{w} :

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right) + \frac{\partial}{\partial \mathbf{w}} (-\mathbf{w}^T \mathbf{X}^T \mathbf{y}) + \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{y}^T \mathbf{y} \right) = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y}$$

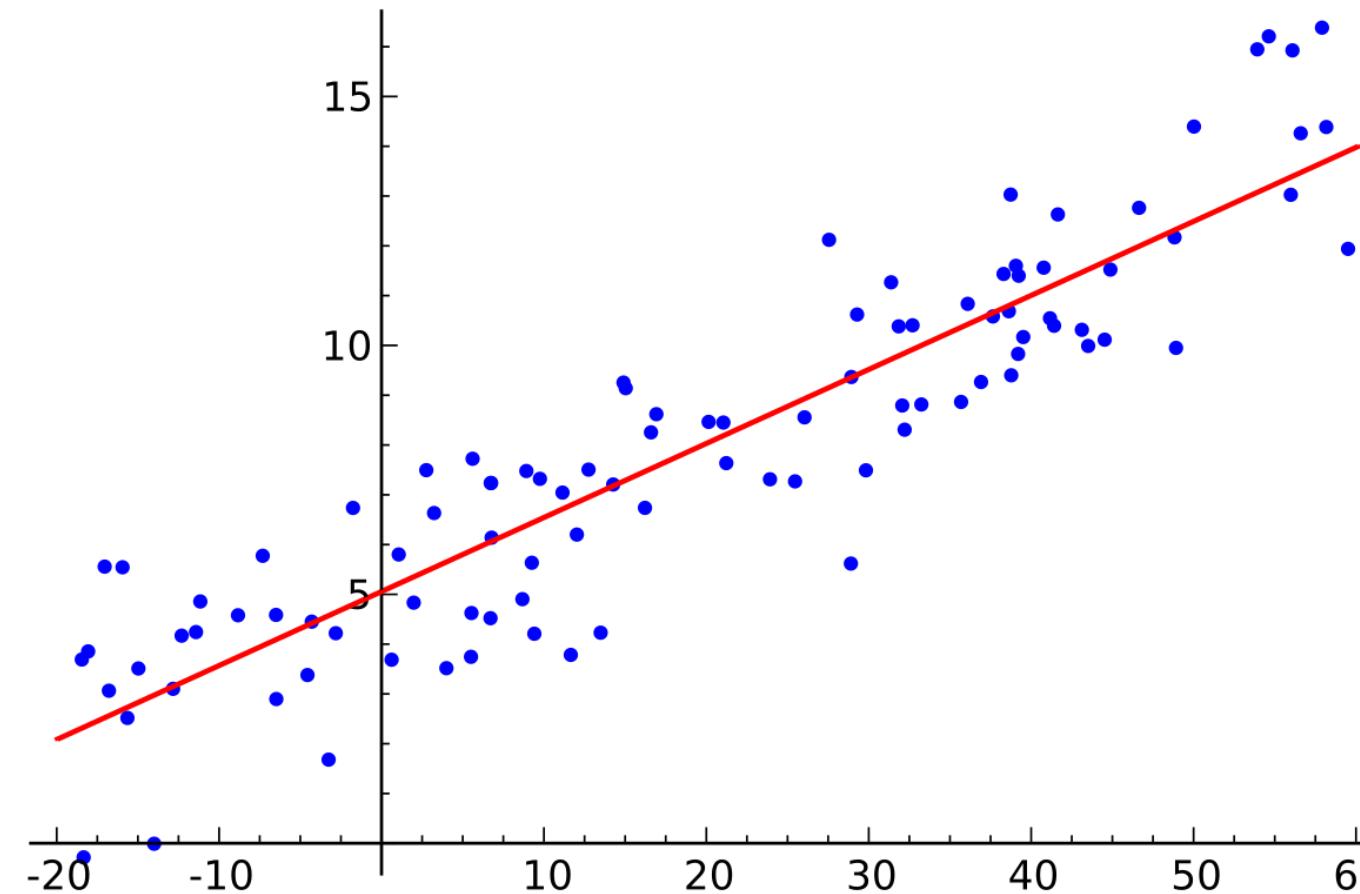
Buscamos minimizar $J(\mathbf{w})$, entonces igualamos la derivada a cero:

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y} = 0$$



Ordinary Least-Squares Estimator

El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ y los valores reales \mathbf{y} .



$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \frac{1}{2} [\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}]$$

Derivando respecto a \mathbf{w} :

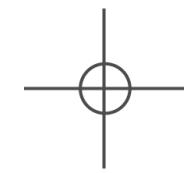
$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right) + \frac{\partial}{\partial \mathbf{w}} (-\mathbf{w}^T \mathbf{X}^T \mathbf{y}) + \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{y}^T \mathbf{y} \right) = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y}$$

Buscamos minimizar $J(\mathbf{w})$, entonces igualamos la derivada a cero:

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y} = 0$$

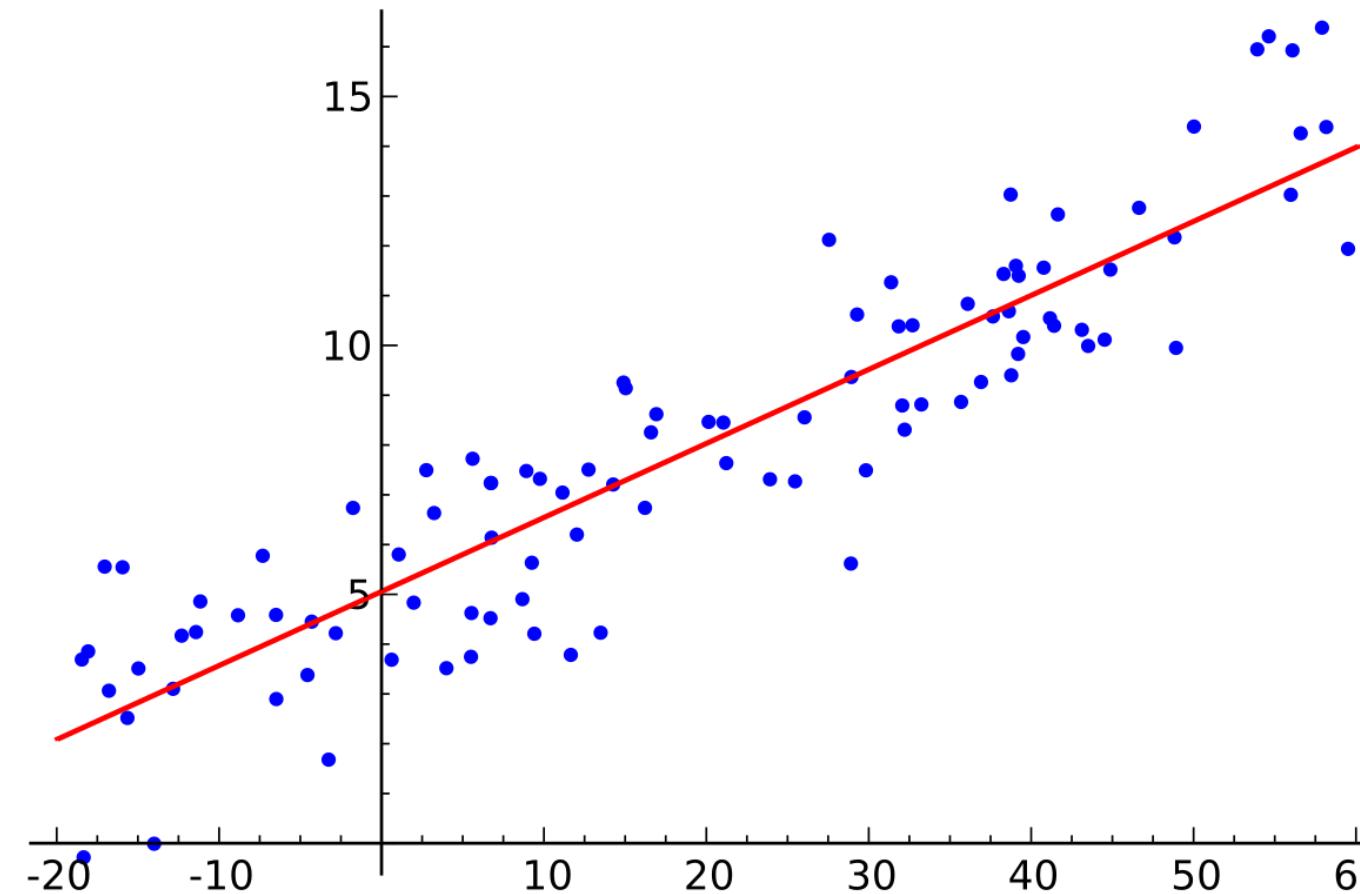
Reorganizando:

$$\mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{y}$$



Ordinary Least-Squares Estimator

El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}$ y los valores reales \mathbf{y} .



$$J(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \frac{1}{2} (\mathbf{X}\mathbf{w} - \mathbf{y})^T (\mathbf{X}\mathbf{w} - \mathbf{y}) = \frac{1}{2} [\mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} - 2\mathbf{w}^T \mathbf{X}^T \mathbf{y} + \mathbf{y}^T \mathbf{y}]$$

Derivando respecto a \mathbf{w} :

$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{X}^T \mathbf{X} \mathbf{w} \right) + \frac{\partial}{\partial \mathbf{w}} (-\mathbf{w}^T \mathbf{X}^T \mathbf{y}) + \frac{\partial}{\partial \mathbf{w}} \left(\frac{1}{2} \mathbf{y}^T \mathbf{y} \right) = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y}$$

Buscamos minimizar $J(\mathbf{w})$, entonces igualamos la derivada a cero:

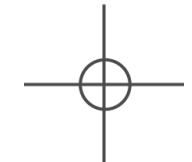
$$\frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{X}^T \mathbf{X} \mathbf{w} - \mathbf{X}^T \mathbf{y} = 0$$

Reorganizando:

$$\mathbf{X}^T \mathbf{X} \mathbf{w} = \mathbf{X}^T \mathbf{y}$$

Despejamos \mathbf{w} :

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



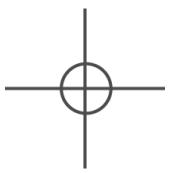
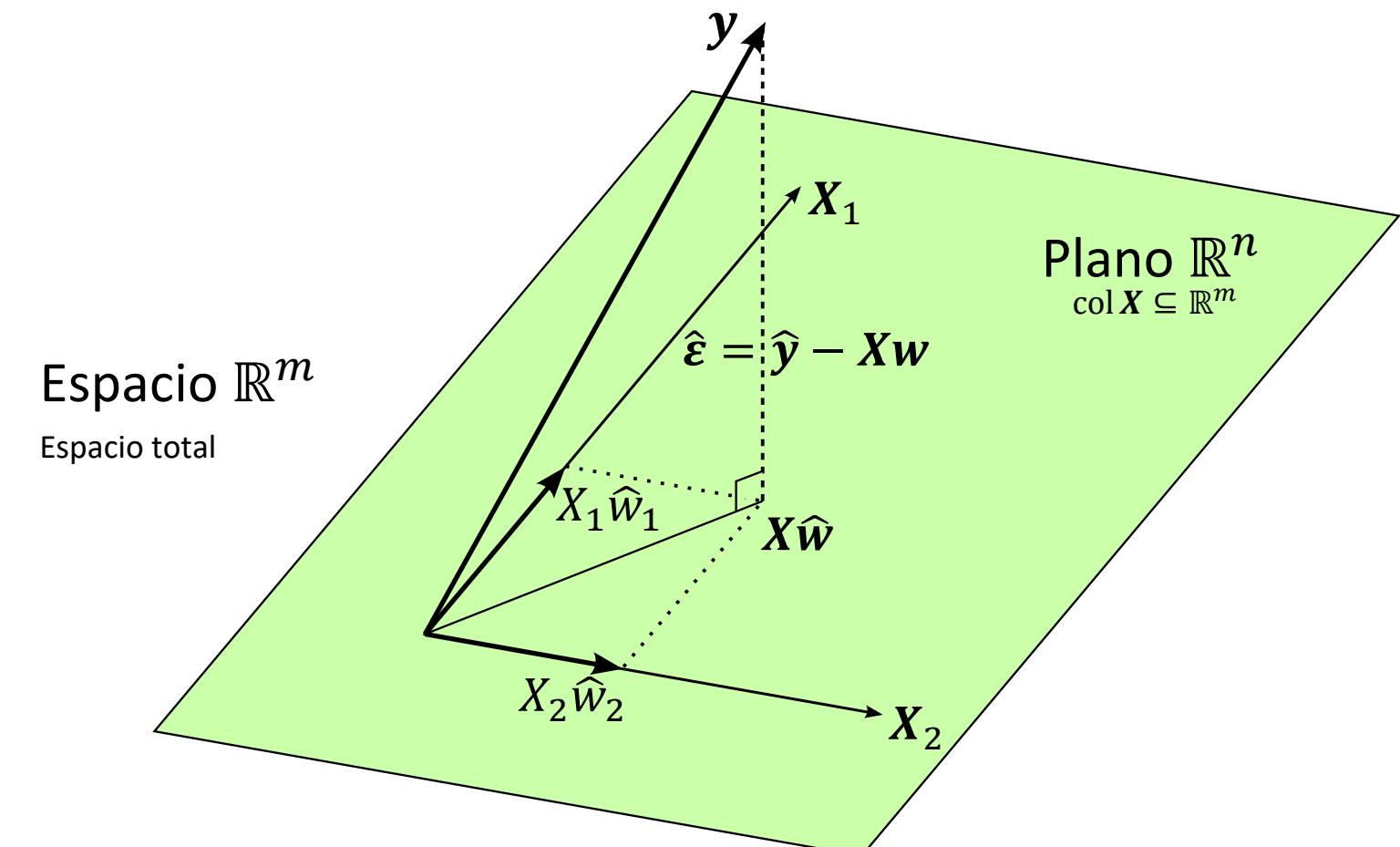
Ordinary Least-Squares Estimator

El objetivo de OLS es minimizar la suma de los errores al cuadrado entre las predicciones $\hat{y} = Xw$ y los valores reales y .

Solución analítica:

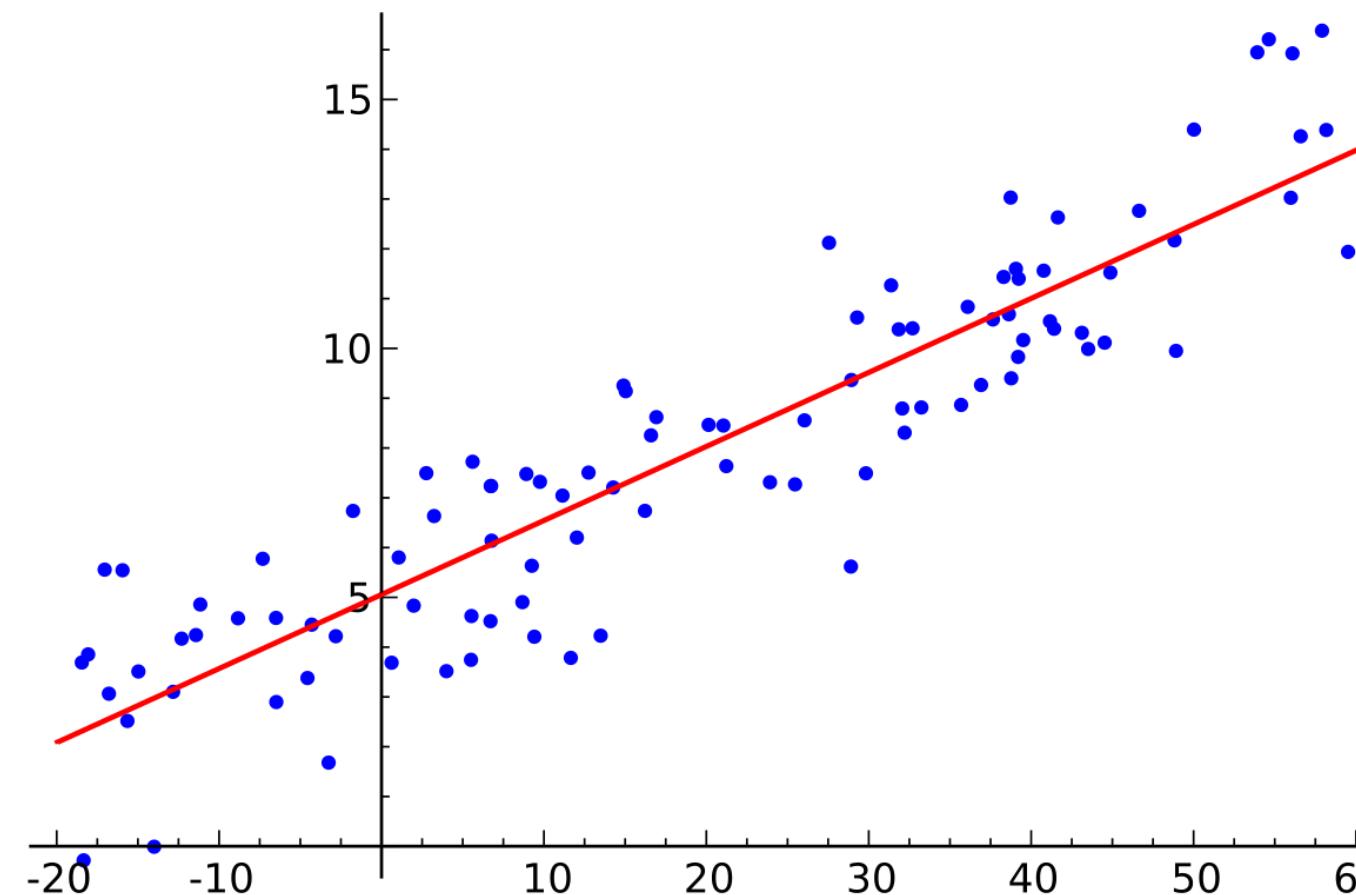
$$w = (X^T X)^{-1} X^T y$$

- m : número de muestras
- n : número de features
- $X \in \mathbb{R}^{m \times n}$
- $w \in \mathbb{R}^{n \times 1}$
- $y \in \mathbb{R}^{m \times 1}$



Ordinary Least-Squares Estimator

Considerando el modelo lineal $\mathbf{y} = \mathbf{X}\mathbf{w}^* + \epsilon$, donde \mathbf{w}^* son los verdaderos parámetros y ϵ es el ruido estocástico.
El estimador OLS se define como $\hat{\mathbf{w}}_{OLS} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.



Exogeneidad Estricta $\mathbb{E}[\epsilon | \mathbf{X}] = 0$

El error esperado es cero para cualquier valor de \mathbf{X} . El ruido no tiene patrones correlacionados con nuestras variables.

Garantiza que el estimador sea Insesgado ($\mathbb{E}[\hat{\mathbf{w}}] = \mathbf{w}^*$). No subestimamos ni sobreestimamos sistemáticamente.

Homoscedasticidad y No-Autocorrelación $\text{Var}[\epsilon | \mathbf{X}] = \sigma^2 \mathbb{I}_m$

La varianza del error es constante σ^2 para todas las observaciones (homoscedasticidad) y los errores de diferentes muestras no están correlacionados entre sí (matriz diagonal).

Es crucial para la eficiencia del estimador.

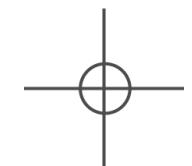
Rango Completo $\text{rank}(\mathbf{X}) = n$

No hay multicolinealidad perfecta entre los features.

Asegura que la matriz $\mathbf{X}^T \mathbf{X}$ sea invertible y la solución sea única.

Bajo estos tres supuestos, el estimador OLS es el Best Linear Unbiased Estimator (BLUE).

Esto significa que, de todos los posibles estimadores lineales e insesgados que existen, OLS es el que tiene la menor varianza (es el más preciso).



Ordinary Least-Squares Estimator

El estimador OLS proyecta ortogonalmente el vector objetivo y sobre el espacio generado por las columnas de X (Col X).

Predicción: $\hat{y} = X\hat{w} = \mathcal{H}y$, donde $\mathcal{H} = X(X^T X)^{-1}X^T$ es la matriz de proyección (Hat Matrix).

El vector de residuos \hat{e} es ortogonal a cualquier característica: $X^T \hat{e} = 0$. Esto implica que hemos extraído toda la información linealmente posible de X .

Incertidumbre en los Parámetros (Varianza):

Bajo los supuestos de Gauss-Markov, la covarianza del estimador es: $\text{Var}[\hat{w} | X] = \sigma^2 (X^T X)^{-1}$

La precisión de \hat{w} depende de la fuerza de la señal ($X^T X$) y del ruido intrínseco σ^2 .

Problema: Si las columnas de X son colineales, los autovalores de $(X^T X)$ son cercanos a 0, haciendo que $(X^T X)^{-1}$ explote. Esto genera una alta varianza en los pesos (inestabilidad numérica).



Ordinary Least-Squares Estimator

El estimador OLS proyecta ortogonalmente el vector objetivo \mathbf{y} sobre el espacio generado por las columnas de \mathbf{X} (Col \mathbf{X}).

Predicción: $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathcal{H}\mathbf{y}$, donde $\mathcal{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ es la matriz de proyección (Hat Matrix).

El vector de residuos $\hat{\mathbf{e}}$ es ortogonal a cualquier característica: $\mathbf{X}^\top\hat{\mathbf{e}} = 0$. Esto implica que hemos extraído toda la información linealmente posible de \mathbf{X} .

Incertidumbre en los Parámetros (Varianza):

Bajo los supuestos de Gauss-Markov, la covarianza del estimador es: $\text{Var}[\hat{\mathbf{w}} | \mathbf{X}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$

La precisión de $\hat{\mathbf{w}}$ depende de la fuerza de la señal ($\mathbf{X}^\top\mathbf{X}$) y del ruido intrínseco σ^2 .

Problema: Si las columnas de \mathbf{X} son colineales, los autovalores de $(\mathbf{X}^\top\mathbf{X})$ son cercanos a 0, haciendo que $(\mathbf{X}^\top\mathbf{X})^{-1}$ explote. Esto genera una alta varianza en los pesos (inestabilidad numérica).

Estimación de la Escala del Ruido:

Dado que σ^2 es desconocido, usamos un estimador insesgado basado en la Suma de Errores al Cuadrado (RSS): $\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2}{m - n}$

- **Grados de Libertad ($m - n$):** Penalizamos por los n parámetros ya estimados para evitar subestimar el ruido (overfitting).
- **Error Estándar:** $\text{SE}(\hat{w}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^\top\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^{-1}]_{jj}}$ Fundamental para pruebas de hipótesis (t -tests) sobre la relevancia de cada feature.



Ordinary Least-Squares Estimator

El estimador OLS proyecta ortogonalmente el vector objetivo \mathbf{y} sobre el espacio generado por las columnas de \mathbf{X} (Col \mathbf{X}).

Predicción: $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = \mathcal{H}\mathbf{y}$, donde $\mathcal{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$ es la matriz de proyección (Hat Matrix).

El vector de residuos $\hat{\mathbf{e}}$ es ortogonal a cualquier característica: $\mathbf{X}^\top\hat{\mathbf{e}} = 0$. Esto implica que hemos extraído toda la información linealmente posible de \mathbf{X} .

Incertidumbre en los Parámetros (Varianza):

Bajo los supuestos de Gauss-Markov, la covarianza del estimador es: $\text{Var}[\hat{\mathbf{w}} | \mathbf{X}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$

La precisión de $\hat{\mathbf{w}}$ depende de la fuerza de la señal ($\mathbf{X}^\top\mathbf{X}$) y del ruido intrínseco σ^2 .

Problema: Si las columnas de \mathbf{X} son colineales, los autovalores de $(\mathbf{X}^\top\mathbf{X})$ son cercanos a 0, haciendo que $(\mathbf{X}^\top\mathbf{X})^{-1}$ explote. Esto genera una alta varianza en los pesos (inestabilidad numérica).

Estimación de la Escala del Ruido:

Dado que σ^2 es desconocido, usamos un estimador insesgado basado en la Suma de Errores al Cuadrado (RSS): $\hat{\sigma}^2 = \frac{\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|^2}{m - n}$

- **Grados de Libertad ($m - n$):** Penalizamos por los n parámetros ya estimados para evitar subestimar el ruido (overfitting).
- **Error Estándar:** $\text{SE}(\hat{w}_j) = \sqrt{\hat{\sigma}^2[(\mathbf{X}^\top\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^{-1}]_{jj}}$ Fundamental para pruebas de hipótesis (t -tests) sobre la relevancia de cada feature.

Diagnóstico de Residuos ($\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}$):

Los residuos son el sistema de alerta temprana de una mala especificación del modelo:

- **Patrones no aleatorios (vs. $\hat{\mathbf{y}}$):** Indican que el modelo es insuficiente (ej. falta un término cuadrático o interacción).
- **Heterocedasticidad (Embudo):** La varianza del error crece con y . Viola Gauss-Markov. Solución: Transformación logarítmica o Weighted Least Squares.
- **No-Normalidad:** Afecta la validez de los intervalos de confianza (aunque no necesariamente la predicción puntual).

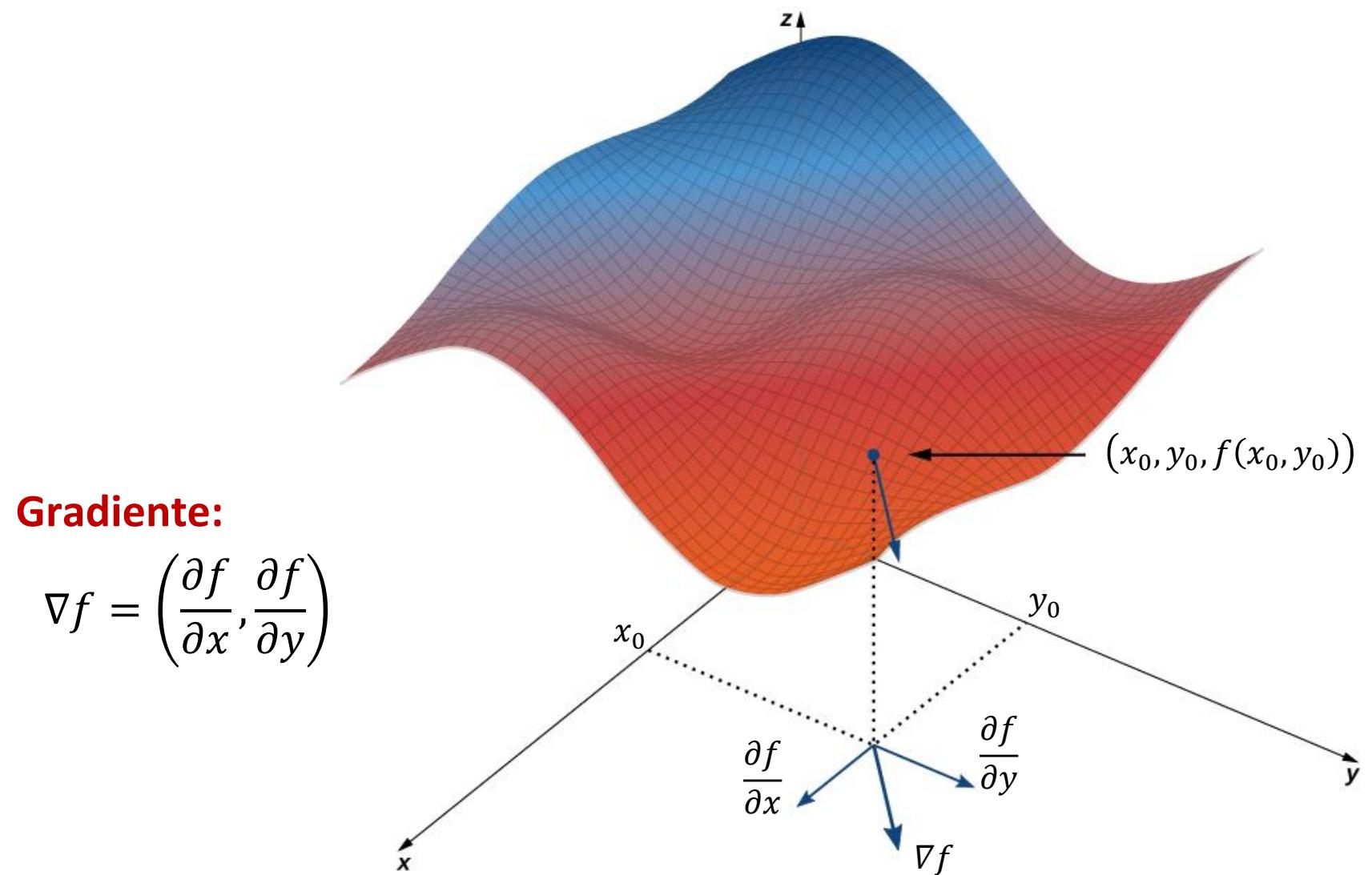


Gradient Descent

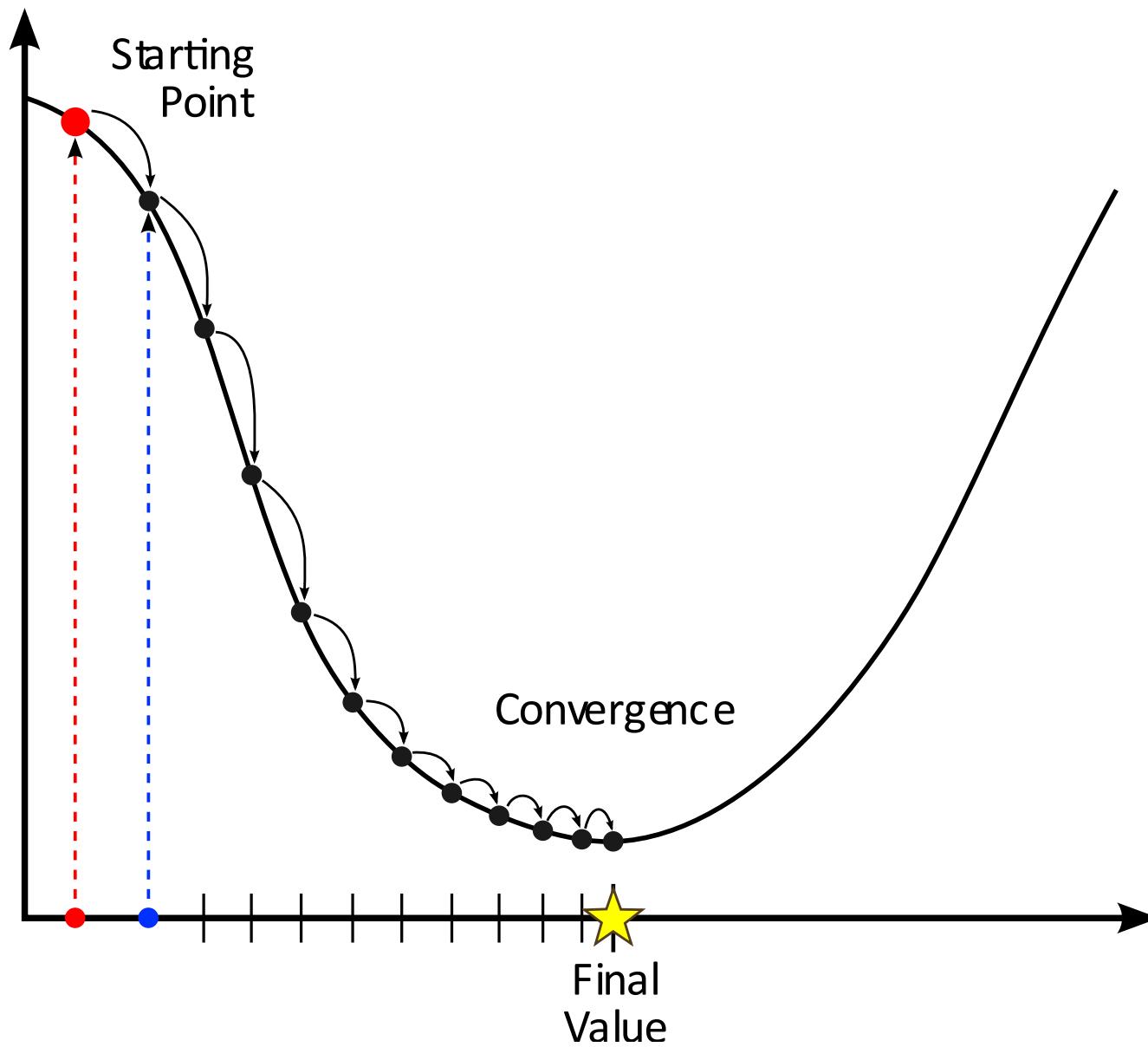
Método iterativo



Gradient descent

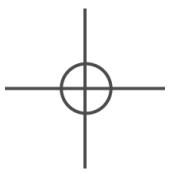


Gradient descent

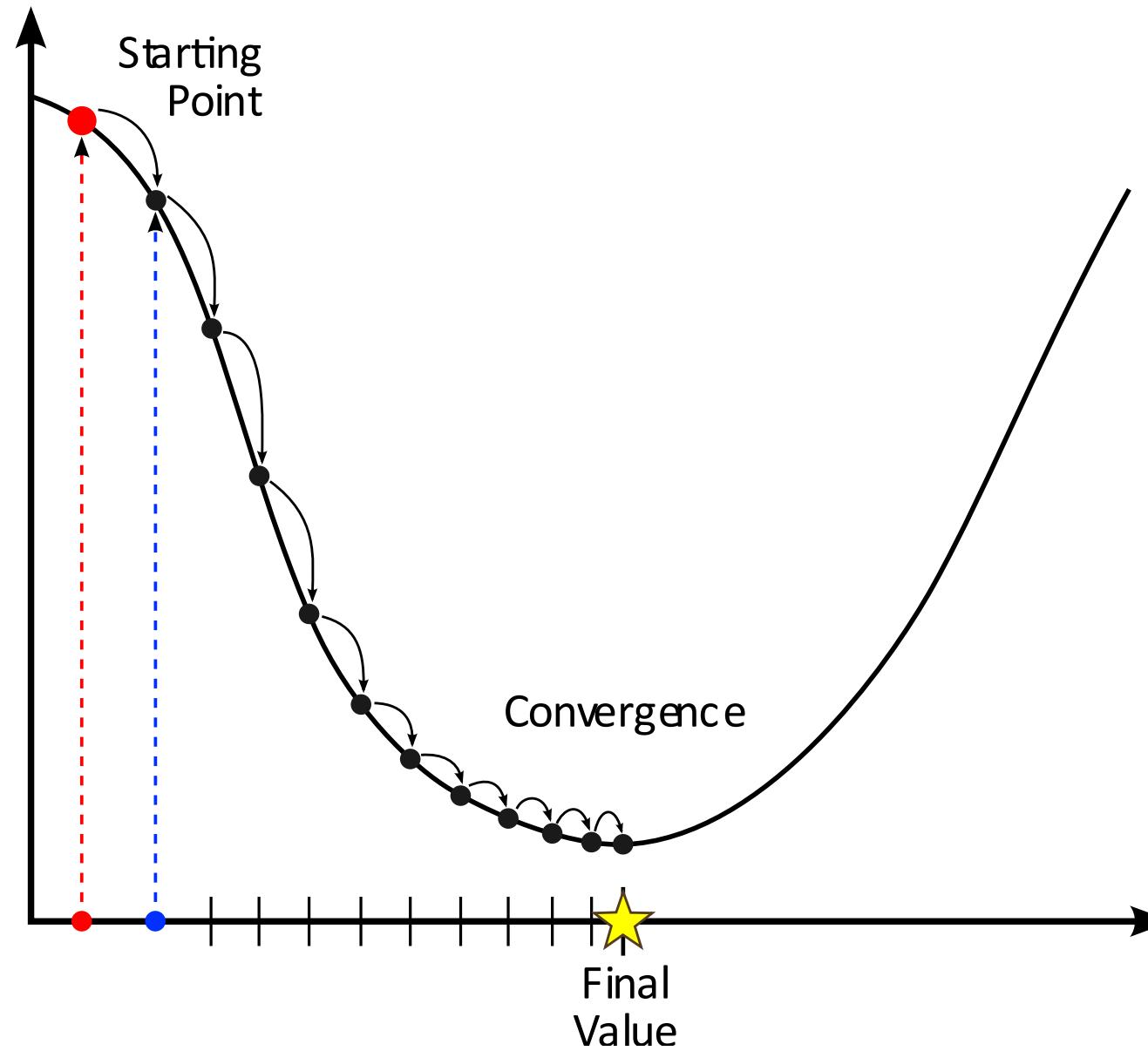


$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.



Gradient descent

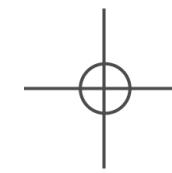
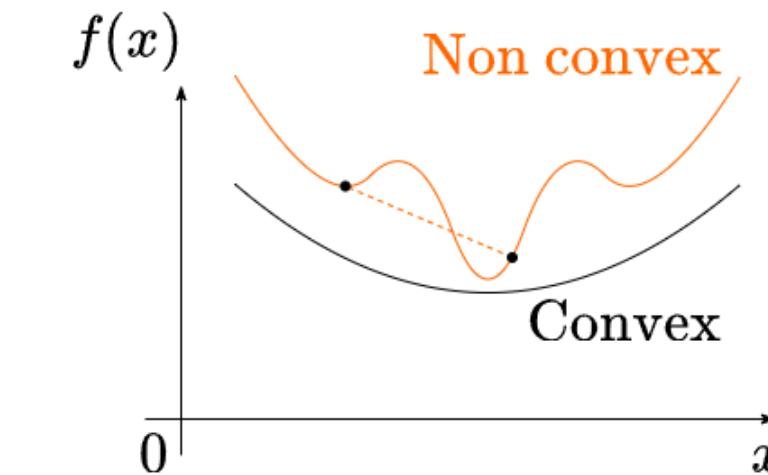
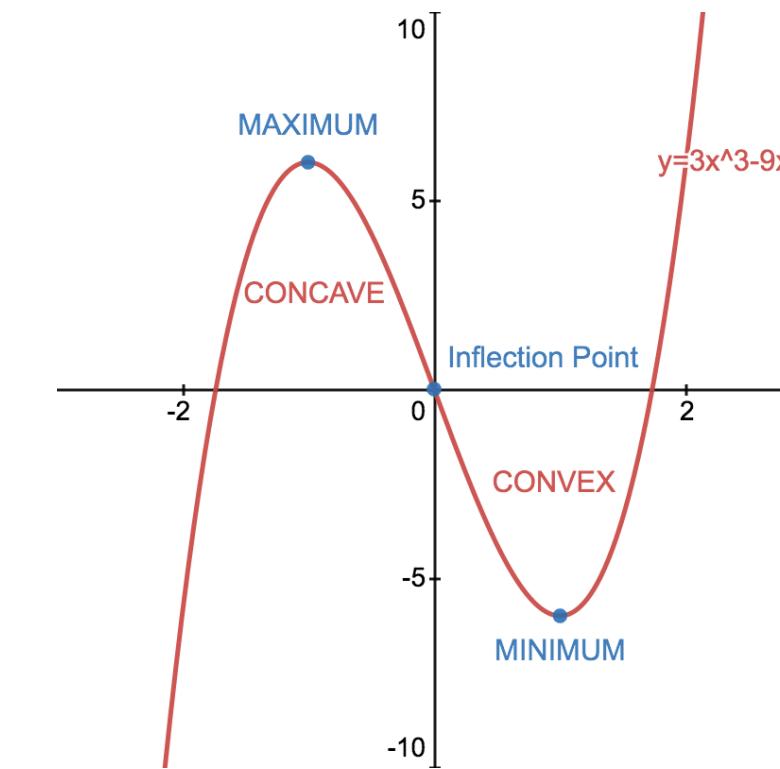


$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

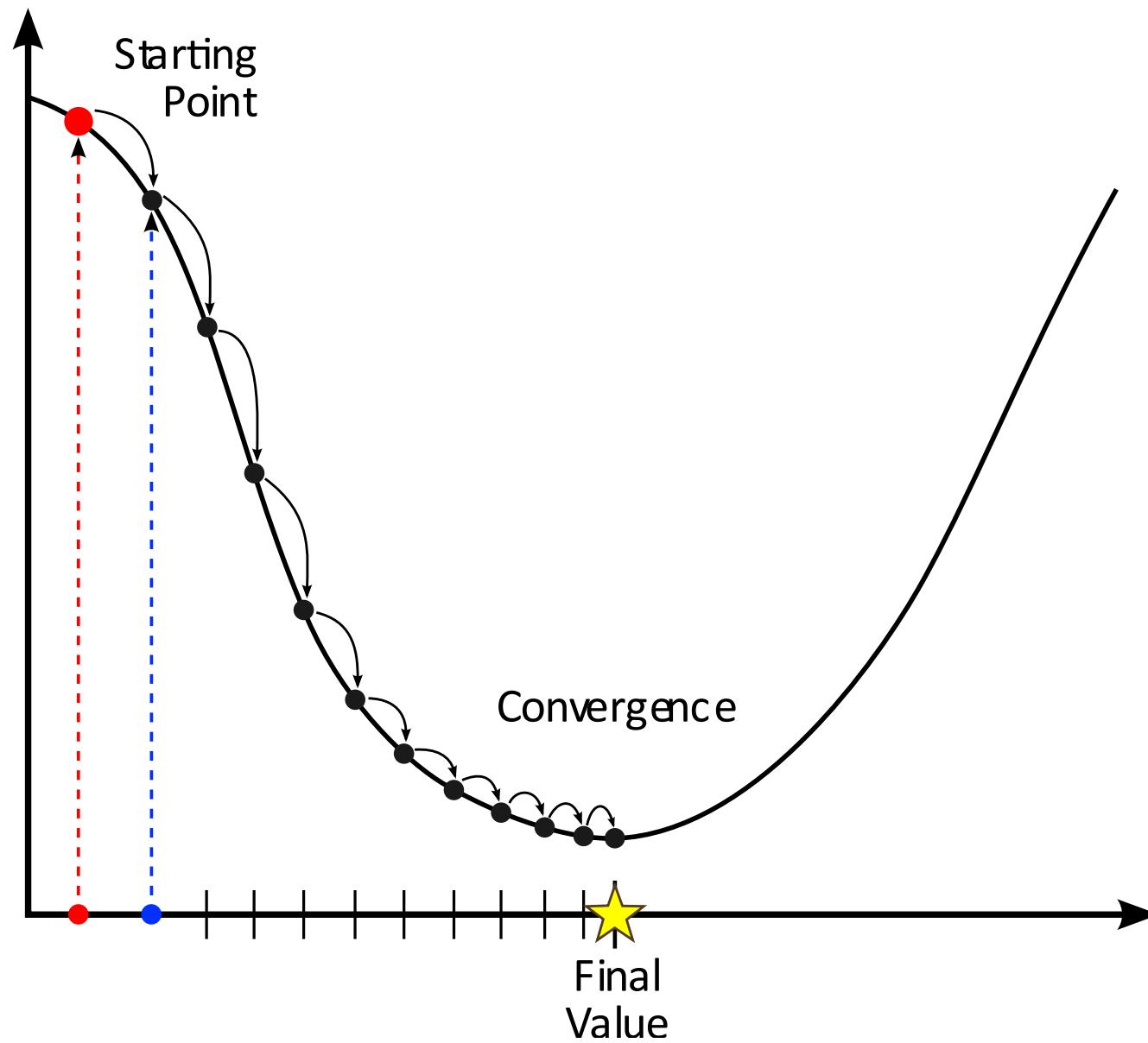
donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.

Se garantiza convergencia en el óptimo global si $J(\theta_k)$ tiene las siguientes propiedades:

Función convexa:



Gradient descent

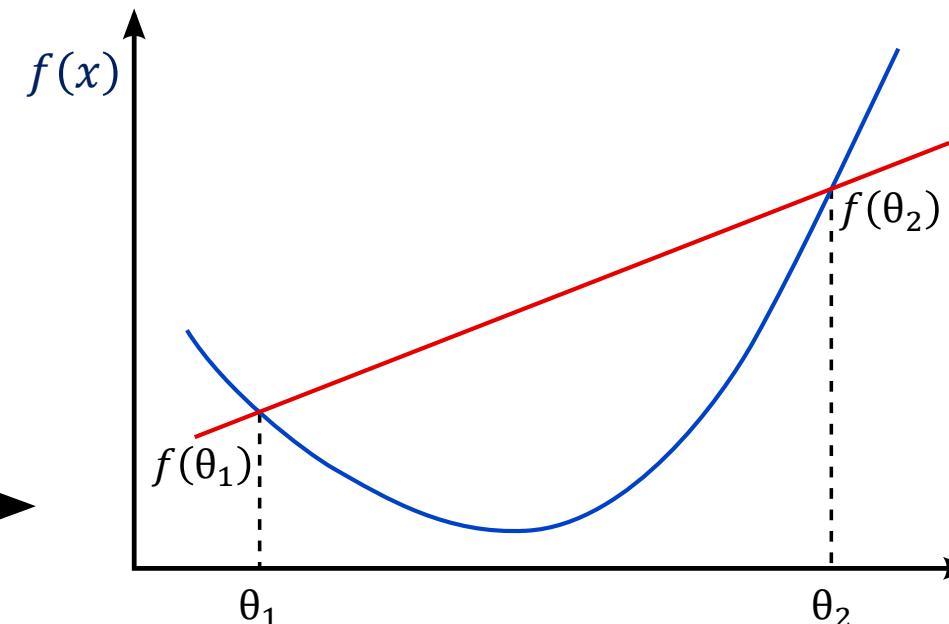


$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

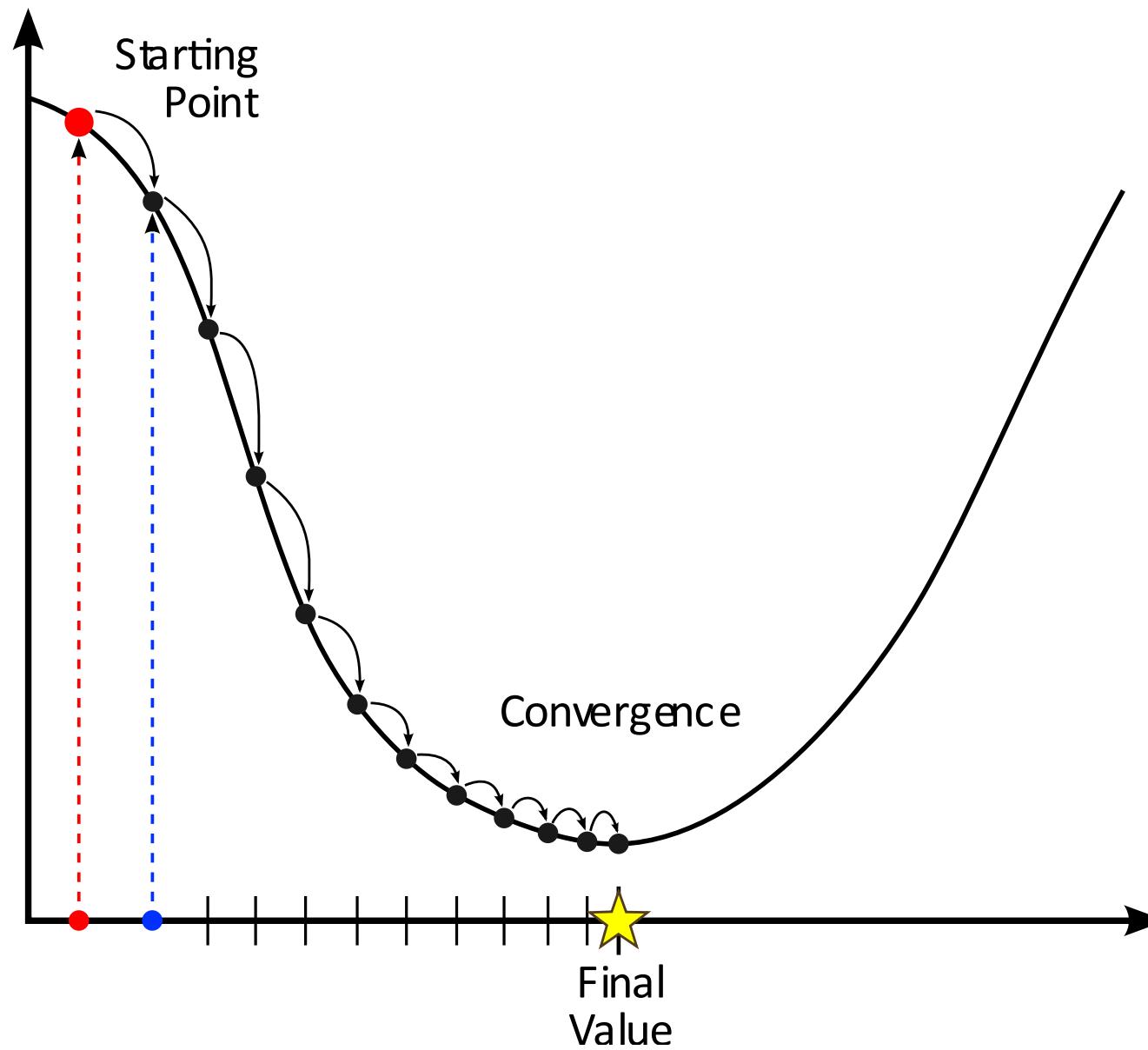
donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.

Se garantiza convergencia en el óptimo global si $J(\theta_k)$ tiene las siguientes propiedades:

Función convexa:



Gradient descent

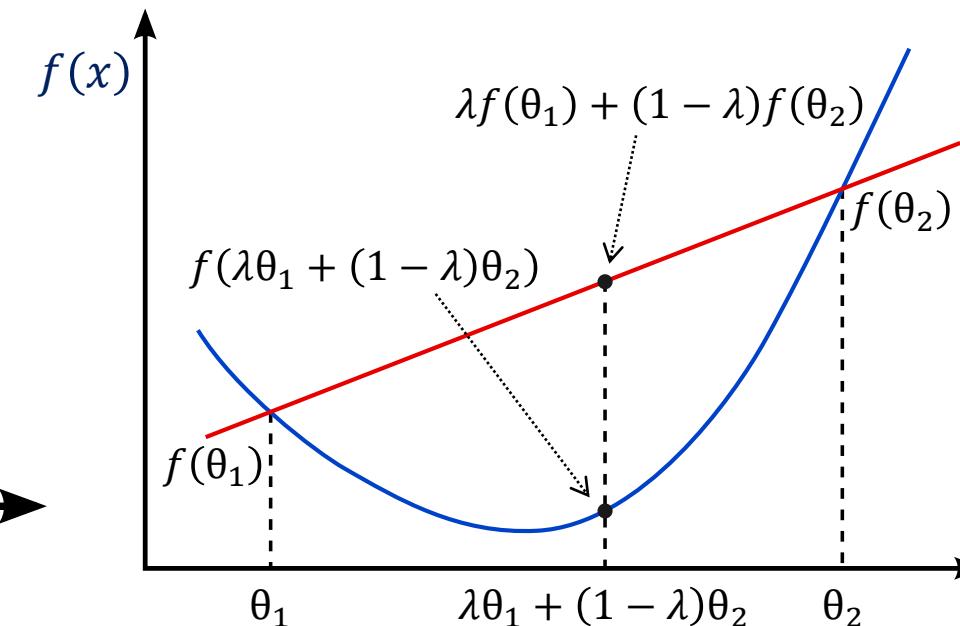


$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

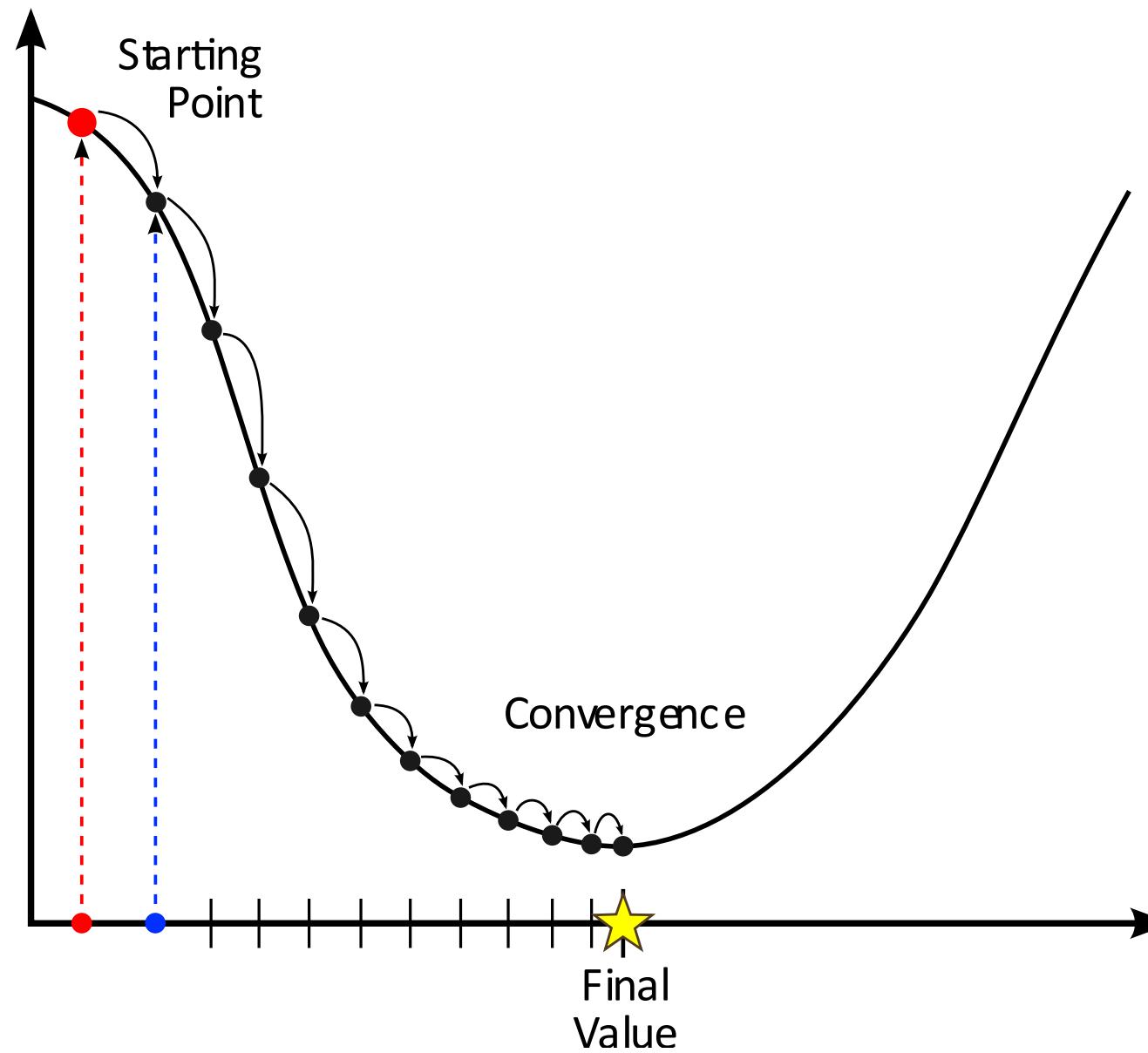
donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.

Se garantiza convergencia en el óptimo global si $J(\theta_k)$ tiene las siguientes propiedades:

Función convexa:



Gradient descent

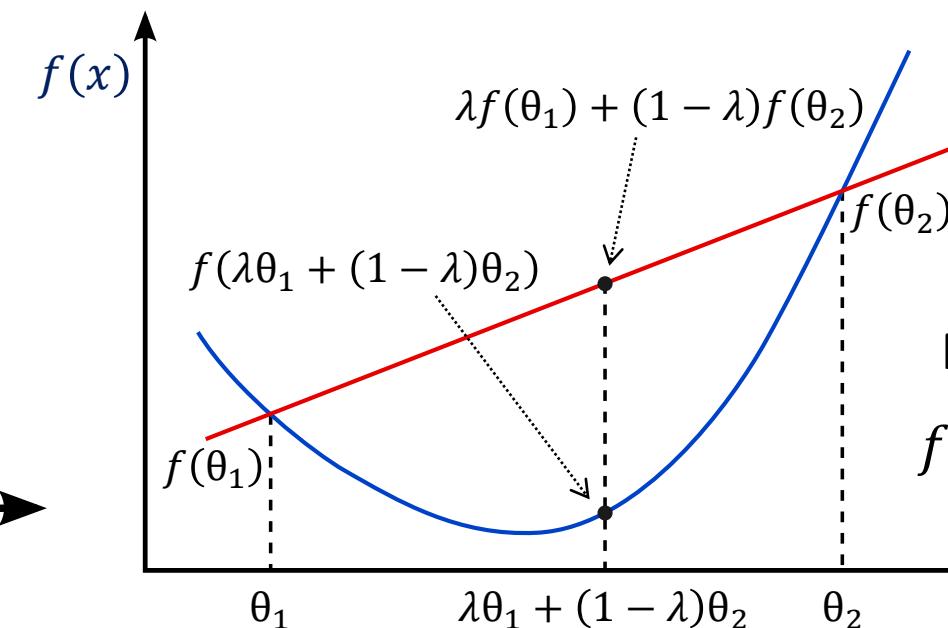


$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

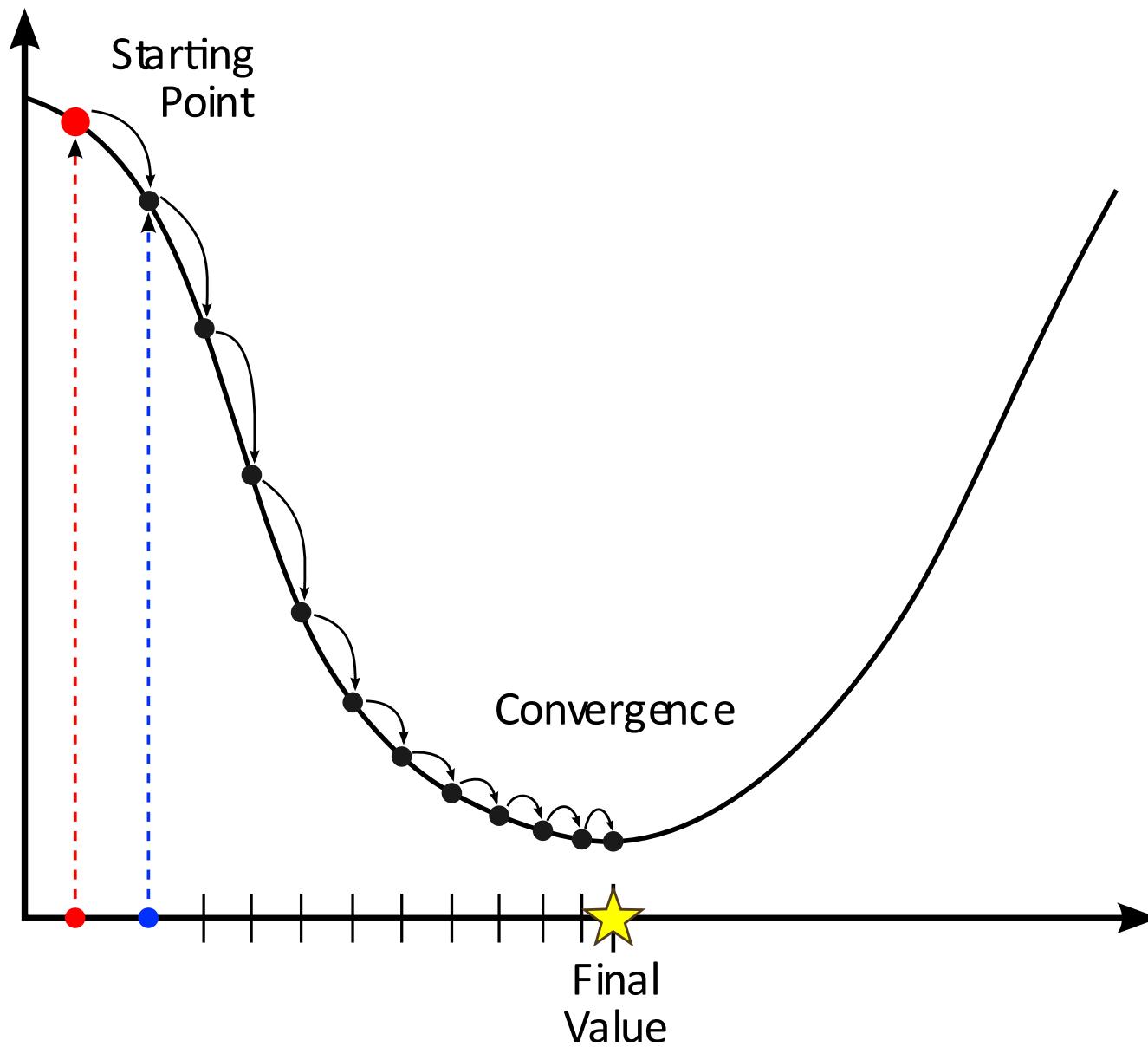
donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.

Se garantiza convergencia en el óptimo global si $J(\theta_k)$ tiene las siguientes propiedades:

Función convexa:



Gradient descent



$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.

Se garantiza convergencia en el óptimo global si $J(\theta_k)$ tiene las siguientes propiedades:

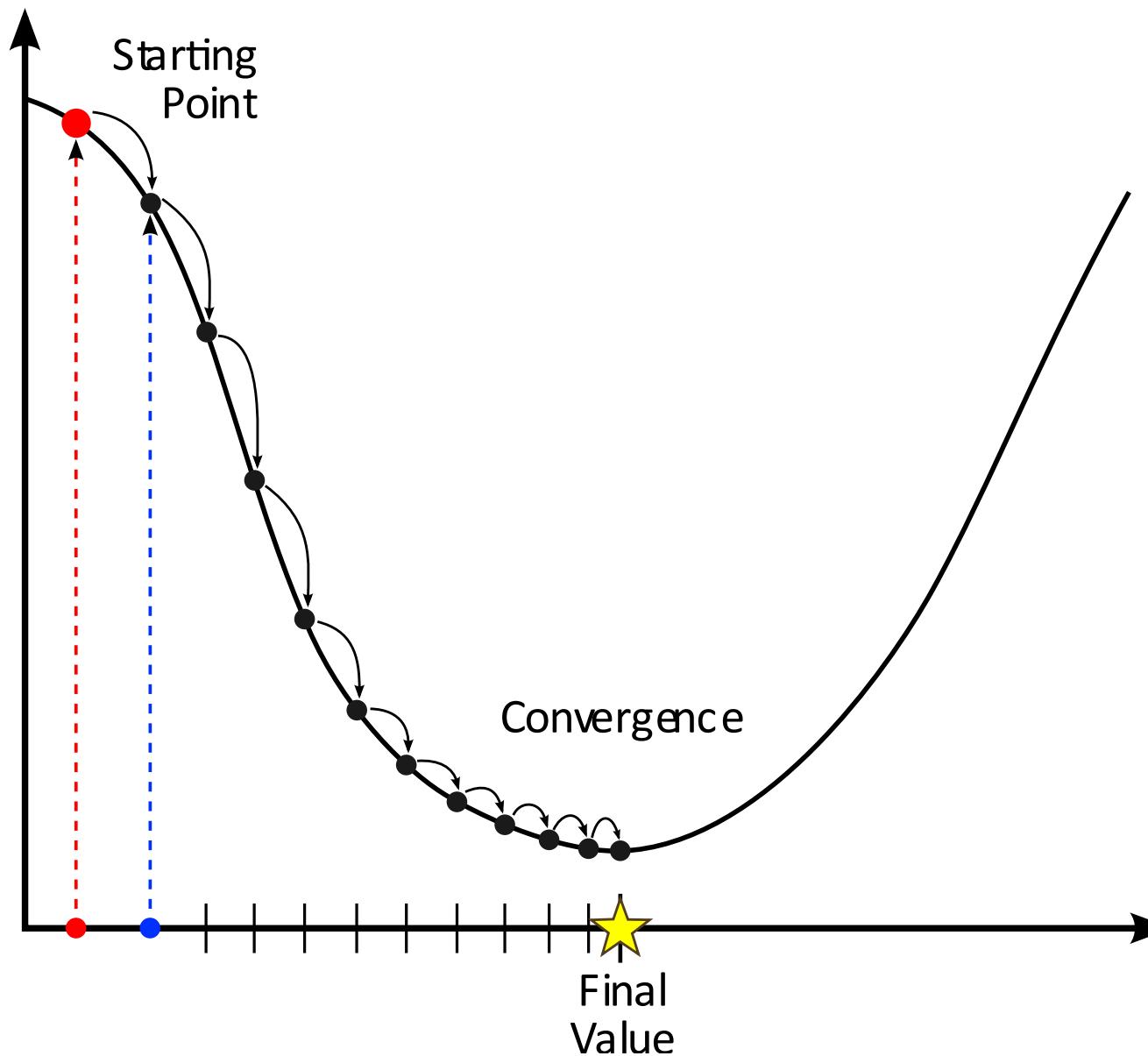
Función convexa: Para cualquier θ_1, θ_2 y $\lambda \in [0,1]$, se cumple:

$$J(\lambda\theta_1 + (1 - \lambda)\theta_2) \leq \lambda J(\theta_1) + (1 - \lambda)J(\theta_2)$$

Si $J(\theta)$ es estrictamente convexa, el mínimo será único.



Gradient descent



$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

donde θ_k representa los parámetros en la iteración k , y el learning rate $\alpha > 0$.

Se garantiza convergencia en el óptimo global si $J(\theta_k)$ tiene las siguientes propiedades:

Función convexa: Para cualquier θ_1, θ_2 y $\lambda \in [0,1]$, se cumple:

$$J(\lambda\theta_1 + (1 - \lambda)\theta_2) \leq \lambda J(\theta_1) + (1 - \lambda)J(\theta_2)$$

Si $J(\theta)$ es estrictamente convexa, el mínimo será único.

Lipschitz Continuity del Gradiente: Existe una constante $L > 0$ tal que:

$$\|\nabla J(\theta_1) - \nabla J(\theta_2)\| \leq L\|\theta_1 - \theta_2\|, \quad \forall \theta_1, \theta_2$$

Esto implica que $J(\theta)$ es una función suavemente diferenciable con una curvatura controlada.

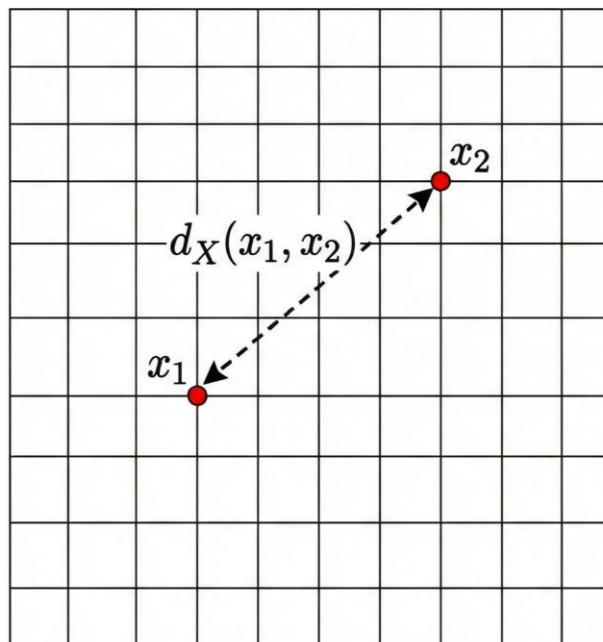


Lipschitz Continuity

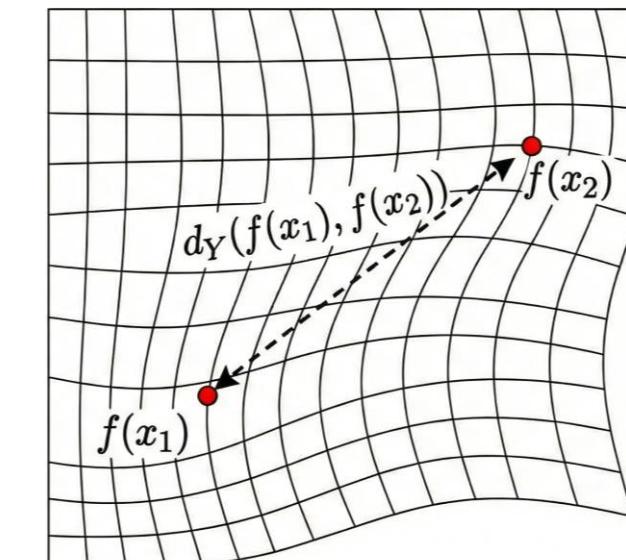
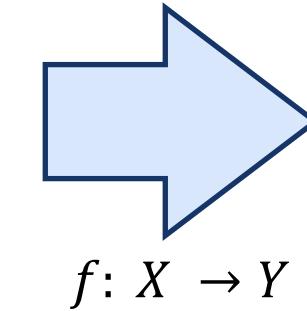
Sean (X, d_X) e (Y, d_Y) dos espacios métricos. Una función $f: X \rightarrow Y$ se dice Lipschitz continua si existe una constante real $L \geq 0$ tal que, para todo $x_1, x_2 \in X$:

$$d_Y(f(x_1), f(x_2)) \leq L \cdot d_X(x_1, x_2)$$

A la constante L se le denomina constante de Lipschitz de la función f .

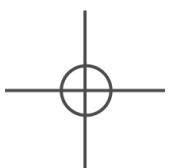


Espacio original X



Espacio Final Y

La deformación del espacio es acotada por la constante de Lipschitz L .



Lipschitz Continuity

Sean (X, d_X) e (Y, d_Y) dos espacios métricos. Una función $f: X \rightarrow Y$ se dice Lipschitz continua si existe una constante real $L \geq 0$ tal que, para todo $x_1, x_2 \in X$:

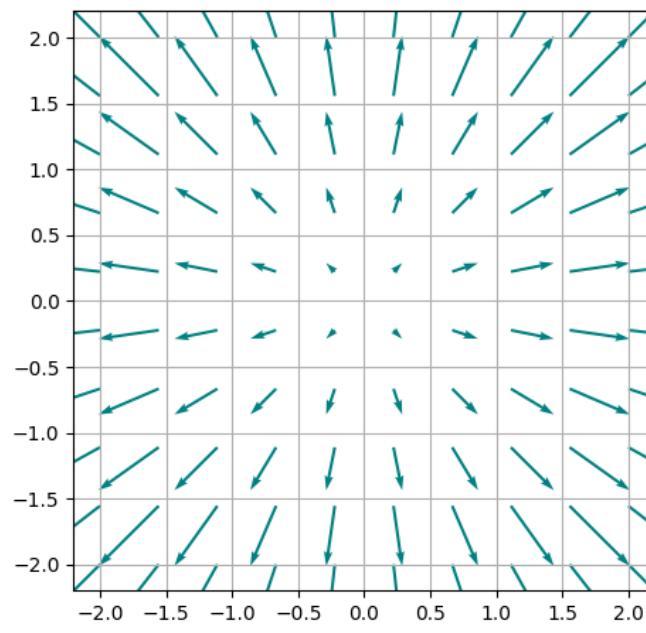
$$d_Y(f(x_1), f(x_2)) \leq K \cdot d_X(x_1, x_2)$$

A la constante L se le denomina constante de Lipschitz de la función f .

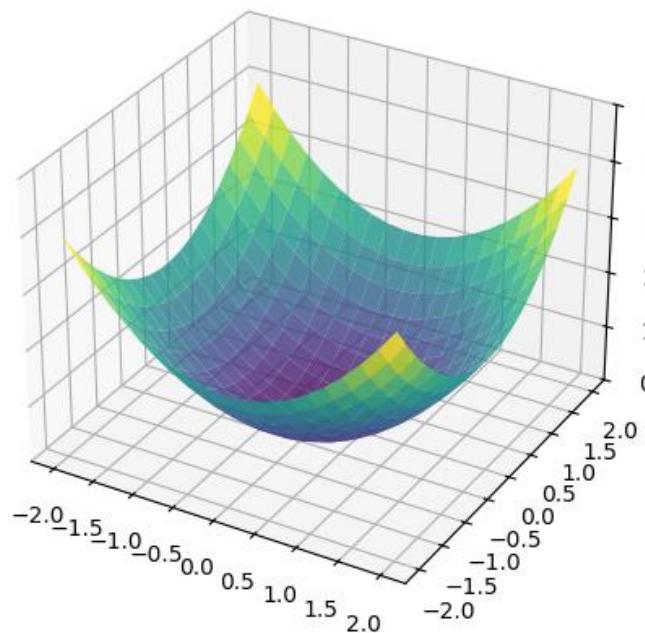
L-smoothness Cuando el gradiente es Lipschitz continuo. Es decir, existe una constante L tal que para todo x, y :

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

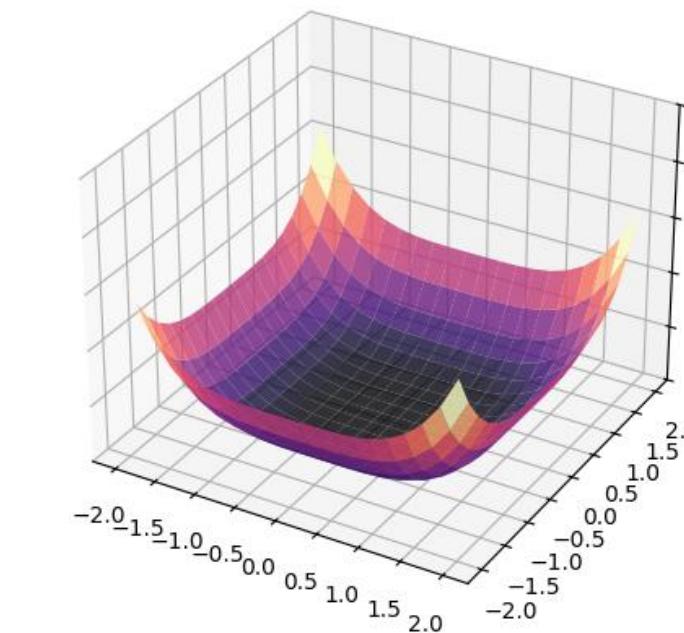
la continuidad de Lipschitz sobre el gradiente limita la curvatura máxima.



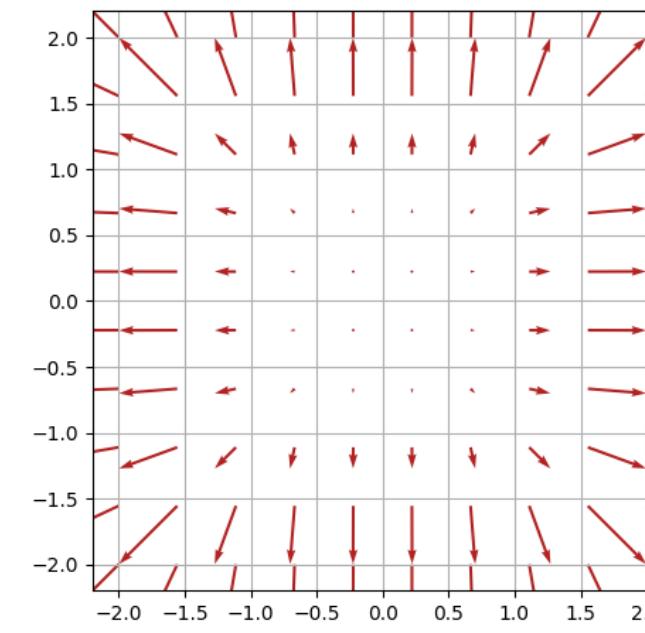
Gradiente Lipschitz
(Crecimiento Lineal)



$x^2 + y^2$
(Curvatura Constante)



$x^4 + y^4$
(Curvatura Explosiva)



Gradiente No Lipschitz
(Crecimiento Cúbico Explosivo)



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

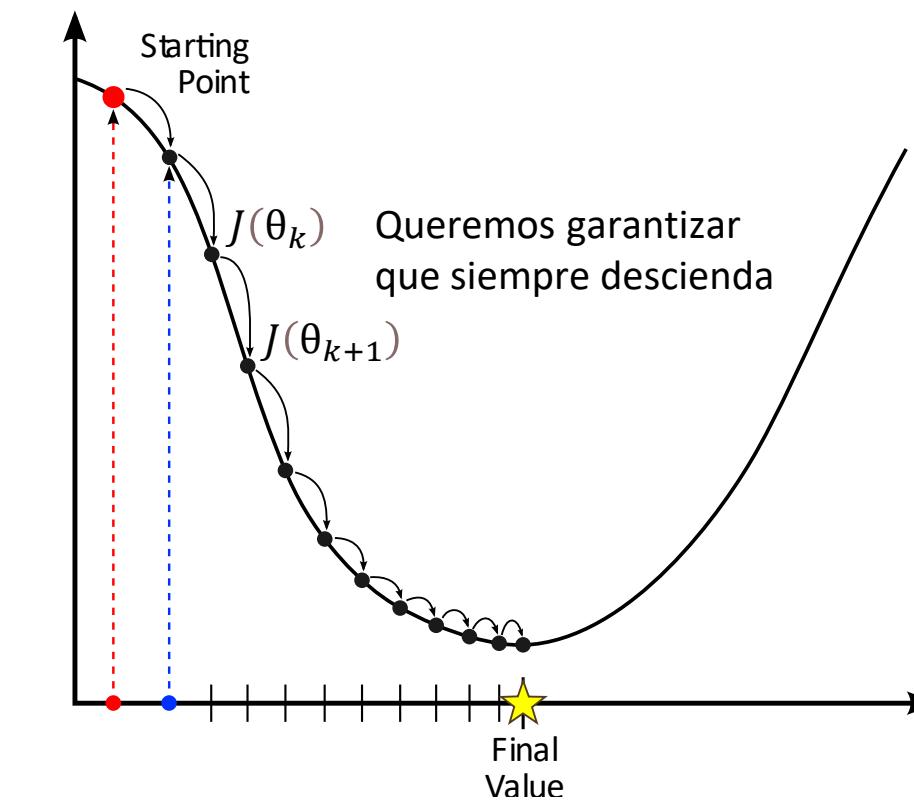


Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Como primer paso, buscaremos garantizar que $J(\theta_{k+1}) < J(\theta_k)$, es decir que la sucesión $\{J(\theta_k)\}_{k=0}^{\infty}$ sea **monótonamente decreciente**.

→ Esto significa que la gradiente se hará más pequeña en cada iteración.



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

Vamos a demostrar este Lema!



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **serie Taylor** de la función $J(\theta)$ alrededor de un punto θ_k se expresa como:

$$J(\theta_{k+1}) = \sum_{r=0}^{\infty} \frac{1}{r!} D^r J(\theta_k) [\theta_{k+1} - \theta_k]^r$$

donde $D^r J(\theta_k)$ representa el tensor de derivadas de orden r evaluado en θ_k .

Por ejemplo:

- cuando $r = 1$, $D^1 J(\theta_k)$ es el gradiente $\nabla J(\theta_k)$
- cuando $r = 2$, $D^2 J(\theta_k)$ es la matriz Hessiana $\nabla^2 J(\theta_k)$.



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

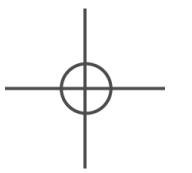
La **expansión de Taylor** de orden p de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \sum_{r=1}^{p-1} \frac{1}{r!} D^r J(\theta_k) [\theta_{k+1} - \theta_k]^r + R_p$$

El término de resto se expresa como:

$$R_p = \int_0^1 \frac{(1-t)^{p-1}}{(p-1)!} D^p J(\theta_t) [\theta_{k+1} - \theta_k]^p dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

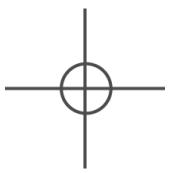
$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla J(\theta + d) - \nabla J(\theta)\| \leq L\|d\|, \quad \forall \theta, d \in \mathbb{R}^n$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla J(\theta + d) - \nabla J(\theta)\| \leq L\|d\|, \quad \forall \theta, d \in \mathbb{R}^n$$

Desarrollamos:
$$\frac{\|\nabla J(\theta + d) - \nabla J(\theta)\|}{\|d\|} \leq L$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla J(\theta + d) - \nabla J(\theta)\| \leq L\|d\|, \quad \forall \theta, d \in \mathbb{R}^n$$

Desarrollamos: $\lim_{d \rightarrow 0} \frac{\|\nabla J(\theta + d) - \nabla J(\theta)\|}{\|d\|} \leq L$ **Definición de derivada!**
 $\|\nabla^2 J(\theta)\| \leq L$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla^2 J(\theta)\| \leq L$$

Vamos a usar esta cota dentro de la integral de R_2 . Ordenamos:

$$(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_k) (\theta_{k+1} - \theta_k) \leq \|\nabla^2 J(\theta_k)\| \|\theta_{k+1} - \theta_k\|^2$$

aquí aprovechamos que $\nabla^2 J(\theta_k)$ simétrica y positiva semidefinida, ya que J es convexa. (**Para estas matrices, se cumple:** $x^T A x \leq \|A\| \|x\|^2$)



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla^2 J(\theta)\| \leq L$$

Vamos a usar esta cota dentro de la integral de R_2 . Ordenamos:

$$(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_k) (\theta_{k+1} - \theta_k) \leq \|\nabla^2 J(\theta_k)\| \|\theta_{k+1} - \theta_k\|^2$$

aquí aprovechamos que $\nabla^2 J(\theta_k)$ simétrica y positiva semidefinida, ya que J es convexa. (**Para estas matrices, se cumple:** $x^T A x \leq \|A\| \|x\|^2$)

En esta segunda inecuación podemos usar la cota de $\|\nabla^2 J(\theta)\|$:

$$(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) \leq L \|\theta_{k+1} - \theta_k\|^2$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla^2 J(\theta)\| \leq L$$

Usamos esta cota dentro de la integral de R_2 :

$$(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) \leq L \|\theta_{k+1} - \theta_k\|^2$$

Sustituyendo este resultado en la cota en $R_2(\theta, \theta_{k+1})$:

$$\begin{aligned} R_2(\theta_k, \theta_{k+1}) &\leq \int_0^1 (1-t) L \|\theta_{k+1} - \theta_k\|^2 dt \\ &= L \|\theta_{k+1} - \theta_k\|^2 \int_0^1 (1-t) dt = \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \end{aligned}$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

La **expansión de Taylor** de orden 2 de una función $J(\theta)$ alrededor de un punto θ_k es obtiene:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

El término de resto de segundo orden es:

$$R_2 = \int_0^1 (1-t)(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) dt$$

donde $\theta_t = \theta_k + t(\theta_{k+1} - \theta_k)$ es un punto intermedio entre θ_k y θ_{k+1} .

Este término integra la contribución de la curvatura (la matriz Hessiana) a lo largo del trayecto que une θ_k y θ_{k+1} , ponderando la contribución según $1-t$.

Sabemos que si $\nabla J(\theta)$ es L-Lipschitz continuo, entonces existe $L > 0$:

$$\|\nabla^2 J(\theta)\| \leq L$$

Usamos esta cota dentro de la integral de R_2 :

$$(\theta_{k+1} - \theta_k)^T \nabla^2 J(\theta_t) (\theta_{k+1} - \theta_k) \leq L \|\theta_{k+1} - \theta_k\|^2$$

Sustituyendo este resultado en la cota en $R_2(\theta, \theta_{k+1})$:

$$\begin{aligned} R_2(\theta_k, \theta_{k+1}) &\leq \int_0^1 (1-t) L \|\theta_{k+1} - \theta_k\|^2 dt \\ &= L \|\theta_{k+1} - \theta_k\|^2 \int_0^1 (1-t) dt = \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \end{aligned}$$

Por lo tanto:

$$J(\theta_{k+1}) = J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + R_2$$

$$J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**: $J(\theta_{k+1}) \leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2$

Recordemos: $\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$

Usamos: $\theta_{k+1} - \theta_k = -\alpha \nabla J(\theta_k)$

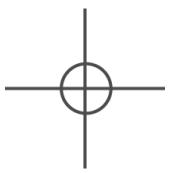


Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**:

$$\begin{aligned} J(\theta_{k+1}) &\leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \\ &= J(\theta_k) + \nabla J(\theta_k)^T (-\alpha \nabla J(\theta_k)) + \frac{L}{2} (-\alpha \nabla J(\theta_k))^2 \end{aligned}$$

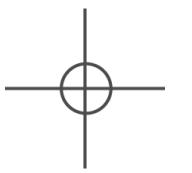


Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**:

$$\begin{aligned} J(\theta_{k+1}) &\leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \\ &= J(\theta_k) + \nabla J(\theta_k)^T (-\alpha \nabla J(\theta_k)) + \frac{L}{2} (-\alpha \nabla J(\theta_k))^2 \\ &= J(\theta_k) - \alpha \|\nabla J(\theta_k)\|^2 + \frac{L\alpha^2}{2} \|\nabla J(\theta_k)\|^2 \\ &= J(\theta_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \quad \leftarrow \text{Aquí solo agrupamos} \end{aligned}$$

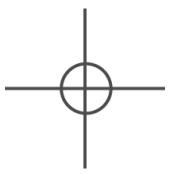


Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Usamos el **Lema de Descenso**:

$$\begin{aligned} J(\theta_{k+1}) &\leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \\ &= J(\theta_k) + \nabla J(\theta_k)^T (-\alpha \nabla J(\theta_k)) + \frac{L}{2} (-\alpha \nabla J(\theta_k))^2 \\ &= J(\theta_k) - \alpha \|\nabla J(\theta_k)\|^2 + \frac{L\alpha^2}{2} \|\nabla J(\theta_k)\|^2 \\ &= J(\theta_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \\ \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 &\leq J(\theta_k) - J(\theta_{k+1}) \end{aligned}$$

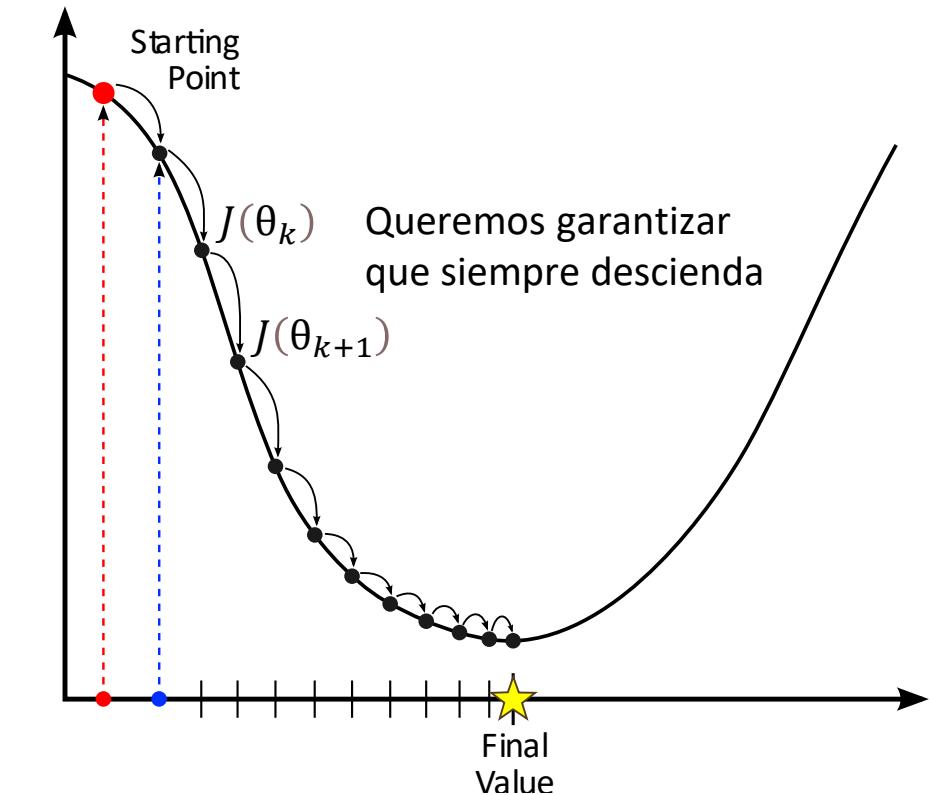


Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

$$\begin{aligned}
 \text{Usamos el Lema de Descenso: } J(\theta_{k+1}) &\leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \\
 &= J(\theta_k) + \nabla J(\theta_k)^T (-\alpha \nabla J(\theta_k)) + \frac{L}{2} (-\alpha \nabla J(\theta_k))^2 \\
 &= J(\theta_k) - \alpha \|\nabla J(\theta_k)\|^2 + \frac{L\alpha^2}{2} \|\nabla J(\theta_k)\|^2 \\
 &= J(\theta_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2
 \end{aligned}$$

?
 $\alpha \left(1 - \frac{L\alpha}{2}\right)$ Positivo $\|\nabla J(\theta_k)\|^2$ Positivo



Si queremos que la gradiente decrezca en cada iteración, es decir $J(\theta_{k+1}) < J(\theta_k)$, hacemos que $0 < \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2$ ya que implicaría que $0 < J(\theta_k) - J(\theta_{k+1})$ por transitividad.

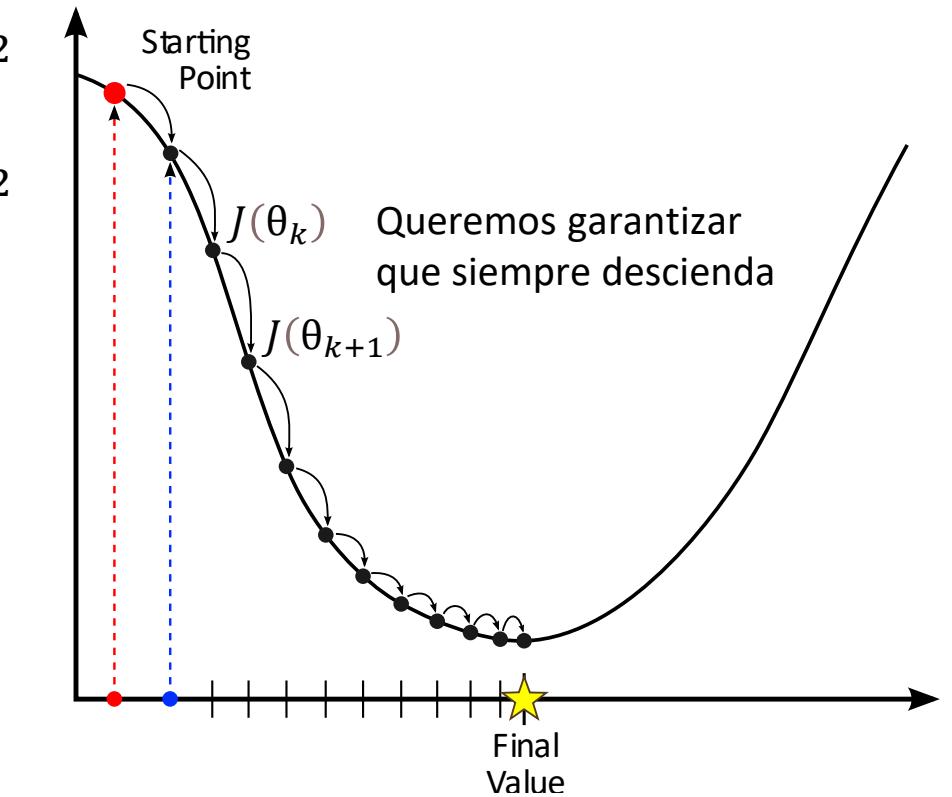


Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

$$\begin{aligned}
 \text{Usamos el Lema de Descenso: } J(\theta_{k+1}) &\leq J(\theta_k) + \nabla J(\theta_k)^T (\theta_{k+1} - \theta_k) + \frac{L}{2} \|\theta_{k+1} - \theta_k\|^2 \\
 &= J(\theta_k) + \nabla J(\theta_k)^T (-\alpha \nabla J(\theta_k)) + \frac{L}{2} (-\alpha \nabla J(\theta_k))^2 \\
 &= J(\theta_k) - \alpha \|\nabla J(\theta_k)\|^2 + \frac{L\alpha^2}{2} \|\nabla J(\theta_k)\|^2 \\
 &= J(\theta_k) - \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2
 \end{aligned}$$

?
 $\alpha \left(1 - \frac{L\alpha}{2}\right)$ Positivo $\|\nabla J(\theta_k)\|^2$ Positivo



Si queremos que la gradiente decrezca en cada iteración, es decir $J(\theta_{k+1}) < J(\theta_k)$, hacemos que $0 < \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2$ ya que implicaría que $0 < J(\theta_k) - J(\theta_{k+1})$ por transitividad.

Solo basta hacer: $1 - \frac{L\alpha}{2} > 0 \iff 0 < \alpha < \frac{2}{L}$

Entonces, esta condición garantiza que la gradiente decrezca en cada paso.



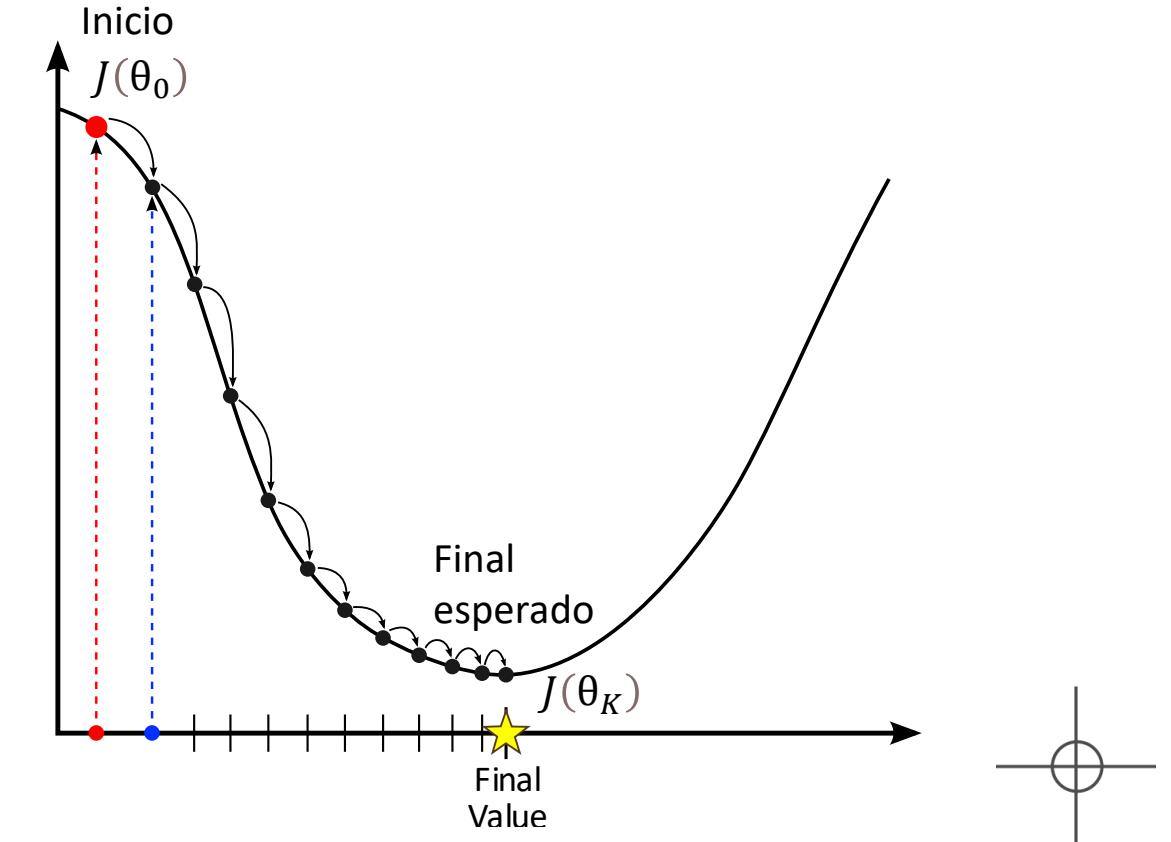
Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

$$\text{Para } 0 < \alpha < \frac{2}{L}: \quad 0 < \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k) - J(\theta_{k+1}) \quad \Rightarrow \quad J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

Nuestro siguiente objetivo será garantizar que la **gradiente descenderá hasta el mínimo global**.



Gradient descent

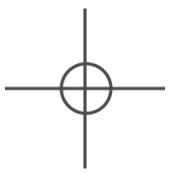
El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_{k+1}) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-2})\|^2 \leq J(\theta_{K-2}) - J(\theta_{K-1})$$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-3})\|^2 \leq J(\theta_{K-3}) - J(\theta_{K-2})$$

⋮

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_0)\|^2 \leq J(\theta_0) - J(\theta_1)$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-2})\|^2 \leq J(\theta_{K-2}) - J(\theta_{K-1})$$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-3})\|^2 \leq J(\theta_{K-3}) - J(\theta_{K-2})$$

⋮

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_0)\|^2 \leq J(\theta_0) - J(\theta_1)$$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \sum_{k=0}^{K-1} \|\nabla J(\theta_k)\|^2 \leq \sum_{k=0}^{K-1} [J(\theta_k) - J(\theta_{k+1})] = J(\theta_0) - J(\theta_K)$$

(suma telescopica)



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

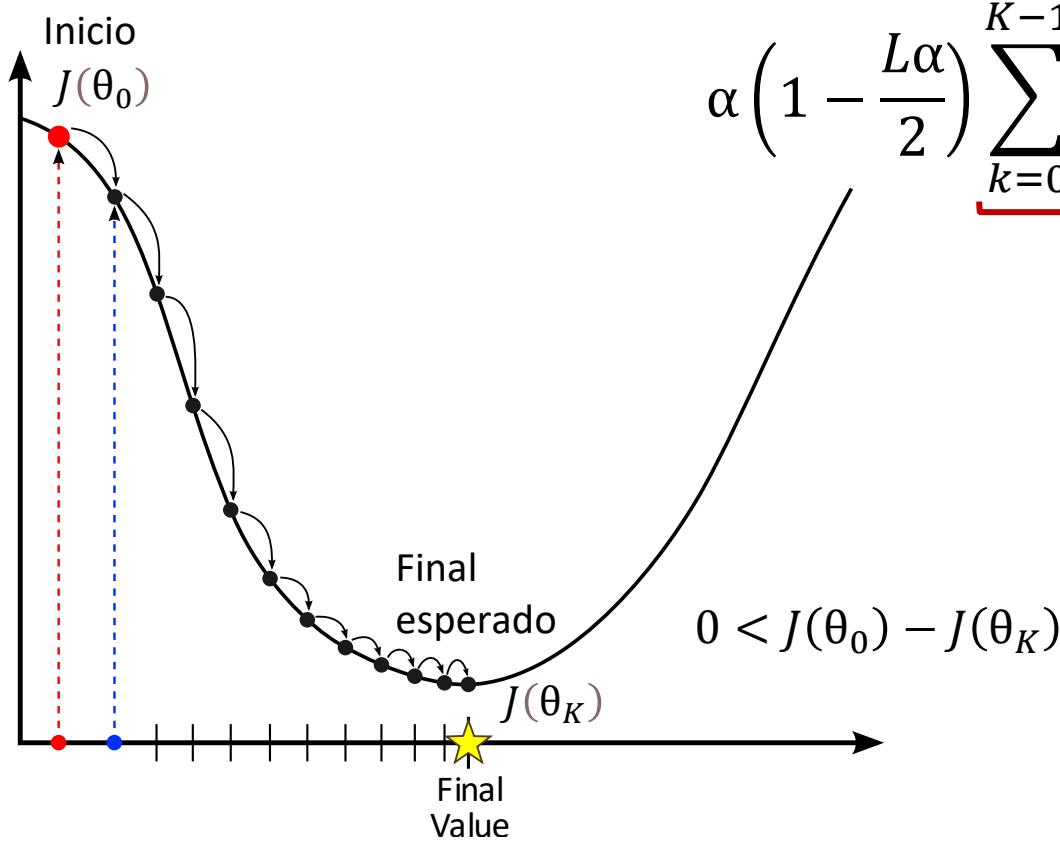
Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \sum_{k=0}^{K-1} \|\nabla J(\theta_k)\|^2 \leq J(\theta_0) - J(\theta_K)$$

Serie

Acotada porque $J(\theta_k)$ es monótona decreciente y tiene un límite inferior (función convexa)



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \sum_{k=0}^{K-1} \|\nabla J(\theta_k)\|^2 \leq J(\theta_0) - J(\theta_K)$$

Acotada porque $J(\theta_k)$ es monótona decreciente y tiene un límite inferior (función convexa)

Entonces: $\sum_{k=0}^{\infty} \|\nabla J(\theta_k)\|^2 \leq \infty$

Convergencia de una serie:

Si $\sum_{k=0}^{\infty} a_k \leq \infty$, entonces $a_k \rightarrow 0$ cuando $k \rightarrow \infty$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

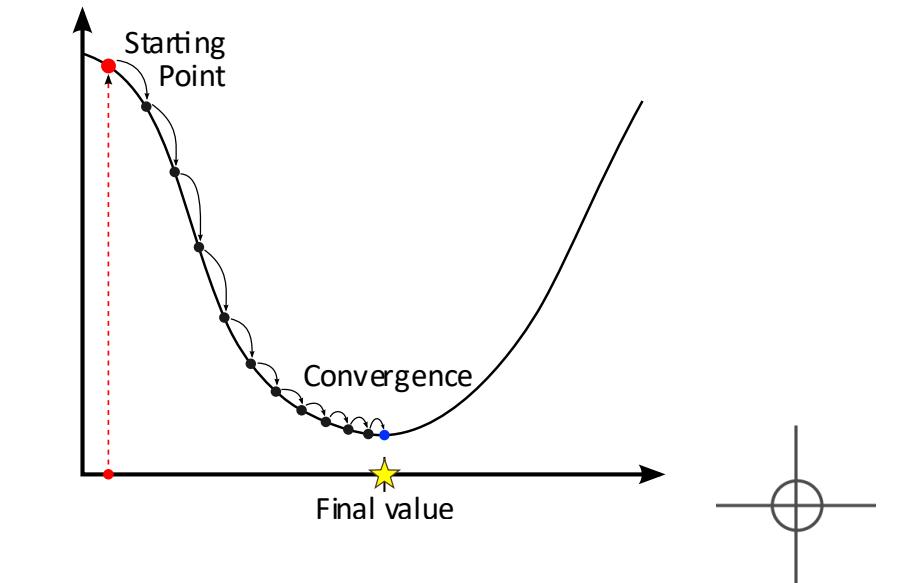
De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

$$\alpha \left(1 - \frac{L\alpha}{2}\right) \sum_{k=0}^{K-1} \|\nabla J(\theta_k)\|^2 \leq J(\theta_0) - J(\theta_K)$$

Acotada porque $J(\theta_k)$ es monótona decreciente y tiene un límite inferior (función convexa)

Entonces: $\sum_{k=0}^{\infty} \|\nabla J(\theta_k)\|^2 \leq \infty$. Por lo tanto $\lim_{k \rightarrow \infty} \|\nabla J(\theta_k)\|^2 = 0$

y por lo tanto $\lim_{k \rightarrow \infty} \|\nabla J(\theta_k)\| = 0$



Gradient descent

El objetivo es demostrar que el descenso de gradiente converge al óptimo global θ^* , donde: $\theta^* = \arg \min_{\theta} J(\theta)$

Para $0 < \alpha < \frac{2}{L}$:

$$J(\theta_{k+1}) < J(\theta_k) + \alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_k)\|^2 \leq J(\theta_k)$$

Por lo tanto, $\{J(\theta_k)\}$ es una sucesión decreciente.

De la desigualdad, ordenamos: $\alpha \left(1 - \frac{L\alpha}{2}\right) \|\nabla J(\theta_{K-1})\|^2 \leq J(\theta_{K-1}) - J(\theta_K)$

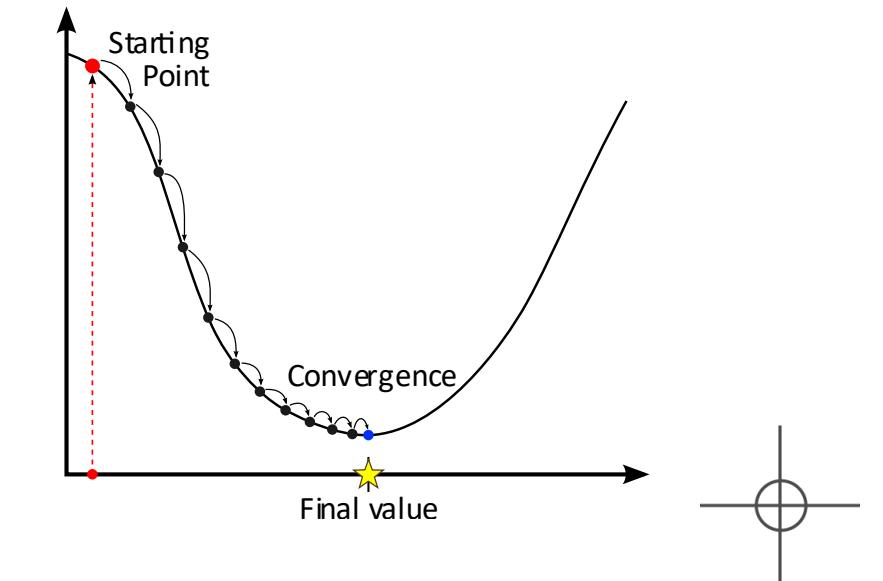
$$\alpha \left(1 - \frac{L\alpha}{2}\right) \sum_{k=0}^{K-1} \|\nabla J(\theta_k)\|^2 \leq J(\theta_0) - J(\theta_K)$$

Acotada porque $J(\theta_k)$ es monótona decreciente y tiene un límite inferior (función convexa)

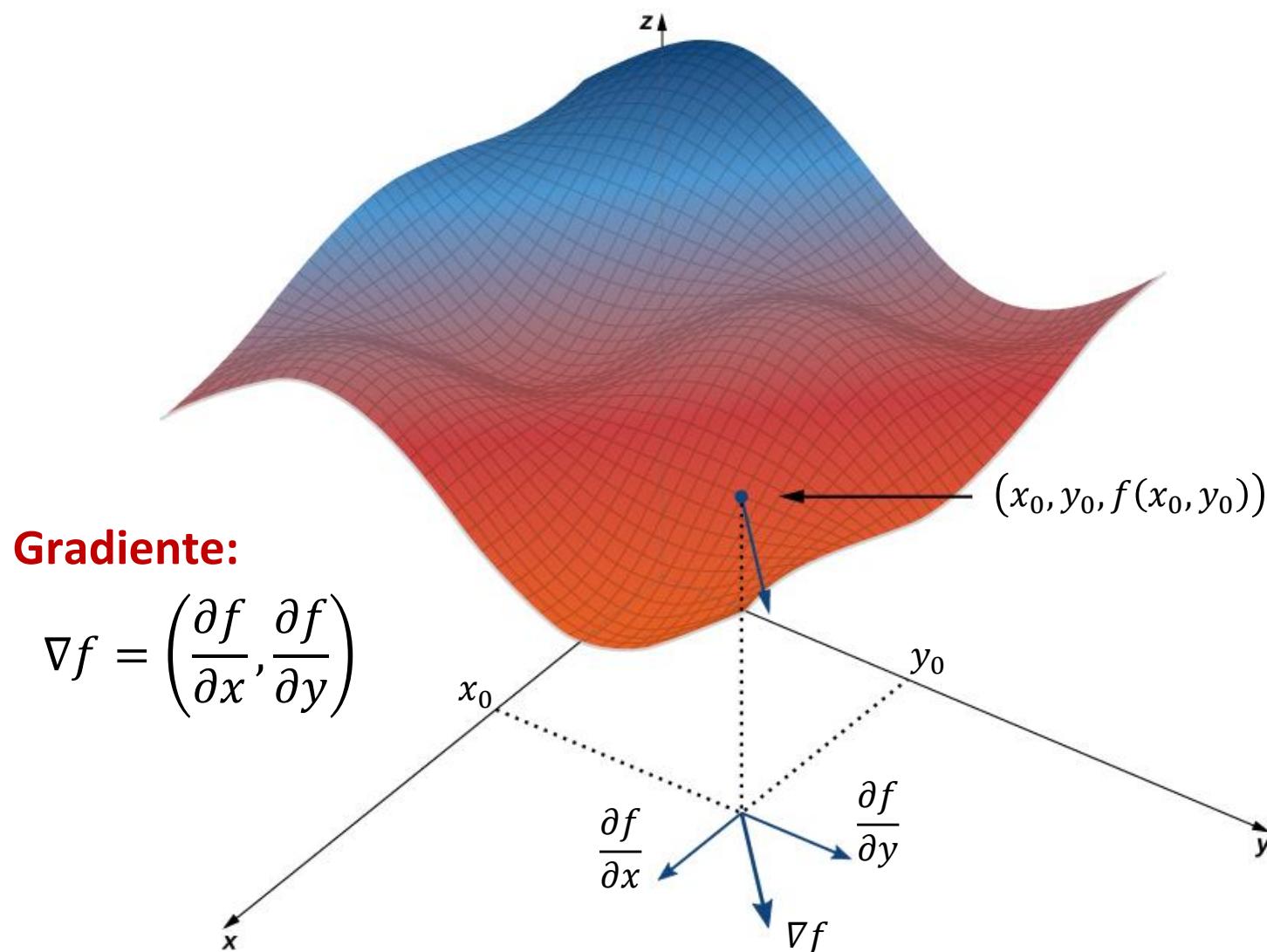
Entonces: $\sum_{k=0}^{\infty} \|\nabla J(\theta_k)\|^2 \leq \infty$. Por lo tanto $\lim_{k \rightarrow \infty} \|\nabla J(\theta_k)\|^2 = 0$

$$\text{y por lo tanto } \lim_{k \rightarrow \infty} \|\nabla J(\theta_k)\| = 0$$

Esto implica que **cuando $k \rightarrow \infty$ se alcanza el mínimo global** ya que $J(\theta_k)$ es **estRICTAMENTE CONVEXA**.



Gradient descent



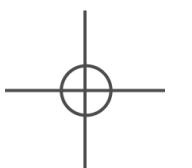
$$\theta_{k+1} = \theta_k - \alpha \nabla J(\theta_k)$$

donde θ_k representa los parámetros en la iteración k .

Se garantiza convergencia en el óptimo global si:

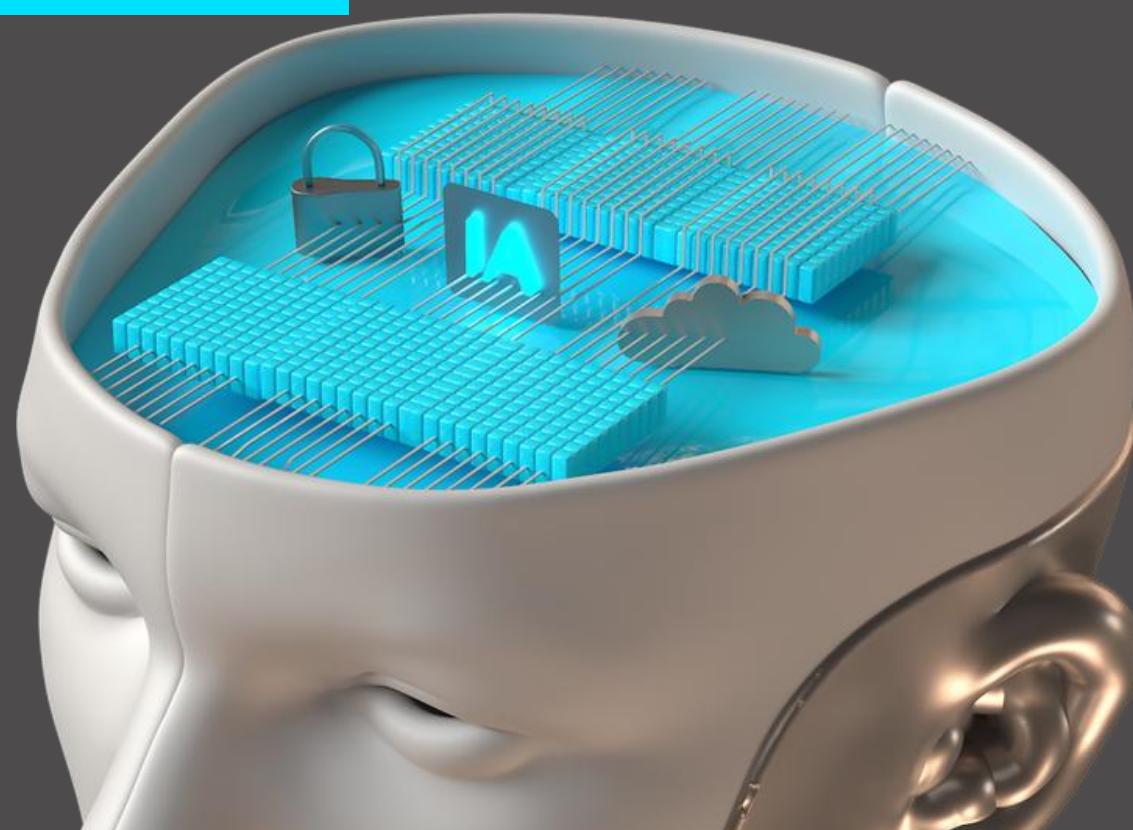
- Si $J(\theta_k)$ es convexa.
- La gradiente $\nabla J(\theta_k)$ es Lipschitz continua.
- El learning rate está en el rango $0 < \alpha < \frac{2}{L}$

Estos objetivos son fáciles de lograr con una regresión lineal.



Métricas

Absolutas y relativas.



Métricas absolutas

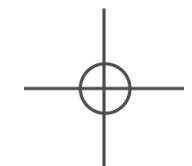
- Mean Squared Error (MSE):** $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$
- Penaliza fuertemente errores grandes.
 - Diferenciable y convexa, lo que facilita optimización.
 - Sensible a outliers.
 - Tiene unidades al cuadrado de la variable objetivo.

Es la métrica implícita detrás de OLS y de la regresión lineal clásica.

- Root Mean Squared Error (RMSE):** $RMSE = \sqrt{MSE}$
- Misma interpretación que MSE, pero en las mismas unidades que y .
 - Más interpretable en contextos físicos o económicos.
 - Mantiene la alta penalización de errores grandes.

- Mean Absolute Error (MAE):** $MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$
- Penaliza errores de forma lineal.
 - Más robusta a outliers que MSE/RMSE.
 - No es diferenciable en cero.
 - Corresponde a un supuesto implícito de ruido Laplaciano.

- En resumen:
- MSE/RMSE favorecen modelos que evitan errores grandes.
 - MAE favorece estabilidad frente a valores atípicos.
 - La elección refleja una decisión implícita sobre la distribución del ruido.



Métricas relativas

- **Las métricas absolutas:** ¿cuánto error hay?
- **Las métricas relativas:** ¿cuánto mejor es esto que una referencia simple?
¿cuánta variación del target estamos capturando?

La referencia más usada en regresión es el predictor constante: $\hat{y}_i^{(\text{base})} = \bar{y}$
que minimiza el MSE entre todos los predictores constantes.

$$\mathbf{R^2: Se define: SSE (sum of squared errors): } \quad \text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad \text{SSR (total sum of residuals): } \quad \text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\text{SST (total sum of squares): } \quad \text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$$

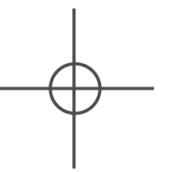
Entonces: $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$

Interpretación:

- $R^2 = 1$: ajuste perfecto ($\text{SSE} = 0$).
- $R^2 = 0$: tan bueno como predecir la media ($\text{SSE} = \text{SST}$).
- $R^2 < 0$: peor que predecir la media ($\text{SSE} > \text{SST}$). Esto ocurre con facilidad en test, o si el modelo está mal regularizado.

En OLS con intercepto se cumple: $\text{SST} = \text{SSR} + \text{SSE}$

Entonces, $R^2 = \frac{\text{SSR}}{\text{SST}}$ lo cual se relaciona con fracción de varianza explicada.



Métricas relativas

- **Las métricas absolutas:** ¿cuánto error hay?
- **Las métricas relativas:** ¿cuánto mejor es esto que una referencia simple?
¿cuánta variación del target estamos capturando?

La referencia más usada en regresión es el predictor constante: $\hat{y}_i^{(\text{base})} = \bar{y}$
que minimiza el MSE entre todos los predictores constantes.

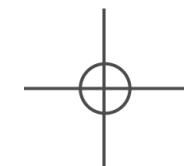
Adjusted R^2 : penalización por complejidad (pero sigue siendo in-sample)

Con p features (sin contar intercepto):

$$R_{\text{adj}}^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

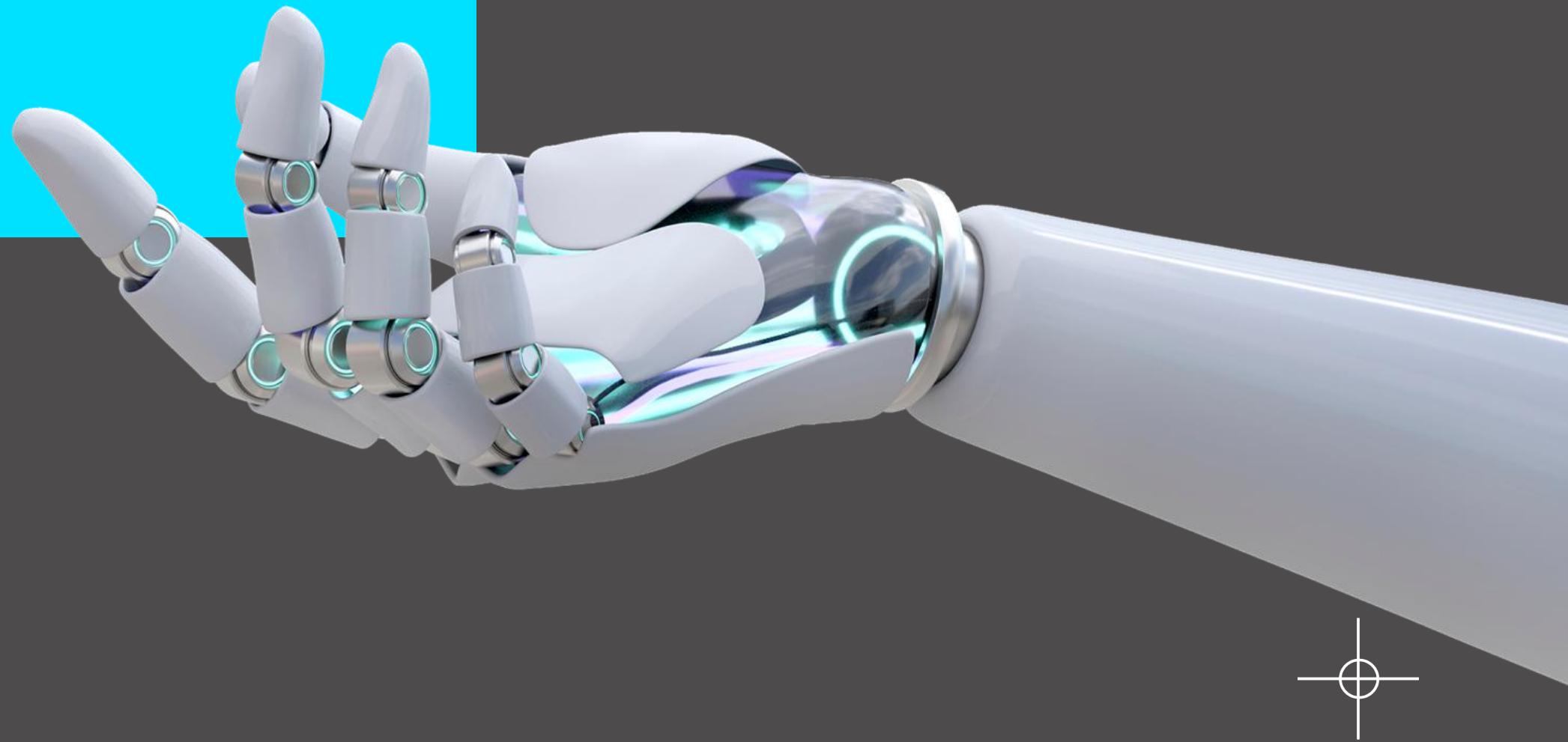
Comparar modelos lineales entrenados en el mismo dataset con distintos p . Detectar señales de sobreparametrización en un análisis exploratorio.

- Ajusta por grados de libertad, penalizando meter variables que solo mejoran por azar.
- Puede disminuir al agregar features irrelevantes, a diferencia de R^2 .
- Sigue siendo una métrica basada en el mismo set sobre el que se ajusta el modelo.

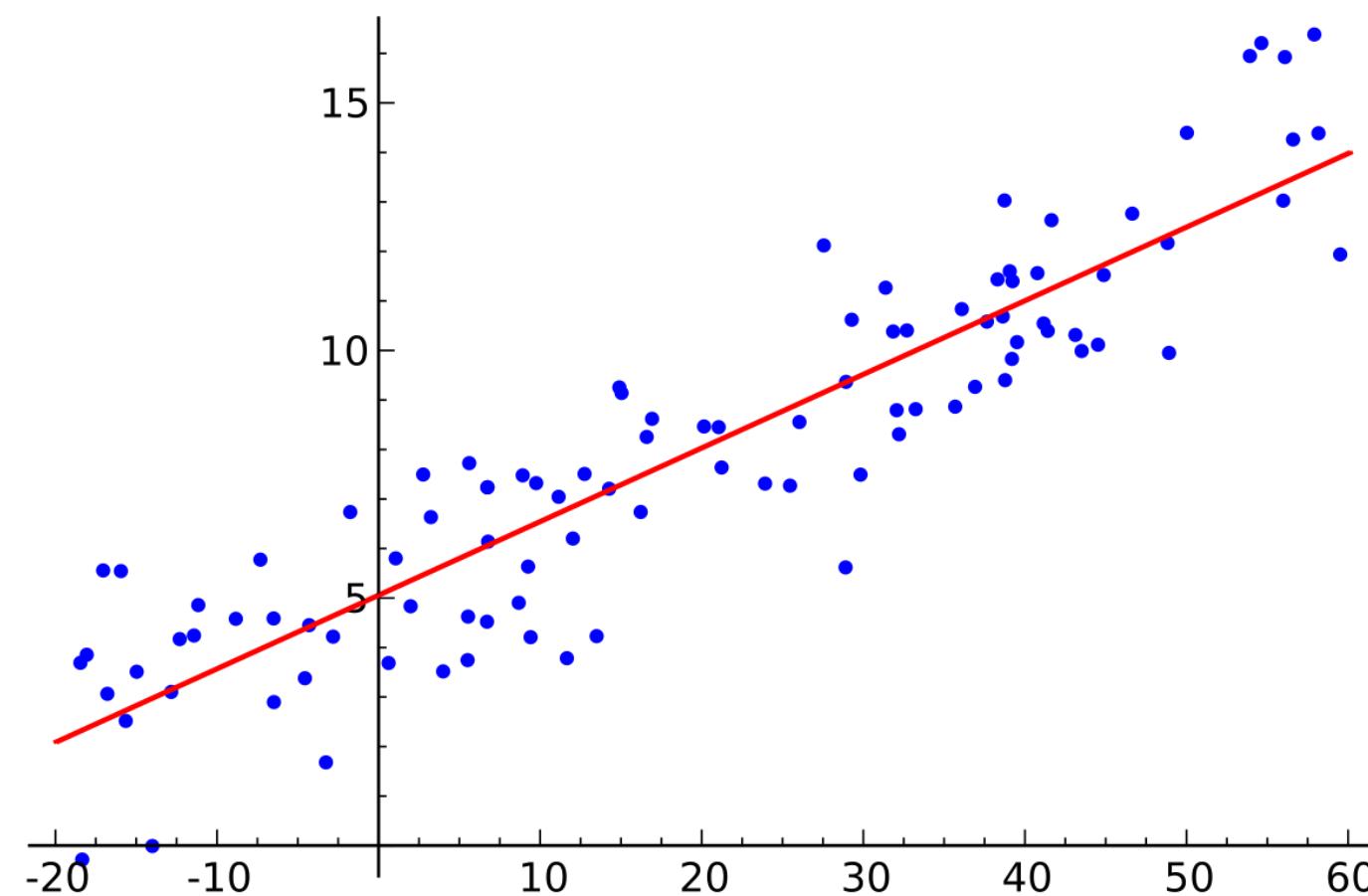


Bayesian linear model

Maximum a Posteriori (MAP). Regularization



Linear model



$$y_i = \beta_0 + \sum_{j=1}^p x_{ij}\beta_j + \varepsilon_i$$

number of features → p

noise term → $\varepsilon_i \stackrel{iid}{\sim} \mathbf{N}(0, \sigma^2)$

independence assumption → $\varepsilon_i \stackrel{iid}{\sim} \mathbf{N}(0, \sigma^2)$

noise level → σ^2

response → y_i

global intercept → β_0

feature j of observation i → x_{ij}

coefficient for feature j → β_j



Probability

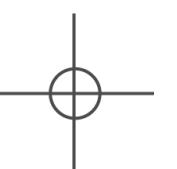
Enfoque Laplaciano (clásico)

Si todos los resultados posibles son igualmente probables, la probabilidad de un evento es:

$$P(A) = \frac{\text{\# casos favorables}}{\text{\# total de casos}}$$

Limitación

Solo funciona cuando los resultados equiprobables están claramente definidos.



Probability

Enfoque Laplaciano (clásico)

Si todos los resultados posibles son igualmente probables, la probabilidad de un evento es:

$$P(A) = \frac{\# \text{ casos favorables}}{\# \text{ total de casos}}$$

Enfoque frecuentista

La probabilidad de un evento es el límite de su frecuencia relativa cuando el experimento se repite muchas veces.

$$P(A) = \lim_{n \rightarrow \infty} \frac{N_A(n)}{n}$$

donde $N_A(n)$ es el número de veces que ocurre el evento A en n repeticiones

Ventaja

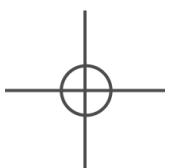
Intuitivo, empírico. Base del análisis estadístico clásico.

Limitación

Solo funciona cuando los resultados equiprobables están claramente definidos.

Limitación

Requiere un número infinito de repeticiones en teoría, y no define la probabilidad de eventos únicos.



Probability

Enfoque Laplaciano (clásico)

Si todos los resultados posibles son igualmente probables, la probabilidad de un evento es:

$$P(A) = \frac{\text{\# casos favorables}}{\text{\# total de casos}}$$

Enfoque frecuentista

La probabilidad de un evento es el límite de su frecuencia relativa cuando el experimento se repite muchas veces.

$$P(A) = \lim_{n \rightarrow \infty} \frac{N_A(n)}{n}$$

donde $N_A(n)$ es el número de veces que ocurre el evento A en n repeticiones

Ventaja

Intuitivo, empírico. Base del análisis estadístico clásico.

Limitación

Requiere un número infinito de repeticiones en teoría, y no define la probabilidad de eventos únicos.

Limitación

Solo funciona cuando los resultados equiprobables están claramente definidos.

Enfoque bayesiano

La probabilidad representa un grado de creencia o confianza en que ocurra un evento, dado cierto conocimiento.

La probabilidad es subjetiva, pero actualizable con nueva evidencia mediante el Teorema de Bayes:

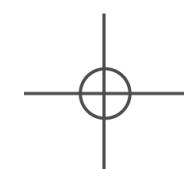
$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Ventaja

Permite modelar incertidumbre sobre eventos únicos o parámetros desconocidos.

Limitación

La elección del prior (probabilidad inicial) puede ser subjetiva.



Probability

Enfoque Laplaciano (clásico)

Si todos los resultados posibles son igualmente probables, la probabilidad de un evento es:

$$P(A) = \frac{\text{\# casos favorables}}{\text{\# total de casos}}$$

Enfoque frecuentista

La probabilidad de un evento es el límite de su frecuencia relativa cuando el experimento se repite muchas veces.

$$P(A) = \lim_{n \rightarrow \infty} \frac{N_A(n)}{n}$$

donde $N_A(n)$ es el número de veces que ocurre el evento A en n repeticiones

Ventaja

Intuitivo, empírico. Base del análisis estadístico clásico.

Limitación

Requiere un número infinito de repeticiones en teoría, y no define la probabilidad de eventos únicos.

Limitación

Solo funciona cuando los resultados equiprobables están claramente definidos.

Enfoque bayesiano

La probabilidad representa un grado de creencia o confianza en que ocurra un evento, dado cierto conocimiento.

La probabilidad es subjetiva, pero actualizable con nueva evidencia mediante el Teorema de Bayes:

$$P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)}$$

Ventaja

Permite modelar incertidumbre sobre eventos únicos o parámetros desconocidos.

Limitación

La elección del prior (probabilidad inicial) puede ser subjetiva.

Enfoque axiomático

La probabilidad es una medida matemática definida sobre un conjunto de eventos, y debe cumplir ciertos axiomas (llamados axiomas de Kolmogorov).

Se define un espacio de probabilidad como un triplete

$$(\Omega, \mathcal{F}, P)$$

Ventaja

Es el enfoque más general y riguroso, y sirve como base para toda la teoría moderna de probabilidad.

Limitación

Falta de interpretación empírica directa. No dice qué es la probabilidad, solo cómo debe comportarse una función que la representa.



Probability

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad:

1. Ω es el espacio muestral, es decir, el conjunto de todos los posibles resultados de un experimento aleatorio.
2. \mathcal{F} es una σ -álgebra sobre Ω , también llamada familia de eventos. Es un conjunto de subconjuntos (eventos) de Ω que cumple:
 - $\Omega \in \mathcal{F}$,
 - Si $A \in \mathcal{F}$, entonces su complemento $A^c \in \mathcal{F}$,
 - Si $A_1, A_2, A_3, \dots \in \mathcal{F}$, entonces $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.
3. \mathbb{P} es una medida de probabilidad, es decir, una función $\mathbb{P}: \mathcal{F} \rightarrow [0,1]$ tal que:
 - **No negatividad:** $\mathbb{P}(A) \geq 0, \forall A \in \mathcal{F}$.
 - **Normalización:** $\mathbb{P}(\Omega) = 1$,
 - **Aditividad numerable (σ -aditividad):** Si $\{A_i\}_{i=1}^{\infty} \subset \mathcal{F}$ es una colección de eventos mutuamente disjuntos, es decir $A_i \cap A_j = \emptyset$ para $i \neq j$, entonces:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

Axiomas de Kolmogorov



Probability

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

- $X: \Omega \rightarrow \mathbb{R}^d$, la cual representa la variable de entrada.
- $w: \Omega \rightarrow \mathbb{R}^d$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

donde:

Ω Es el espacio muestral abstracto que indexa toda la aleatoriedad del modelo.
 Ω Es el conjunto de posibles resultados elementales; sobre él viven las variables aleatorias.

X, w, ϵ Son funciones medibles definidas sobre Ω .



Bayesian model

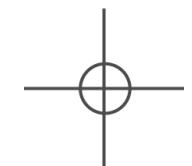
Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$



Bayesian model

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

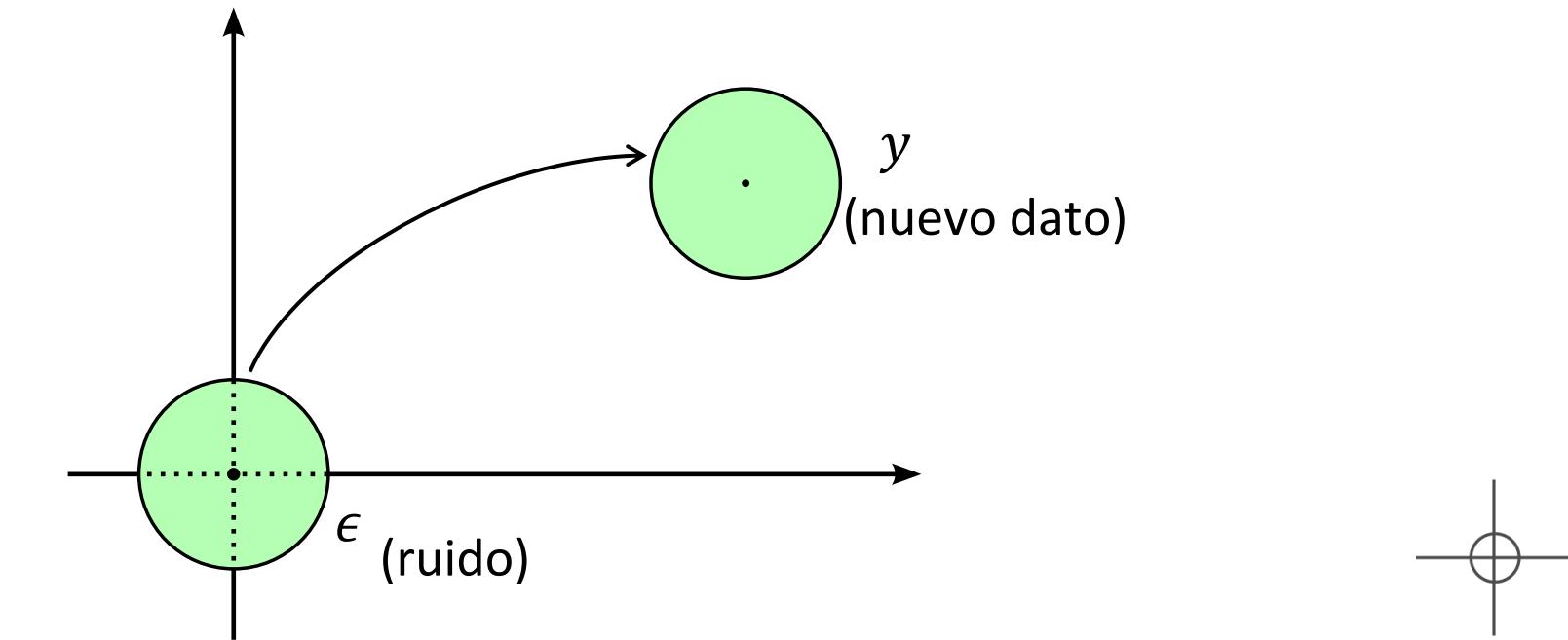
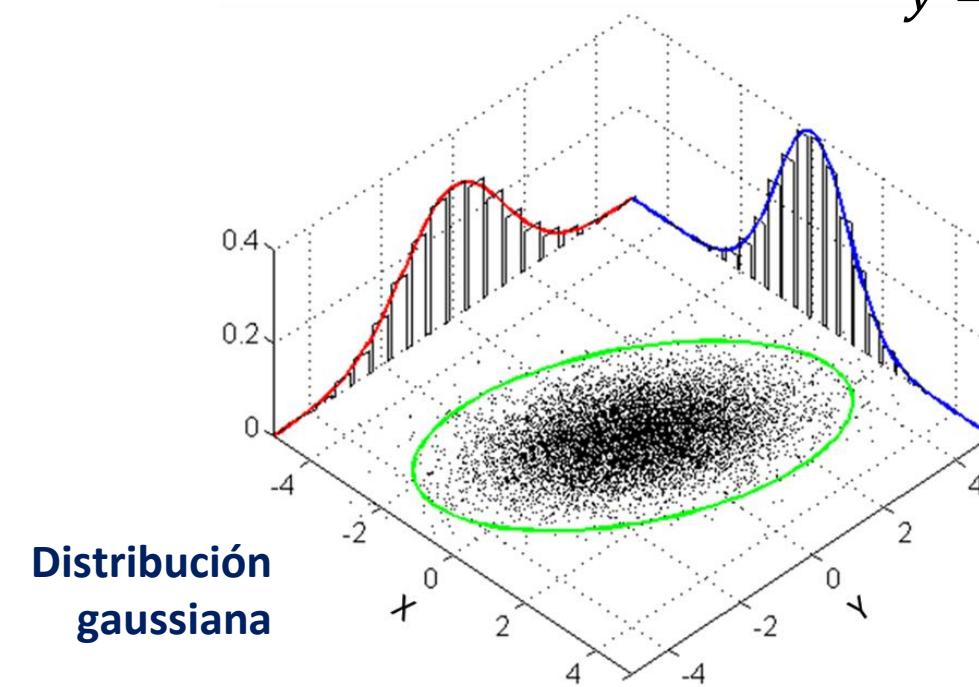
- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$

$$y = w^\top x + \epsilon$$



Bayesian model

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

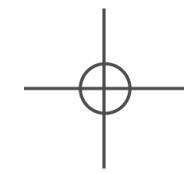
- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$

Aquí W es la variable de interés y es optimizado a partir de los datos X ; por otro lado, ϵ es la fuente exógena de ruido.



Bayesian model

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$

Aquí W es la variable de interés y es optimizado a partir de los datos X ; por otro lado, ϵ es la fuente exógena de ruido.

Fijamos $X(\omega) = x, w(\omega) = w$, entonces la variable aleatoria Y tiene distribución $p(Y|X(\omega) = x, w(\omega) = w)$.



Bayesian model

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$

Aquí W es la variable de interés y es optimizado a partir de los datos X ; por otro lado, ϵ es la fuente exógena de ruido.

Fijamos $X(\omega) = x, w(\omega) = w$, entonces la variable aleatoria Y tiene distribución $p(Y|X(\omega) = x, w(\omega) = w)$.

Entones:

$$p(y|x, w) \sim \mathcal{N}(\mu_y, \sigma_y^2)$$



Bayesian model

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$

Aquí W es la variable de interés y es optimizado a partir de los datos X ; por otro lado, ϵ es la fuente exógena de ruido.

Fijamos $X(\omega) = x, w(\omega) = w$, entonces la variable aleatoria Y tiene distribución $p(Y|X(\omega) = x, w(\omega) = w)$.

Entones:

$$p(y|x, w) \sim \mathcal{N}(\mu_y, \sigma_y^2) = \mathcal{N}(w^\top x, \sigma^2)$$

- $\mu_y = \mathbb{E}[y] = \mathbb{E}[w^\top x + n] = w^\top x + \mathbb{E}[n] = w^\top x + 0 = w^\top x$
- $\text{Var}(y) = \text{Var}(w^\top x + n) = 0 + \text{Var}(n) = \sigma^2$



Bayesian model

Sea (Ω, \mathcal{F}, P) un espacio de probabilidad donde definimos tres objetos probabilísticos (variables aleatorias):

- $X: \Omega \rightarrow \mathbb{R}^n$, la cual representa la variable de entrada.
- $W: \Omega \rightarrow \mathbb{R}^{n \times m}$, parámetros del modelo.
- $\epsilon: \Omega \rightarrow \mathbb{R}^m$, una variable aleatoria tal que $\epsilon \sim \mathcal{N}(0, \Sigma)$

El ruido aleatorio ϵ es independiente a la variable de entrada y los parámetros del modelo: $\epsilon \perp (X, W)$

Definimos la variable de salida mediante la **ecuación estructural** del modelo:

$$Y: \Omega \rightarrow \mathbb{R}^m, \quad Y(\omega) = W^\top(\omega)X(\omega) + \epsilon(\omega) \quad \text{modelo generativo}$$

Aquí W es la variable de interés y es optimizado a partir de los datos X ; por otro lado, ϵ es la fuente exógena de ruido.

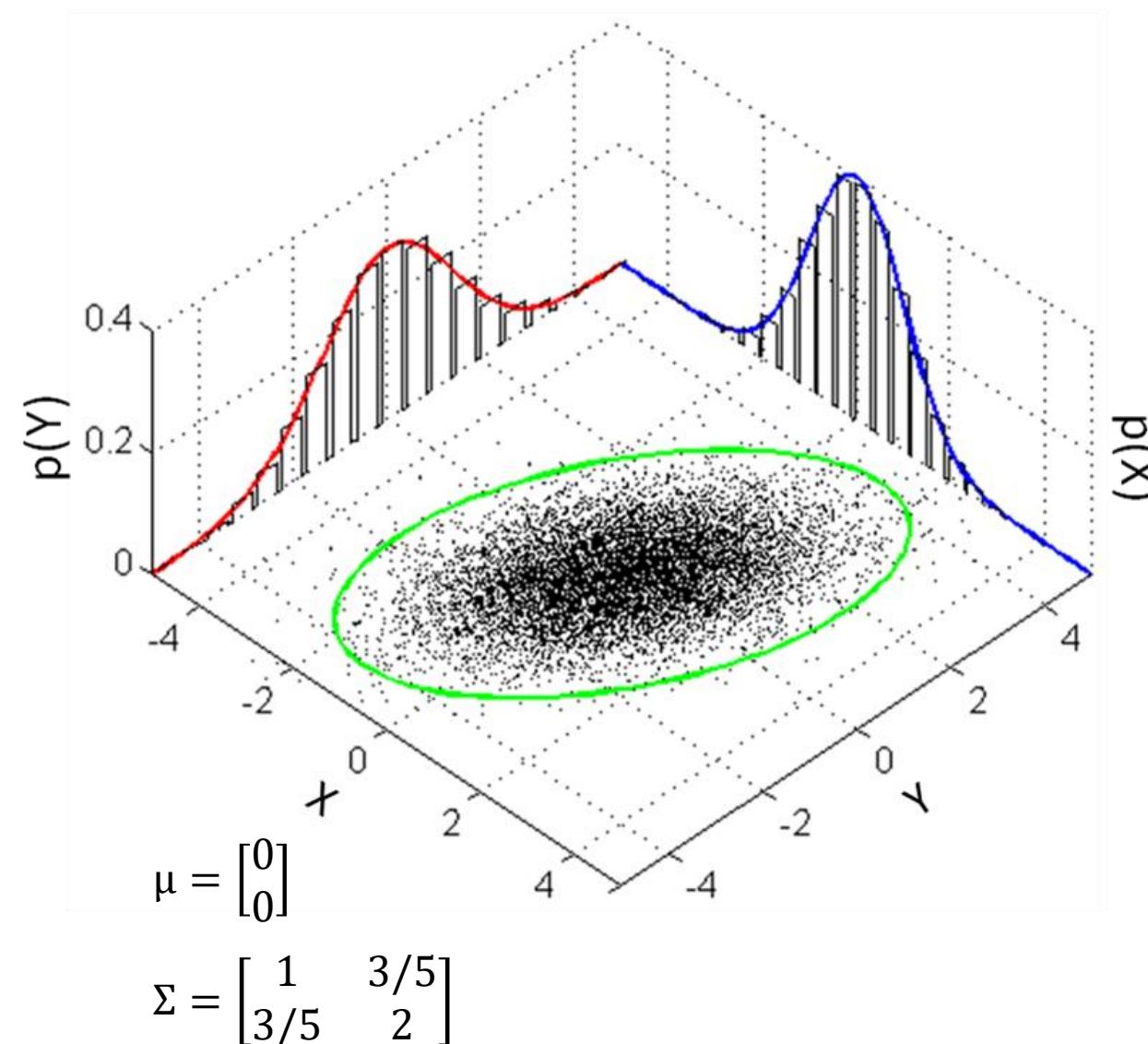
Fijamos $X(\omega) = x, w(\omega) = w$, entonces la variable aleatoria Y tiene distribución $p(Y|X(\omega) = x, w(\omega) = w)$.

Entones, para una variable:

$$p(y|x, w) \sim \mathcal{N}(w^\top x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-w^\top x)^2}{2\sigma^2}}$$



Multivariate normal distribution



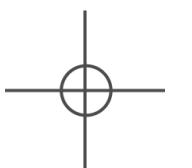
$$\mathcal{N}(x; \mu, \Sigma) = \frac{1}{\sqrt{\det(\Sigma)} (2\pi)^{\frac{d}{2}}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

donde $\mu \in \mathbb{R}^D$ es la media y $\Sigma > 0$ es la matriz de covarianzas.

- Las marginales y condicionales de una Gaussiana son Gaussianas.
- Completamente definida por primer y segundo orden.
- Máxima entropía entre todas las distribuciones con media y covarianza fijas.

Casos comunes:

- **Covarianza diagonal heterocedástica** $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_D^2)$
Variables no correlacionadas, escalas distintas.
- **Varianza isótropa** $\Sigma = \sigma^2 I$
Simetría rotacional, densidad depende solo de $\|x - \mu\|_2$.



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados los datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(w|D)$$

Es la distribución que cuantifica nuestra incertidumbre sobre el valor de w , luego de incorporar la evidencia proporcionada por los datos



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados los datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(w|D)$$

Usando el teorema de Bayes, la distribución posterior se puede expresar como:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

- $p(D|w)$: **Verosimilitud**, mide qué tan bien los datos D son explicados por los parámetros w .
- $p(w)$: **Prior**, refleja nuestras creencias iniciales sobre w antes de observar los datos.
- $p(D)$: **Evidencia**, es un factor de normalización que asegura que $p(w|D)$ sea una distribución de probabilidad válida.



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados los datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(w|D)$$

Usando el teorema de Bayes, la distribución posterior se puede expresar como:

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)}$$

- $p(D|w)$: **Verosimilitud**, mide qué tan bien los datos D son explicados por los parámetros w .
- $p(w)$: **Prior**, refleja nuestras creencias iniciales sobre w antes de observar los datos.
- $p(D)$: **Evidencia**, es un factor de normalización que asegura que $p(w|D)$ sea una distribución de probabilidad válida.

Reemplazando:

$$\hat{w} = \arg \max_w p(w|D) = \arg \max_w \frac{p(D|w)p(w)}{p(D)}$$

Dado que la evidencia $p(D)$ no depende de w , entonces podemos simplificar:

$$\hat{w} = \arg \max_w p(D|w)p(w)$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = p((x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) | w) = \prod_{i=1}^N p((x_i, y_i) | w)$$

Asumiendo que los datos son independientes.



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = p((x_1, y_1), (x_2, y_2), \dots, (x_N, y_N) | w) = \prod_{i=1}^N p((x_i, y_i) | w)$$

Desarrollamos $p((x_i, y_i) | w)$:

$$p((x_i, y_i) | w) = p(x_i | w)p(y_i | x_i, w) = p(x_i)p(y_i | x_i, w)$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = \prod_{i=1}^N p(x_i) p(y_i|x_i, w) = \prod_{i=1}^N p(x_i) \prod_{i=1}^N p(y_i|x_i, w)$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = \prod_{i=1}^N p(x_i) p(y_i|x_i, w) = \prod_{i=1}^N p(x_i) \prod_{i=1}^N p(y_i|x_i, w)$$

Aplicamos logaritmo a $p(D|w)$:

$$\log p(D|w) = \sum_{i=1}^N \log p(x_i) + \sum_{i=1}^N \log p(y_i|x_i, w)$$

Nota que $\log p(x_i)$ es independiente de w , por lo que participa en MAP.

Además: $p(y|x, w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-w^T x)^2}{2\sigma^2}}$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = \prod_{i=1}^N p(x_i) p(y_i|x_i, w) = \prod_{i=1}^N p(x_i) \prod_{i=1}^N p(y_i|x_i, w)$$

Aplicamos logaritmo a $p(D|w)$:

$$\begin{aligned} \log p(D|w) &= \sum_{i=1}^N \log p(x_i) + \sum_{i=1}^N \log p(y_i|x_i, w) \\ &= k_1 + \sum_{i=1}^N \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w_i^\top x_i)^2}{2\sigma^2}} \right] \\ &= k_1 - \frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w_i^\top x_i)^2 \end{aligned}$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = \prod_{i=1}^N p(x_i) p(y_i|x_i, w) = \prod_{i=1}^N p(x_i) \prod_{i=1}^N p(y_i|x_i, w)$$

Aplicamos logaritmo a $p(D|w)$:

$$\begin{aligned} \log p(D|w) &= \sum_{i=1}^N \log p(x_i) + \sum_{i=1}^N \log p(y_i|x_i, w) \\ &= k_1 + \sum_{i=1}^N \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w_i^\top x_i)^2}{2\sigma^2}} \right] \\ &= k_1 - \frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w_i^\top x_i)^2 \end{aligned}$$

Introducimos un prior gaussiano sobre w :

$$p(w) = \frac{1}{(2\pi\tau^2)^{d/2}} e^{-\frac{\|w\|^2}{2\tau^2}}$$

Aplicamos logaritmo a $p(w)$:

$$\log p(w) = -\frac{d}{2} \log(2\pi\tau^2) - \frac{\|w\|^2}{2\tau^2}$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w p(D | w) p(w)$$

Expandimos:

$$p(D|w) = \prod_{i=1}^N p(x_i) p(y_i|x_i, w) = \prod_{i=1}^N p(x_i) \prod_{i=1}^N p(y_i|x_i, w)$$

Aplicamos logaritmo a $p(D|w)$:

$$\begin{aligned} \log p(D|w) &= \sum_{i=1}^N \log p(x_i) + \sum_{i=1}^N \log p(y_i|x_i, w) \\ &= k_1 + \sum_{i=1}^N \log \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - w_i^\top x_i)^2}{2\sigma^2}} \right] \\ &= k_1 - \frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w_i^\top x_i)^2 \end{aligned}$$

Introducimos un prior gaussiano sobre w :

$$p(w) = \frac{1}{(2\pi\tau^2)^{d/2}} e^{-\frac{\|w\|^2}{2\tau^2}}$$

Aplicamos logaritmo a $p(w)$:

$$\log p(w) = -\frac{d}{2} \log(2\pi\tau^2) - \frac{\|w\|^2}{2\tau^2}$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \max_w \left[-\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w_i^\top x_i)^2 - \frac{\|w\|^2}{2\tau^2} \right]$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \min_w \left[\frac{1}{2\sigma^2} \sum_{i=1}^N (y_i - w_i^\top x_i)^2 + \frac{\|w\|^2}{2\tau^2} \right]$$



Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

$$\hat{w} = \arg \min_w \frac{1}{2\sigma^2} \left[\sum_{i=1}^N (y_i - w_i^\top x_i)^2 + \frac{\sigma^2}{\tau^2} \|w\|^2 \right]$$

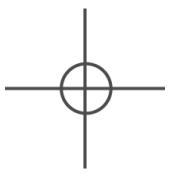
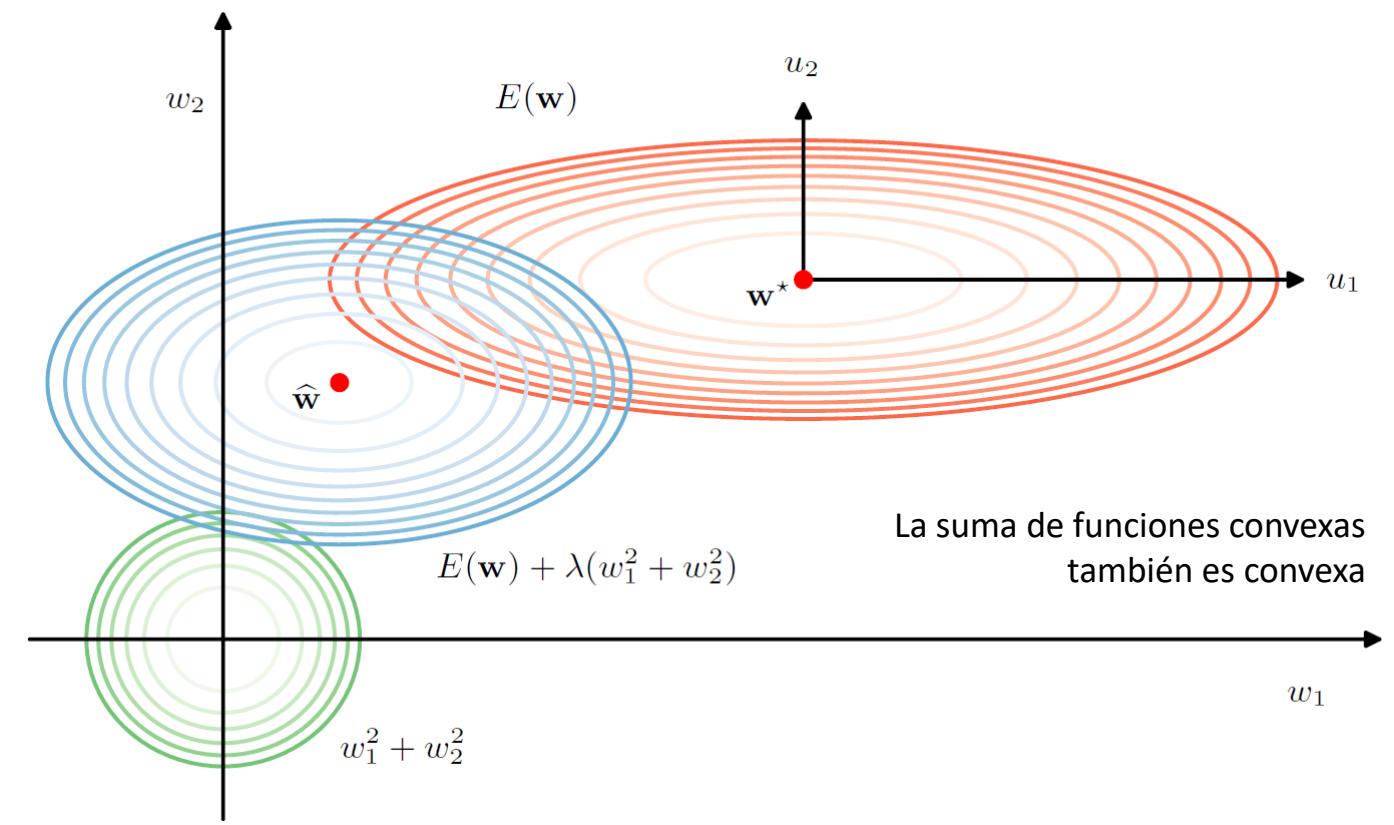


Maximum a Posteriori (MAP)

Maximum a Posteriori (MAP) es un criterio estadístico que buscar los parámetros w que son más probables dados un conjunto de datos observados D . Buscamos:

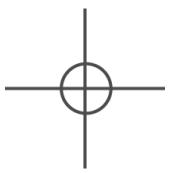
$$\hat{w} = \arg \min_w \left[\sum_{i=1}^N (y_i - w_i^\top x_i)^2 + \lambda \|w\|^2 \right], \quad \text{donde } \lambda = \frac{\sigma^2}{\tau^2} \text{ controla la fuerza de la regularización.}$$

Función Loss Término regularizador



Función Loss

Distribución de errores	Densidad de probabilidad	Función Loss	Propiedades	Aplicaciones típicas
Gaussiana ($\mathcal{N}(0, \sigma^2)$)	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right)$	$\sum_i (y_i - w^\top x_i)^2$ Mínimos cuadrados	Penaliza errores cuadráticamente, sensible a valores atípicos.	Regresión estándar, modelos lineales.
Laplaciano (Laplace($0, b$))	$\frac{1}{2b} \exp\left(-\frac{ x }{b}\right)$	$\sum_i y_i - w^\top x_i $ Error absoluto medio	Penaliza errores de forma lineal. Es más robusta frente a outliers.	Regresión robusto, situaciones con outliers moderados, análisis robusto en economía.
Cauchy (Cauchy($0, \gamma$))	$\frac{1}{\pi\gamma \left(1 + \frac{x^2}{\gamma^2}\right)}$	$\sum_i \ln\left(1 + \frac{(y_i - w^\top x_i)^2}{\gamma^2}\right)$ Cauchy o Lorentzian loss	Colas muy pesadas: alta robustez a valores atípicos.	Regresión muy robusta, visión por computadora con grandes outliers



Regularization

Distribución a priori	Densidad de probabilidad $p(w_j)$	Termino de regularización	Propiedades	Aplicaciones típicas
Gaussiana ($\mathcal{N}(0, \tau^2)$)	$\frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{w_j^2}{2\tau^2}\right)$	$\lambda\ w\ _2^2$ ℓ_2 regularización	Penaliza magnitudes grandes de w . Genera soluciones suaves, sin anular pesos a cero de forma exacta. Convexidad fuerte.	Regresión Ridge. Modelos lineales regulares. No induce sparsity.
Laplaciana (Laplace($0, b$))	$\frac{1}{2b} \exp\left(-\frac{ w_j }{b}\right)$	$\lambda\ w\ _1$ ℓ_1 regularización	Favorece sparsidad (algunos w_j se vuelven exactamente cero). Menos suavidad que la Gaussiana, pero más selectiva (selección de características).	Regresión Lasso. Situaciones donde se busca desactivar pesos irrelevantes.
Cauchy (Cauchy($0, \gamma$))	$\frac{1}{\pi\gamma\left(1 + \frac{w_j^2}{\gamma^2}\right)}$	$\sum_j \log\left(1 + \frac{w_j^2}{\gamma^2}\right)$ Cauchy regularization	Penalización suave para valores grandes de w . En ocasiones se usa como prior robusto, evitando que pesos grandes queden demasiado penalizados.	Regularización robusta, situaciones con colas largas.
t-Student	$\left(1 + \frac{w_j^2}{v}\right)^{-\frac{v+1}{2}}$	$\sum_j \log\left(1 + \frac{w_j^2}{v}\right)$ t -Student regularization	Robusta: colas más pesadas que la Gaussiana, pero no tanto como la Cauchy (depende de v). Para $v \rightarrow \infty$ se approxima a la Gaussiana, mientras que cuando $v \rightarrow 1$ se approxima a Cauchy.	Situaciones con outliers moderados.



Ridge Regression

Técnica de regresión utilizada para datos que sufren de multicolinealidad (cuando las variables independientes están altamente correlacionadas).

En la regresión lineal estándar (OLS), el objetivo es minimizar la suma de los errores al cuadrado. Sin embargo, si las variables de entrada están muy correlacionadas entre sí:

- Las estimaciones de los coeficientes pueden cambiar erráticamente en respuesta a pequeños cambios en los datos.
- El modelo puede volverse demasiado complejo y ajustarse al ruido en lugar de a la señal.
- El modelo tiene alta varianza.

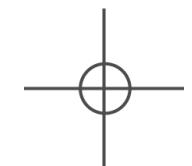
Ridge Regression resuelve esto añadiendo un término de penalización a la función de costo. A esto se le llama Regularización L2.

$$J(\mathbf{w}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p w_j^2$$

Es el error estándar (lo que queremos minimizar para predecir bien).

Es la penalización. Suma los cuadrados de los coeficientes.

- Si $\lambda = 0$: El término de penalización desaparece y volvemos a una regresión lineal estándar (OLS).
- Si $\lambda \rightarrow \infty$: La penalización es tan alta que fuerza a los coeficientes a acercarse mucho a cero (una línea plana).
- El objetivo es encontrar un λ intermedio que equilibre el ajuste de los datos y la simplicidad del modelo.



Lasso Regression

(Least Absolute Shrinkage and Selection Operator)

Al igual que Ridge, busca evitar el overfitting reduciendo la magnitud de los coeficientes, pero utiliza Regularización L1.

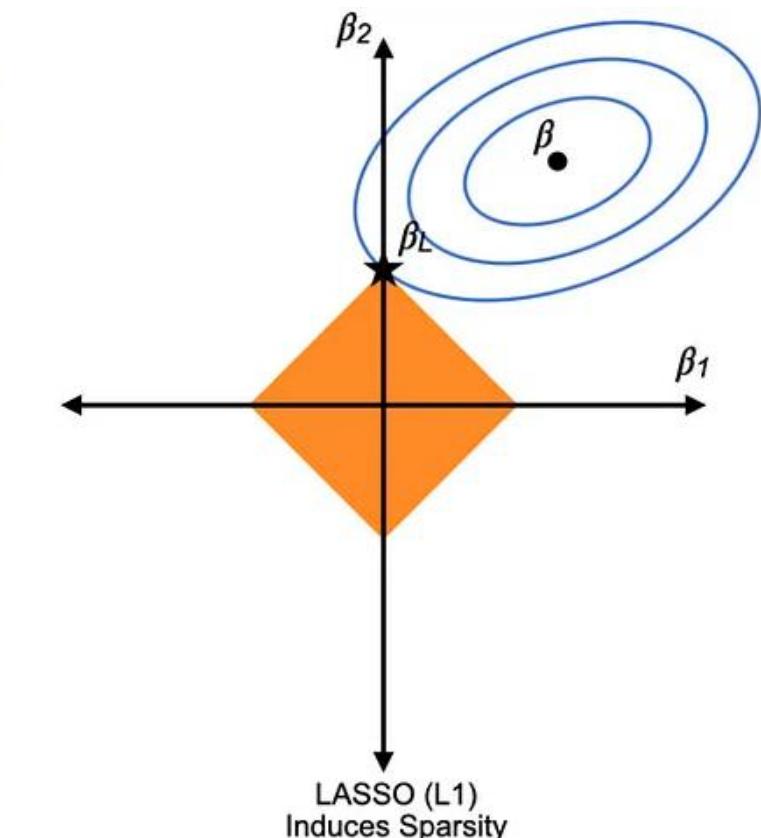
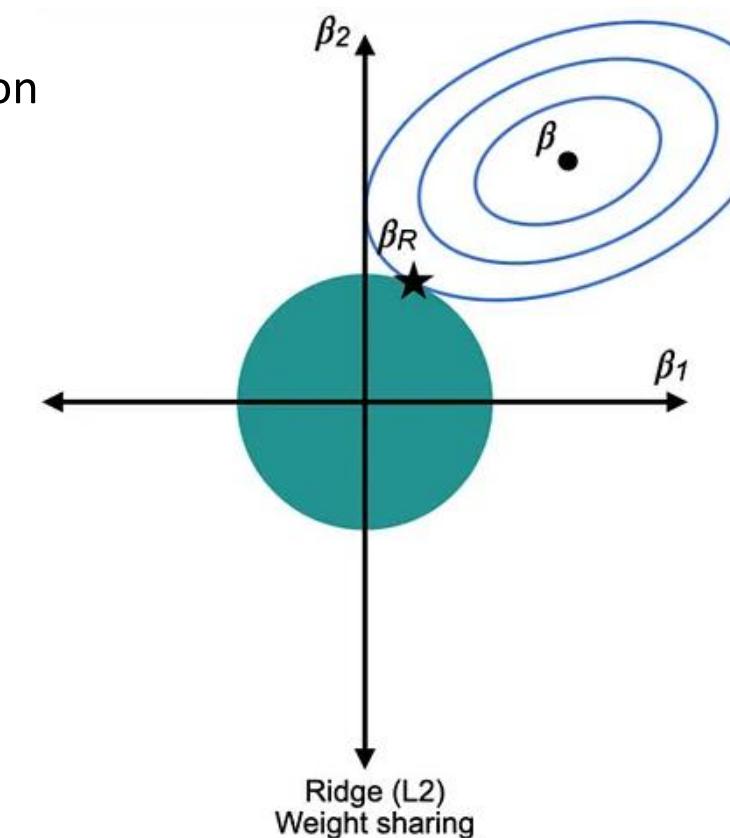
En lugar de elevar los coeficientes al cuadrado (como en Ridge), Lasso suma sus valores absolutos.

$$J(\mathbf{w}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^p |w_j|$$

Lasso puede reducir los coeficientes exactamente a cero, es decir, puede seleccionar de características.

En Ridge: Los coeficientes se vuelven infinitesimalmente pequeños (por ejemplo: 0.00001), pero nunca desaparecen por completo. El modelo sigue usando todas las variables.

En Lasso: Si λ es suficientemente alto, Lasso forzará a los coeficientes de las variables menos importantes a ser 0. Como resultado tenemos un modelo disperso (sparse model).



Linear model

Ridge Regression

$$\min_w \sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_2^2$$

Gaussian likelihood **Gaussian prior**

Lasso Regression

$$\min_w \sum_i (y_i - w^\top x_i)^2 + \lambda \|w\|_1$$

Gaussian likelihood **Laplace prior**

Logistic Regression (L2)

$$\min_w - \sum_i [y_i \log \sigma(w^\top x_i) + (1 - y_i) \log(1 - \sigma(w^\top x_i))] + \lambda \|w\|_2^2$$

Bernoulli likelihood **Gaussian prior**

Logistic Regression (L1)

$$\min_w - \sum_i [y_i \log \sigma(w^\top x_i) + (1 - y_i) \log(1 - \sigma(w^\top x_i))] + \lambda \|w\|_1$$

Bernoulli likelihood **Laplace prior**



ElasticNet

Lasso es excelente para eliminar variables, pero tiene un defecto grave cuando hay alta multicolinealidad (variables muy correlacionadas entre sí):

- Si dos variables están altamente correlacionadas, Lasso tiende a elegir una arbitrariamente y reducir la otra a cero.
- Esto es inestable y pierde información si ambas variables son importantes en conjunto.

Elastic Net soluciona esto: permite seleccionar grupos de variables correlacionadas juntas (efecto de agrupamiento) en lugar de elegir solo una.

$$J(\mathbf{w}) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^p |w_j| + \lambda_2 \sum_{j=1}^p w_j^2$$

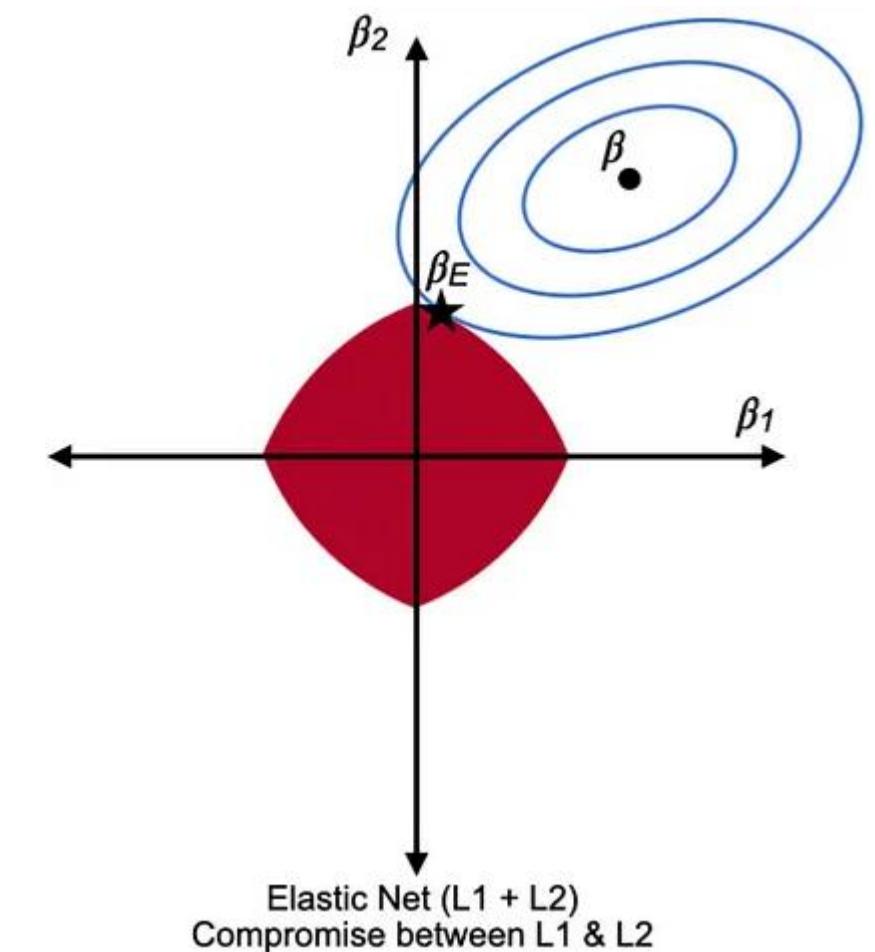
Error (MSE) Penalización L1 Penalización L2

En la práctica, no ajustas λ_1 y λ_2 por separado. Usas dos parámetros más intuitivos:

α : La fuerza total de la regularización ($\lambda_1 + \lambda_2$).

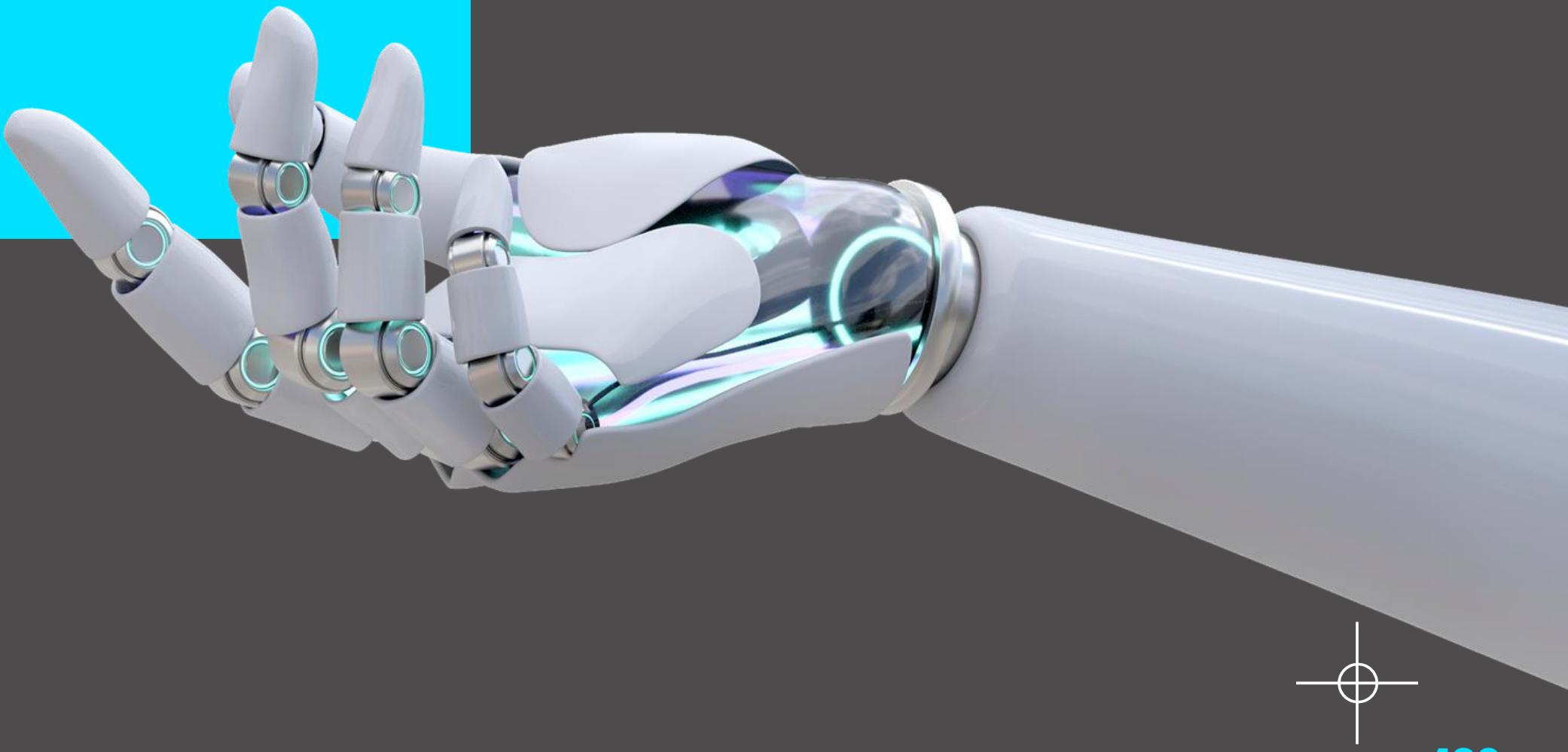
ℓ_1 ratio (ρ): El balance entre Lasso y Ridge.

- ℓ_1 ratio = 1: Es Lasso puro.
- ℓ_1 ratio = 0: Es Ridge puro.
- ℓ_1 ratio = 0.5: Mitad y mitad.



Nonlinear Regression

Polynomial regression. Truly nonlinear.



Polynomial regression

Modela la relación entre una variable independiente x y una variable dependiente y como un polinomio de grado d .

$$\hat{y} = w_0 + w_1x + w_2x^2 + \cdots + w_dx^d$$

donde:

- \hat{y} es la predicción del modelo.
- x es la variable independiente.
- w_0, w_1, \dots, w_d son los coeficientes del modelo que se deben estimar.
- d es el grado del polinomio, que determina la complejidad de la curva.

La regresión polinomial equivale a una regresión lineal realizada en un espacio de características no lineales. Esto se logra transformando la característica original x en un conjunto de nuevas características que son potencias de x .

Definimos el vector de características: $\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^d \end{bmatrix}$ entonces el modelo es: $\hat{y} = w^\top \phi(x)$

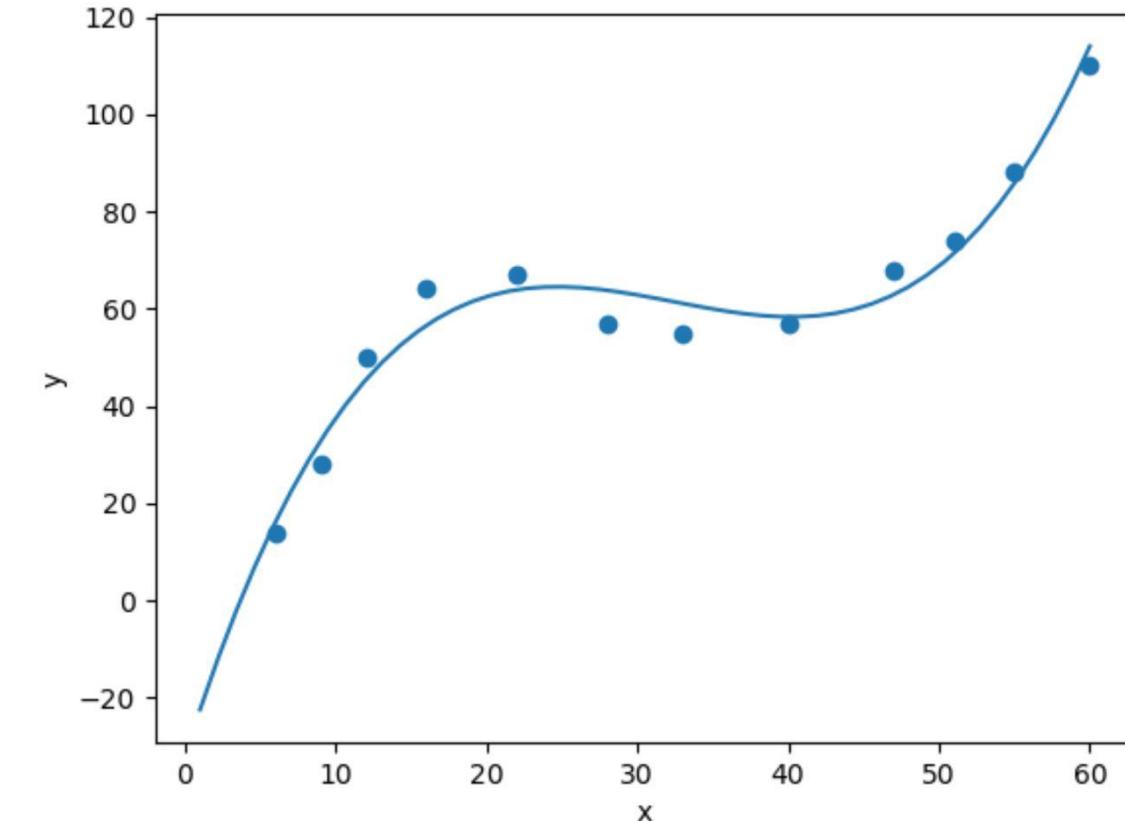
Para un conjunto de n puntos de datos $\{(x_i, y_i)\}_{i=1}^n$, podemos construir la matriz de diseño X , donde cada fila corresponde al vector de características de una observación:

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^d \\ 1 & x_2 & x_2^2 & \cdots & x_2^d \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^d \end{bmatrix}$$

La matriz X tiene dimensiones $n \times (d + 1)$.

Con esta matriz, el modelo para todas las observaciones se puede expresar como: $\hat{y} = Xw$

donde \hat{y} es ahora un vector de predicciones de tamaño $n \times 1$ y w es el vector de coeficientes de tamaño $(d + 1) \times 1$.



Polynomial regression

Modela la relación entre una variable independiente x y una variable dependiente y como un polinomio de grado d .

$$\hat{y} = Xw$$

Al igual que en la regresión lineal estándar, los coeficientes w se pueden estimar utilizando varios métodos:

OLS: $\hat{w} = (X^T X)^{-1} X^T y$

Este método es eficiente para conjuntos de datos pequeños a medianos y proporciona la solución óptima si la matriz $X^T X$ es invertible.

Gradient Descent:

Se suele utilizar cuando el número de características $(d + 1)$ o el número de muestras n es muy grande, lo que hace que la inversión de la matriz en OLS sea computacionalmente costosa.

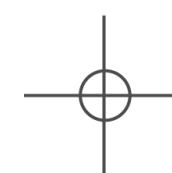
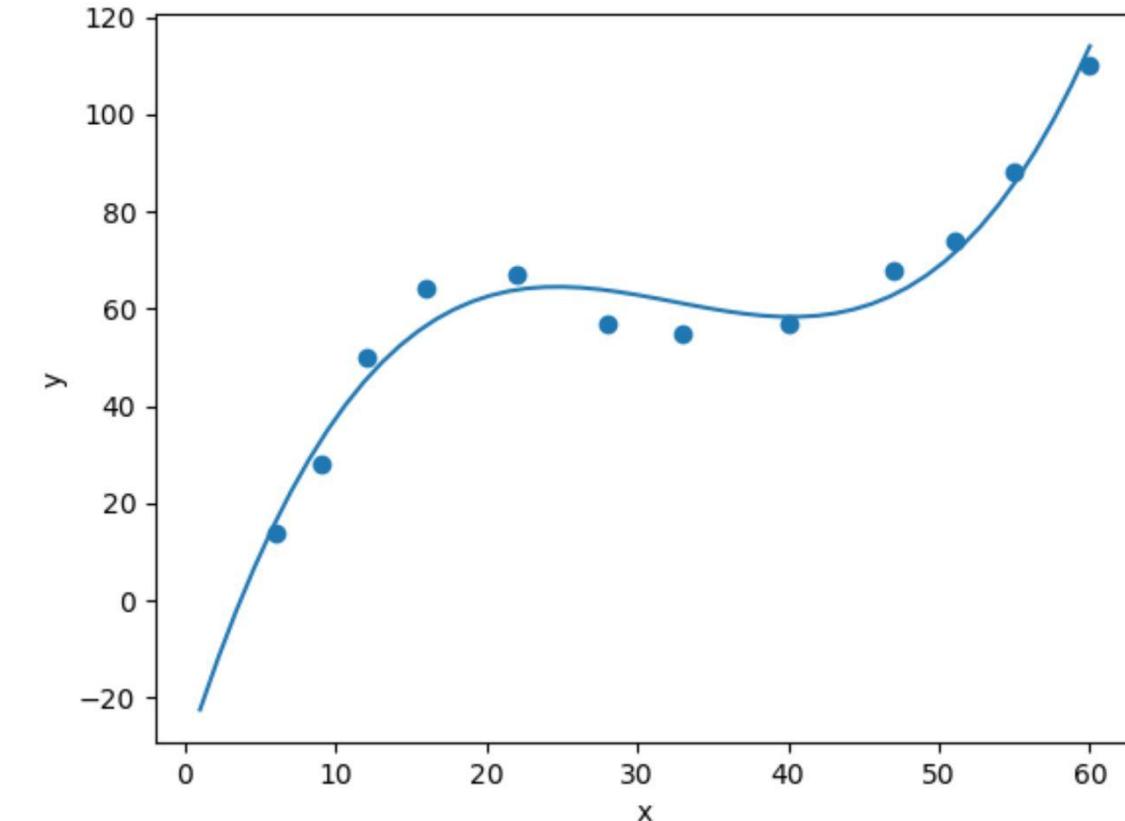
Sin embargo, este enfoque presenta desafíos importantes:

Overfitting: Si el grado d del polinomio es demasiado alto, el modelo puede volverse excesivamente complejo y ajustarse al ruido de los datos de entrenamiento en lugar de capturar la tendencia subyacente.

Solución: Utilizar técnicas de regularización (como Ridge o Lasso) para penalizar coeficientes grandes y usar validación cruzada para seleccionar el grado d óptimo.

Multicolinealidad: Las características x, x^2, \dots, x^d a menudo están altamente correlacionadas, especialmente si el rango de valores de x es grande o si el grado d es alto. Esto puede hacer que la matriz $X^T X$ sea casi singular, lo que lleva a estimaciones de coeficientes inestables y poco fiables.

Solución: Estandarizar (escalar y centrar) la variable x : $x_{std} = \frac{x - \mu}{\sigma}$ donde μ es la media y σ es la desviación estándar de x .



Multivariate Polynomial Regression

Para un polinomio de grado $d = 2$ y dos variables (x_1, x_2) , la ecuación de regresión se expande de la siguiente manera:

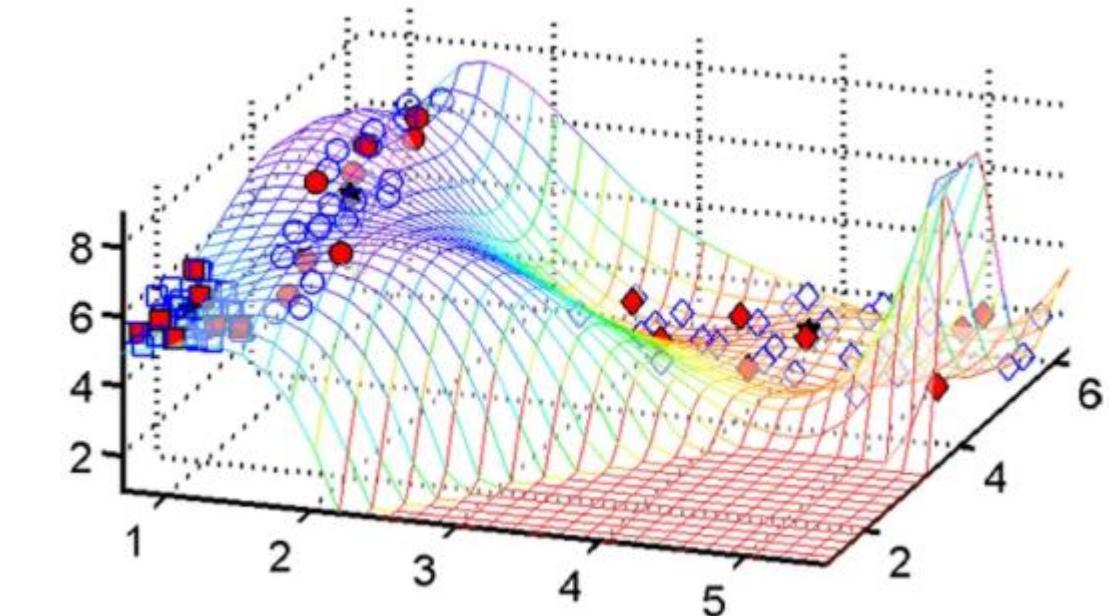
$$\hat{y} = w_0 + \textcolor{red}{w_1 x_1 + w_2 x_2} + \textcolor{blue}{w_3 x_1^2 + w_4 x_2^2} + \textcolor{teal}{w_5 x_1 x_2}$$

Términos Lineales	Términos Cuadráticos	Término de Interacción
-------------------	----------------------	------------------------

Se observa que la complejidad aumenta porque se deben considerar todas las combinaciones posibles de las variables cuyo grado total sea menor o igual a d .

Términos propios (x_i^2): Definen la curvatura de la superficie respecto a un eje específico (cónica o convexa).

Términos de interacción ($x_i x_j$): Definen cómo se tuerce la superficie. Indican que el efecto de la variable x_i sobre y depende del valor actual de x_j .



Explosión Combinatoria

El principal desafío en la regresión polinomial multivariada es el crecimiento rápido del número de características (y por tanto, de coeficientes w a estimar).

El número total de características N para k variables y grado d es:

$$N = \binom{k+d}{d} = \frac{(k+d)!}{k! d!}$$

- 2 variables, grado 2: $\binom{2+2}{2} = 6$ coeficientes.
- 10 variables, grado 2: $\binom{10+2}{2} = 66$ coeficientes.
- 10 variables, grado 5: $\binom{10+5}{5} = 3003$ coeficientes.

Esto incrementa drásticamente el costo computacional y el riesgo de overfitting si no se tiene un número suficiente de datos n respecto a las dimensiones N .

Nonlinear regression

Generalizamos: para los datos $\{(x_i, y_i)\}_{i=1}^n$ con $x_i \in \mathbb{R}^p$, definimos el modelo: $y = f(x) + \epsilon$ con $\mathbb{E}[\epsilon] = 0$

Linear-in-parameters

Un modelo es lineal en los parámetros si la función de predicción $f(x)$ se construye como una combinación lineal de funciones base fijas $\phi(x)$, aunque estas funciones base sean no lineales respecto a la entrada x .

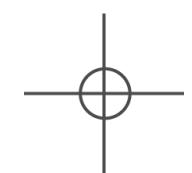
$$f(x) = \sum_{j=1}^m w_j \phi_j(x) = \mathbf{w}^\top \boldsymbol{\Phi}(x)$$

donde:

- \mathbf{w} : Vector de pesos.
- $\boldsymbol{\Phi}(x)$: Vector de funciones base (transformaciones predefinidas de x)

Ejemplos:

• Polinomios	x^j	Tendencias globales suaves y continuas.
• Fourier / wavelets	$\sin(\omega x), \cos(\omega x)$	Datos con comportamiento periódico o estacional (series de tiempo, audio).
• Radial Basis Functions (RBF)	$\exp\left(-\frac{\ x - \mu_j\ ^2}{2\sigma^2}\right)$	Modelado local; la influencia de un punto decae con la distancia.
• Splines / B-splines	Polinomios por tramos	Datos con cambios abruptos de comportamiento en diferentes regiones del dominio.



Nonlinear regression

Generalizamos: para los datos $\{(x_i, y_i)\}_{i=1}^n$ con $x_i \in \mathbb{R}^p$, definimos el modelo: $y = f(x) + \epsilon$ con $\mathbb{E}[\epsilon] = 0$

Nonlinear-in-parameters (Truly nonlinear)

La función de predicción $f(x; \theta)$ no puede factorizarse como un producto escalar $\theta^\top \phi(x)$.

El gradiente $\nabla_\theta f(x)$ depende de los propios parámetros θ .

Ejemplos:

- **Modelos tipo saturación** $y = \alpha(1 - e^{-\beta x})$
- **Redes neuronales** $y = W_2 \cdot \sigma(W_1 x + b)$

La composición de funciones no lineales $\sigma(\cdot)$ hace que el modelo sea no lineal respecto a los pesos internos W_1 .

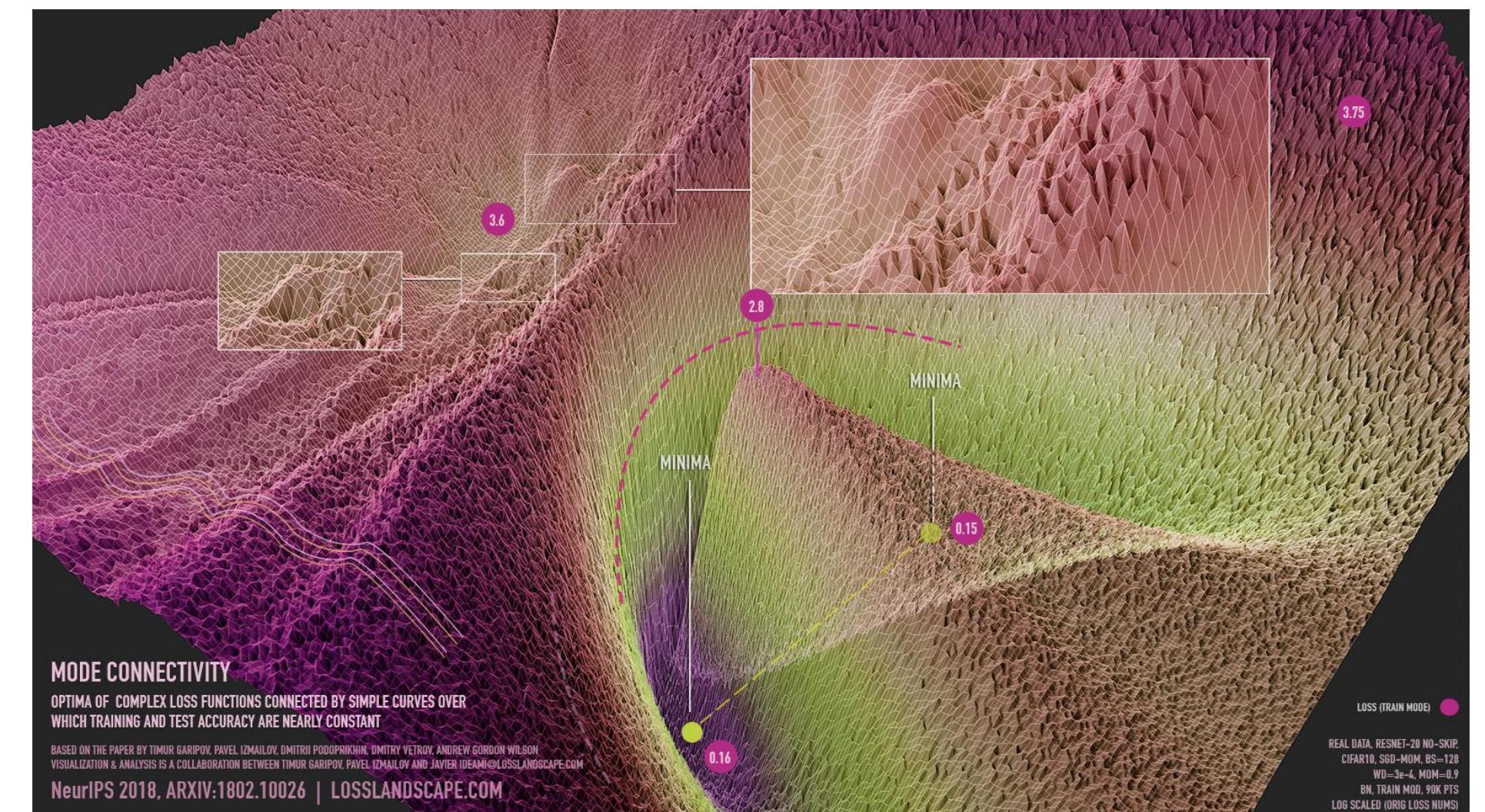
Sin Solución Cerrada:

No existe una fórmula analítica directa para encontrar el óptimo global.

Pérdida de Convexidad:

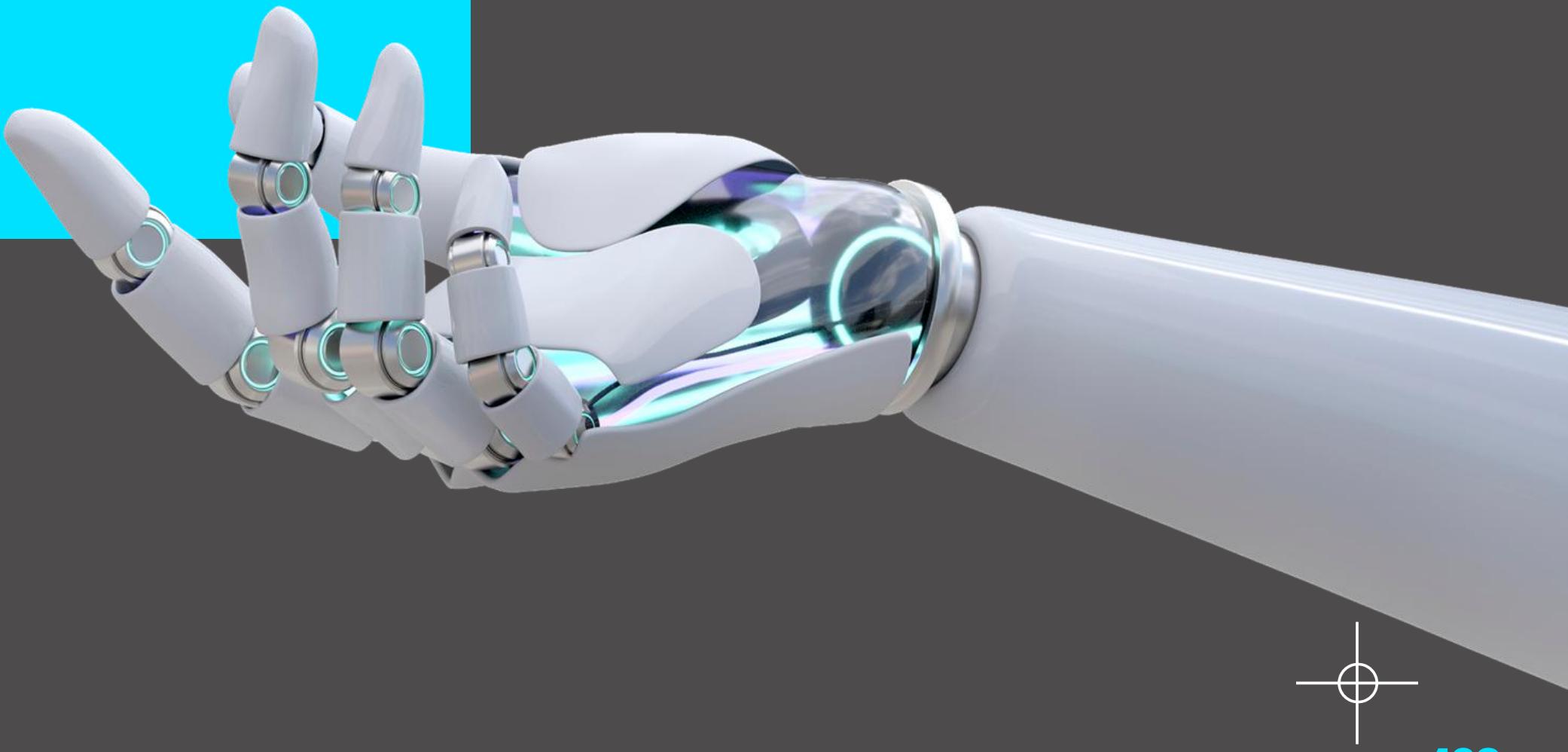
La superficie de error (Loss Landscape) deja de ser convexa.

- Presencia de mínimos locales y puntos de silla (saddle points).
- El resultado depende de la inicialización de los parámetros.



Learning theory

Bias-variance tradeoff. Manifold hypothesis.



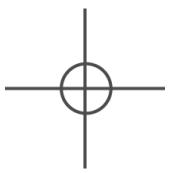
Learning theory

Expected risk

Dada una función $h: \mathcal{X} \rightarrow \mathcal{Y}$ que pertenece a un espacio de hipótesis \mathcal{H} ; una función de pérdida $\mathcal{L}: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$; y una base de datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ con distribución de probabilidad p ; definimos el riesgo esperado de f como:

$$\mathcal{R}(h) = \mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}(h(X), Y)] = \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{L}(h(x), y) \partial p(x, y)$$

- \mathcal{X} : Espacio de entrada.
- \mathcal{Y} : Espacio de salida.



Learning theory

Expected risk

Dada una función $h: \mathcal{X} \rightarrow \mathcal{Y}$ que pertenece a un espacio de hipótesis \mathcal{H} ; una función de pérdida $\mathcal{L}: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$; y una base de datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ con distribución de probabilidad p ; definimos el riesgo esperado de f como:

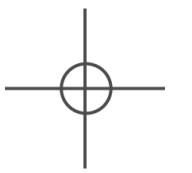
$$\mathcal{R}(h) = \mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}(h(X), Y)] = \int_{\mathcal{X} \times \mathcal{Y}} \mathcal{L}(h(x), y) \partial p(x, y)$$

- \mathcal{X} : Espacio de entrada.
- \mathcal{Y} : Espacio de salida.

Empirical risk

Dada una función $h: \mathcal{X} \rightarrow \mathcal{Y}$ que pertenece a un espacio de hipótesis \mathcal{H} ; una función de pérdida $\mathcal{L}: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}^+$; y una base de datos $(x_i, y_i): \mathcal{X} \times \mathcal{Y}, i = 1, \dots, n$; definimos el riesgo empírico de f como:

$$\mathcal{R}(h) = \mathbb{E}_{(X,Y) \sim \mathcal{D}} [\mathcal{L}(h(X), Y)] = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(h(x_i), y_i)$$



Learning theory

Ahora calculamos el riesgo esperado:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \mathbb{E}_S\left[\mathbb{E}_X\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2 = \mathbb{E}_X\left[\mathbb{E}_S\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2$$



Learning theory

Supongamos que los datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ están definidos mediante la relación $Y = f(X) + \epsilon$, donde:

- $f(X)$ es la función verdadera (o esperanza condicional $f(X) = \mathbb{E}[Y | X]$).
- ϵ es un ruido independiente de X , con $\mathbb{E}[\epsilon] = 0$ y $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Sea $\hat{h}_S(X)$ el predictor construido a partir de una muestra $S \subseteq \mathcal{D}$. Si consideramos $\mathcal{L}(\hat{h}_S(x_i), y_i) = (y_i - \hat{h}_S(x_i))^2$, entonces el riesgo condicional para un predictor entrenado en S está dado por:

$$\mathcal{R}(\hat{h}_S) = \mathbb{E}_{(X,Y)} \left[(Y - \hat{h}_S(X))^2 \right]$$



Learning theory

Supongamos que los datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ están definidos mediante la relación $Y = f(X) + \epsilon$, donde:

- $f(X)$ es la función verdadera (o esperanza condicional $f(X) = \mathbb{E}[Y | X]$).
- ϵ es un ruido independiente de X , con $\mathbb{E}[\epsilon] = 0$ y $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Sea $\hat{h}_S(X)$ el predictor construido a partir de una muestra $S \subseteq \mathcal{D}$. Si consideramos $\mathcal{L}(\hat{h}_S(x_i), y_i) = (y_i - \hat{h}_S(x_i))^2$, entonces el riesgo condicional para un predictor entrenado en S está dado por:

$$\mathcal{R}(\hat{h}_S) = \mathbb{E}_{(X,Y)} \left[(Y - \hat{h}_S(X))^2 \right] = \mathbb{E}_{(X,Y)} \left[(f(X) + \epsilon - \hat{h}_S(X))^2 \right]$$



Learning theory

Supongamos que los datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ están definidos mediante la relación $Y = f(X) + \epsilon$, donde:

- $f(X)$ es la función verdadera (o esperanza condicional $f(X) = \mathbb{E}[Y | X]$).
- ϵ es un ruido independiente de X , con $\mathbb{E}[\epsilon] = 0$ y $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Sea $\hat{h}_S(X)$ el predictor construido a partir de una muestra $S \subseteq \mathcal{D}$. Si consideramos $\mathcal{L}(\hat{h}_S(x_i), y_i) = (y_i - \hat{h}_S(x_i))^2$, entonces el riesgo condicional para un predictor entrenado en S está dado por:

$$\begin{aligned}\mathcal{R}(\hat{h}_S) &= \mathbb{E}_{(X,Y)} \left[(Y - \hat{h}_S(X))^2 \right] = \mathbb{E}_{(X,Y)} \left[(f(X) + \epsilon - \hat{h}_S(X))^2 \right] = \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 + 2\epsilon(f(X) - \hat{h}_S(X)) + \epsilon^2 \right] \\ &= \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 \right] + \mathbb{E}_X \left[2\epsilon(f(X) - \hat{h}_S(X)) \right] + \mathbb{E}_X [\epsilon^2]\end{aligned}$$



Learning theory

Supongamos que los datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ están definidos mediante la relación $Y = f(X) + \epsilon$, donde:

- $f(X)$ es la función verdadera (o esperanza condicional $f(X) = \mathbb{E}[Y | X]$).
- ϵ es un ruido independiente de X , con $\mathbb{E}[\epsilon] = 0$ y $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Sea $\hat{h}_S(X)$ el predictor construido a partir de una muestra $S \subseteq \mathcal{D}$. Si consideramos $\mathcal{L}(\hat{h}_S(x_i), y_i) = (y_i - \hat{h}_S(x_i))^2$, entonces el riesgo condicional para un predictor entrenado en S está dado por:

$$\begin{aligned}\mathcal{R}(\hat{h}_S) &= \mathbb{E}_{(X,Y)} \left[(Y - \hat{h}_S(X))^2 \right] = \mathbb{E}_{(X,Y)} \left[(f(X) + \epsilon - \hat{h}_S(X))^2 \right] = \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 + 2\epsilon(f(X) - \hat{h}_S(X)) + \epsilon^2 \right] \\ &= \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 \right] + \mathbb{E}_X \left[2\epsilon(f(X) - \hat{h}_S(X)) \right] + \mathbb{E}_X [\epsilon^2]\end{aligned}$$

Usamos las propiedades de la esperanza matemática:

- $\mathbb{E}[\epsilon] = 0 \rightarrow \mathbb{E}[\epsilon(f(X) - \hat{h}_S(X))] = (f(X) - \hat{h}_S(X))\mathbb{E}[\epsilon] = 0$.
- $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Por lo tanto, el riesgo condicional se puede reescribir como:

$$\mathcal{R}(\hat{h}_S) = \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 \right] + \sigma_\epsilon^2$$



Learning theory

Supongamos que los datos $\mathcal{D}: \mathcal{X} \times \mathcal{Y}$ están definidos mediante la relación $Y = f(X) + \epsilon$, donde:

- $f(X)$ es la función verdadera (o esperanza condicional $f(X) = \mathbb{E}[Y | X]$).
- ϵ es un ruido independiente de X , con $\mathbb{E}[\epsilon] = 0$ y $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Sea $\hat{h}_S(X)$ el predictor construido a partir de una muestra $S \subseteq \mathcal{D}$. Si consideramos $\mathcal{L}(\hat{h}_S(x_i), y_i) = (y_i - \hat{h}_S(x_i))^2$, entonces el riesgo condicional para un predictor entrenado en S está dado por:

$$\begin{aligned}\mathcal{R}(\hat{h}_S) &= \mathbb{E}_{(X,Y)} \left[(Y - \hat{h}_S(X))^2 \right] = \mathbb{E}_{(X,Y)} \left[(f(X) + \epsilon - \hat{h}_S(X))^2 \right] = \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 + 2\epsilon(f(X) - \hat{h}_S(X)) + \epsilon^2 \right] \\ &= \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 \right] + \mathbb{E}_X \left[2\epsilon(f(X) - \hat{h}_S(X)) \right] + \mathbb{E}_X [\epsilon^2]\end{aligned}$$

Usamos las propiedades de la esperanza matemática:

- $\mathbb{E}[\epsilon] = 0 \rightarrow \mathbb{E}[\epsilon(f(X) - \hat{h}_S(X))] = (f(X) - \hat{h}_S(X))\mathbb{E}[\epsilon] = 0$.
- $\mathbb{E}[\epsilon^2] = \sigma_\epsilon^2$.

Por lo tanto, el riesgo condicional se puede reescribir como:

$$\mathcal{R}(\hat{h}_S) = \mathbb{E}_X \left[(f(X) - \hat{h}_S(X))^2 \right] + \sigma_\epsilon^2$$

Discrepancia entre $f(X)$ y $\hat{h}_S(X)$. **Contribución del ruido irreducible en los datos.**



Learning theory

Ahora calculamos el riesgo esperado:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \mathbb{E}_S\left[\mathbb{E}_X\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2 = \mathbb{E}_X\left[\mathbb{E}_S\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2$$

Despejamos $(f(X) - \hat{h}_S(X))^2$:

$$(f(X) - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)] + \mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2 + 2(f(X) - \mathbb{E}_S[\hat{h}_S(X)]) (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X)) + (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2$$



Learning theory

Ahora calculamos el riesgo esperado:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \mathbb{E}_S\left[\mathbb{E}_X\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2 = \mathbb{E}_X\left[\mathbb{E}_S\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2$$

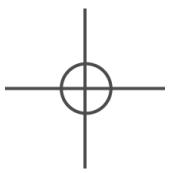
Despejamos $(f(X) - \hat{h}_S(X))^2$:

$$(f(X) - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)] + \mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2 + 2(f(X) - \mathbb{E}_S[\hat{h}_S(X)]) (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X)) + (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2$$

Ahora despejamos la esperanza matemática de cada término de la ecuación

- **Primer término:** $\mathbb{E}_X\left[\mathbb{E}_S\left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2\right]\right] = \mathbb{E}_X\left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2\right]$

Ya que $\mathbb{E}_S[\hat{h}_S(X)]$ es una constante con respecto a S , y $f(X)$ no depende de S .



Learning theory

Ahora calculamos el riesgo esperado:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \mathbb{E}_S\left[\mathbb{E}_X\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2 = \mathbb{E}_X\left[\mathbb{E}_S\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2$$

Despejamos $(f(X) - \hat{h}_S(X))^2$:

$$(f(X) - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)] + \mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2 + 2(f(X) - \mathbb{E}_S[\hat{h}_S(X)]) (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X)) + (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2$$

Ahora despejamos la esperanza matemática de cada término de la ecuación

- **Primer termino:** $\mathbb{E}_X\left[\mathbb{E}_S\left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2\right]\right] = \mathbb{E}_X\left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2\right]$

Ya que $\mathbb{E}_S[\hat{h}_S(X)]$ es una constante con respecto a S , y $f(X)$ no depende de S .

- **Segundo termino:** $\mathbb{E}_X\left[\mathbb{E}_S\left[2(f(X) - \mathbb{E}_S[\hat{h}_S(X)]) (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))\right]\right] = 0$

Son constantes con respecto a S . $\mathbb{E}_S[\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X)] = \mathbb{E}_S[\hat{h}_S(X)] - \mathbb{E}_S[\hat{h}_S(X)] = 0$



Learning theory

Ahora calculamos el riesgo esperado:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \mathbb{E}_S\left[\mathbb{E}_X\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2 = \mathbb{E}_X\left[\mathbb{E}_S\left[\left(f(X) - \hat{h}_S(X)\right)^2\right]\right] + \sigma_\epsilon^2$$

Despejamos $(f(X) - \hat{h}_S(X))^2$:

$$(f(X) - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)] + \mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2 = (f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2 + 2(f(X) - \mathbb{E}_S[\hat{h}_S(X)]) (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X)) + (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2$$

Ahora despejamos la esperanza matemática de cada término de la ecuación

- **Primer termino:** $\mathbb{E}_X\left[\mathbb{E}_S\left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2\right]\right] = \mathbb{E}_X\left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2\right]$

Ya que $\mathbb{E}_S[\hat{h}_S(X)]$ es una constante con respecto a S , y $f(X)$ no depende de S .

- **Segundo termino:** $\mathbb{E}_X\left[\mathbb{E}_S\left[2(f(X) - \mathbb{E}_S[\hat{h}_S(X)]) (\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))\right]\right] = 0$

Son constantes con respecto a S . $\mathbb{E}_S[\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X)] = \mathbb{E}_S[\hat{h}_S(X)] - \mathbb{E}_S[\hat{h}_S(X)] = 0$

- **Tercer termino:** $\mathbb{E}_X\left[\mathbb{E}_S\left[(\mathbb{E}_S[\hat{h}_S(X)] - \hat{h}_S(X))^2\right]\right] = \mathbb{E}_X\left[\text{Var}_S[\hat{h}_S(X)]\right]$



Bias-variance-noise decomposition

Finalmente:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \underbrace{\mathbb{E}_X \left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2 \right]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_X \left[\text{Var}_S[\hat{h}_S(X)] \right]}_{\text{Variance}} + \underbrace{\sigma_\epsilon^2}_{\text{Noise}}$$



Bias-variance-noise decomposition

Finalmente:

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \underbrace{\mathbb{E}_X[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_X[\text{Var}_S[\hat{h}_S(X)]]}_{\text{Variance}} + \underbrace{\sigma_\epsilon^2}_{\text{Noise}}$$

- **Bias²** = $\mathbb{E}_X[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2]$

Mide la diferencia entre la verdadera función objetivo $f(X)$ y la hipótesis promedio $\mathbb{E}_S[\hat{h}_S(X)]$. Representa el error sistemático debido a la incapacidad del modelo para aproximarse a $f(X)$.

- **Variance** = $\mathbb{E}_X[\text{Var}_S[\hat{h}_S(X)]]$

Cuantifica la sensibilidad del modelo a las fluctuaciones en la muestra de entrenamiento S . Un valor alto indican que el modelo cambia significativamente al cambiar de muestras.

- **Noise** = σ_ϵ^2

Es un término irreducible que depende únicamente de la varianza intrínseca del ruido en los datos.

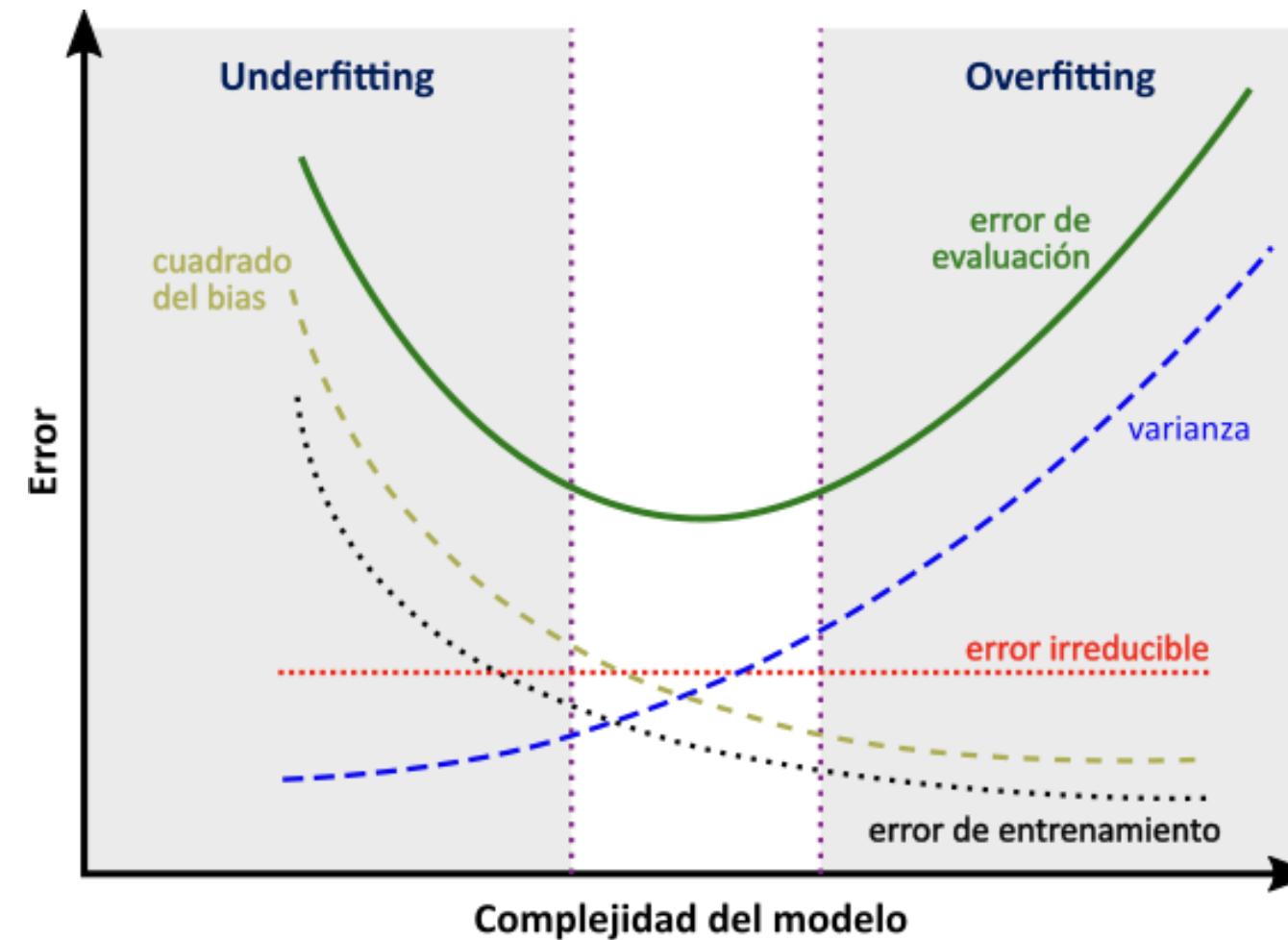


Bias–variance tradeoff

$$\mathbb{E}_S[\mathcal{R}(\hat{h}_S)] = \underbrace{\mathbb{E}_X[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2]}_{\text{Bias}^2} + \underbrace{\mathbb{E}_X[\text{Var}_S[\hat{h}_S(X)]]}_{\text{Variance}} + \underbrace{\sigma_\epsilon^2}_{\text{Noise}}$$

- Bias²** = $\mathbb{E}_X[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2]$

Mide la diferencia entre la verdadera función objetivo $f(X)$ y la hipótesis promedio $\mathbb{E}_S[\hat{h}_S(X)]$. Representa el error sistemático debido a la incapacidad del modelo para aproximarse a $f(X)$.

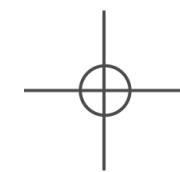


- Variance** = $\mathbb{E}_X[\text{Var}_S[\hat{h}_S(X)]]$

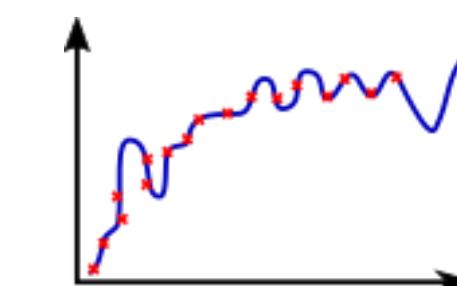
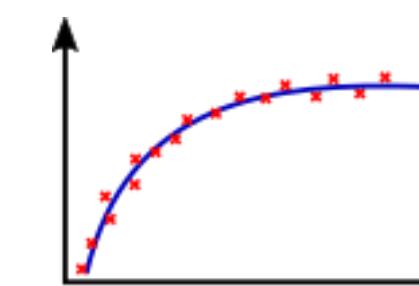
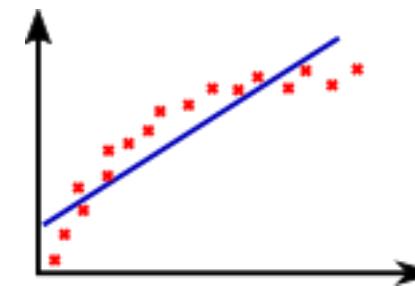
Cuantifica la sensibilidad del modelo a las fluctuaciones en la muestra de entrenamiento S . Un valor alto indica que el modelo cambia significativamente al cambiar de muestras.

- Noise** = σ_ϵ^2

Es un término irreducible que depende únicamente de la varianza intrínseca del ruido en los datos.

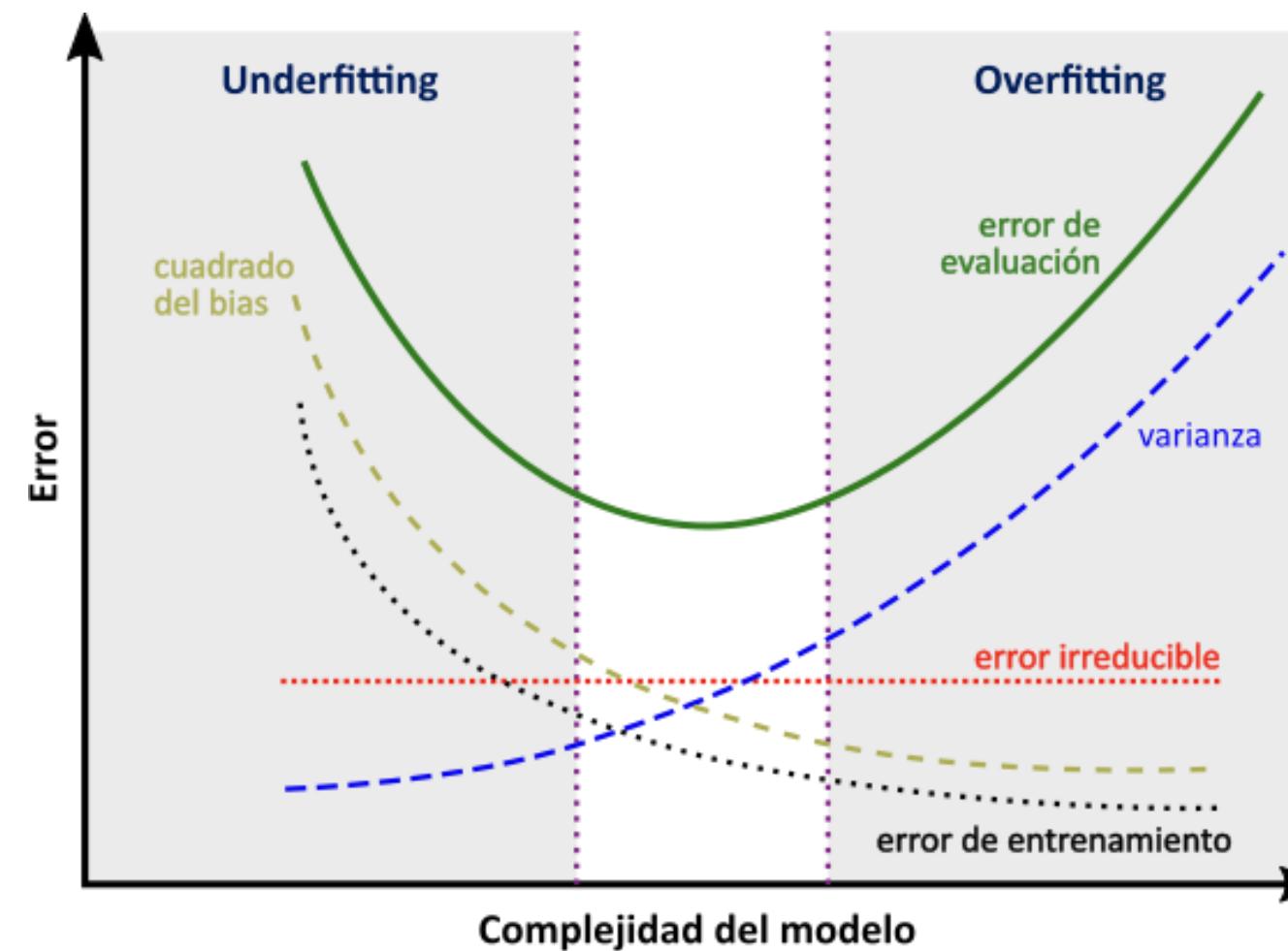


Bias–variance tradeoff



- Bias** $= \mathbb{E}_X \left[(f(X) - \mathbb{E}_S[\hat{h}_S(X)])^2 \right]$

Mide la diferencia entre la verdadera función objetivo $f(X)$ y la hipótesis promedio $\mathbb{E}_S[\hat{h}_S(X)]$. Representa el error sistemático debido a la incapacidad del modelo para aproximarse a $f(X)$.



- Variance** $= \mathbb{E}_X \left[\text{Var}_S[\hat{h}_S(X)] \right]$

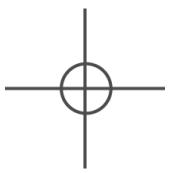
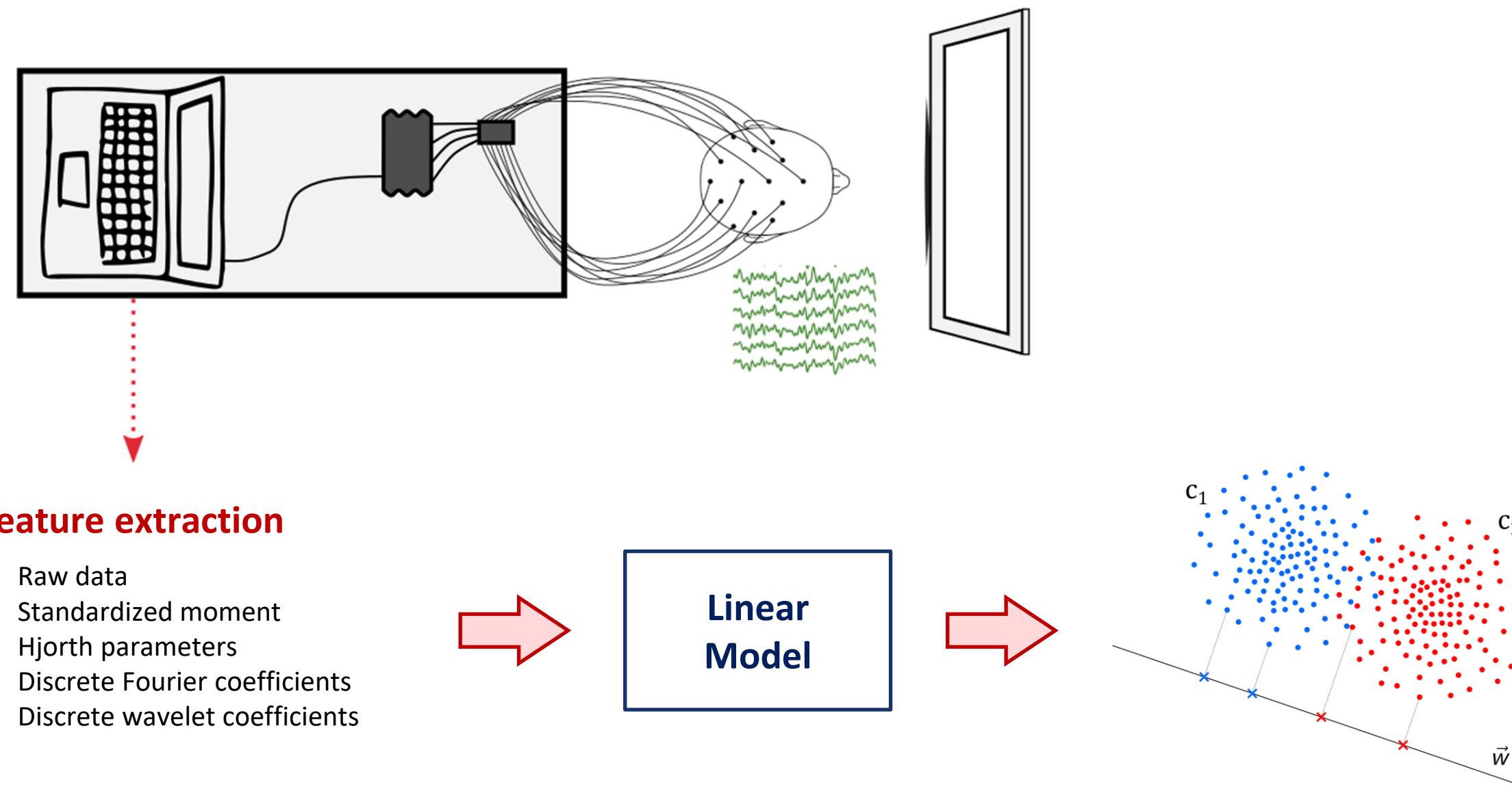
Cuantifica la sensibilidad del modelo a las fluctuaciones en la muestra de entrenamiento S . Un valor alto indican que el modelo cambia significativamente al cambiar de muestras.

- Noise** $= \sigma_\epsilon^2$

Es un término irreducible que depende únicamente de la varianza intrínseca del ruido en los datos.



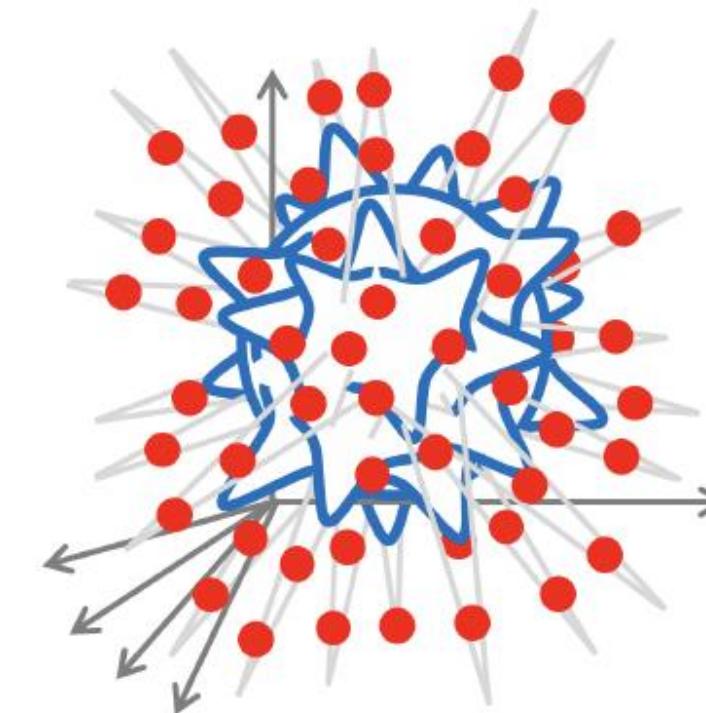
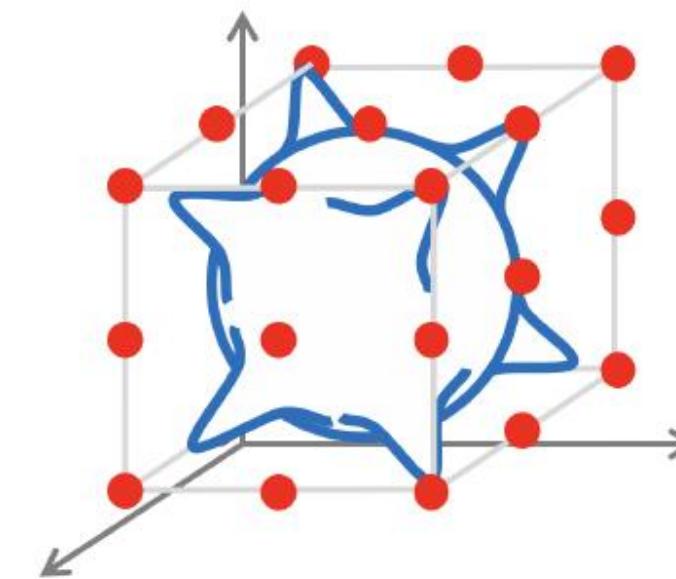
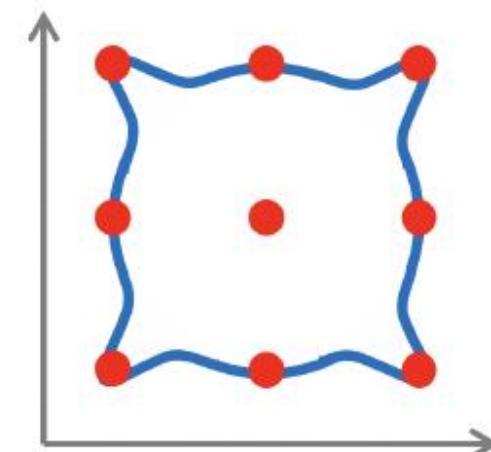
Modeling



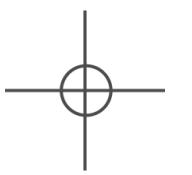
Curse of dimensionality

Fenómenos que surgen cuando el número de dimensiones de los datos aumenta, haciendo que las distancias se vuelvan menos informativas y los datos más dispersos. Esto complica tareas como clasificación, búsqueda de vecinos y generalización de modelos.

En baja dimensión los puntos están relativamente cerca y cubren bien el espacio.



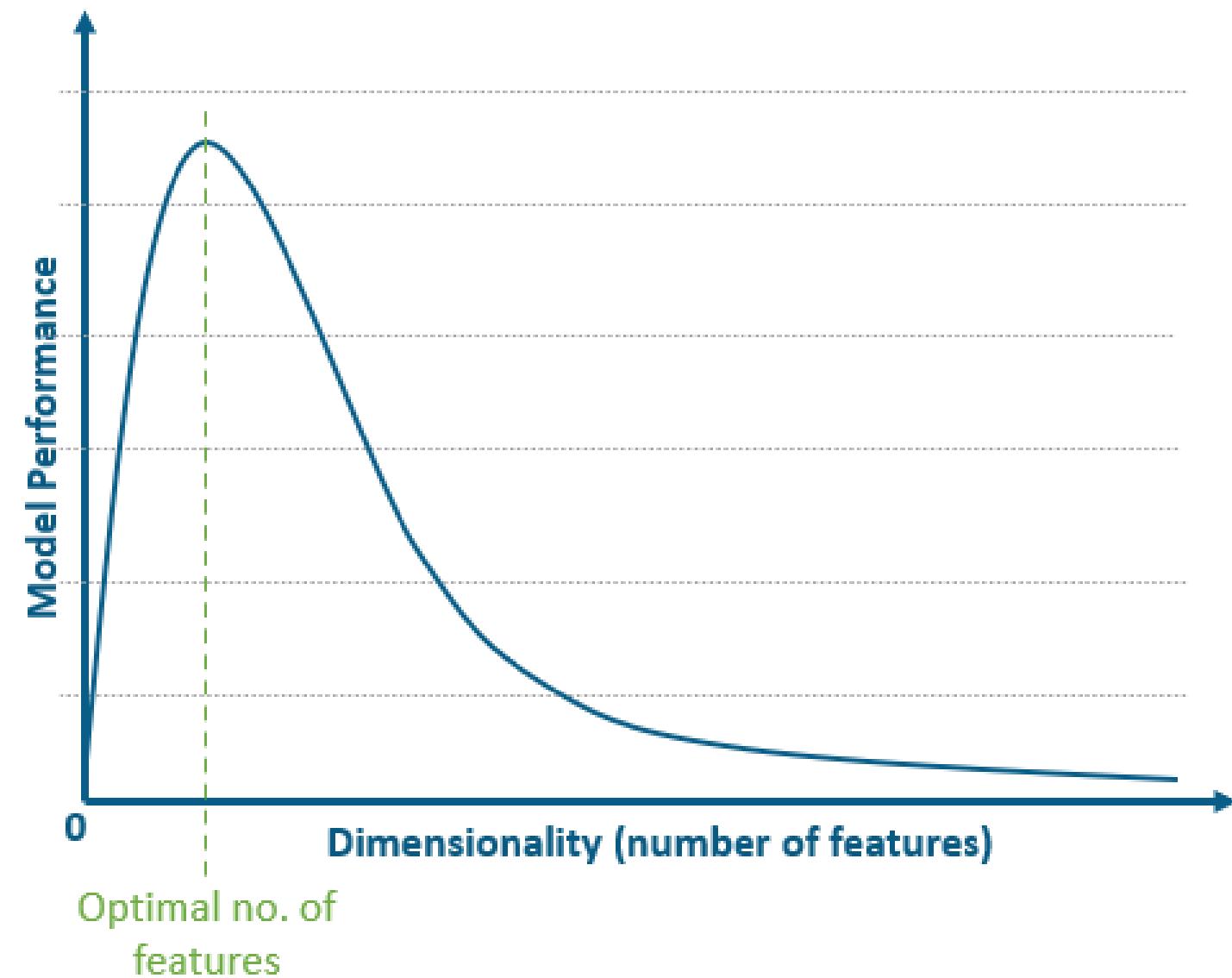
Al aumentar las dimensiones, los puntos se dispersan hacia los bordes y el espacio vacío crece.



Curse of dimensionality

Hughes Phenomenon

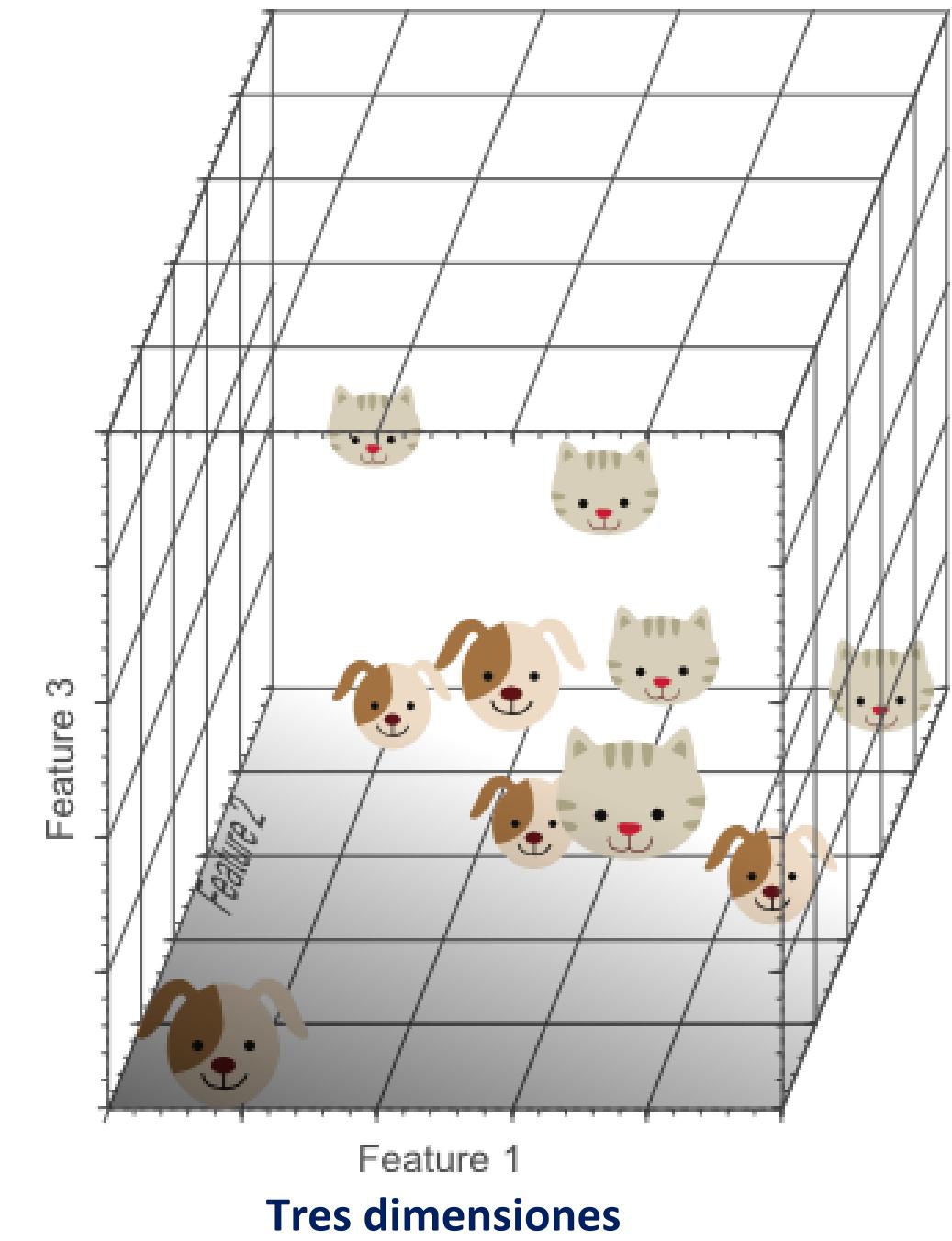
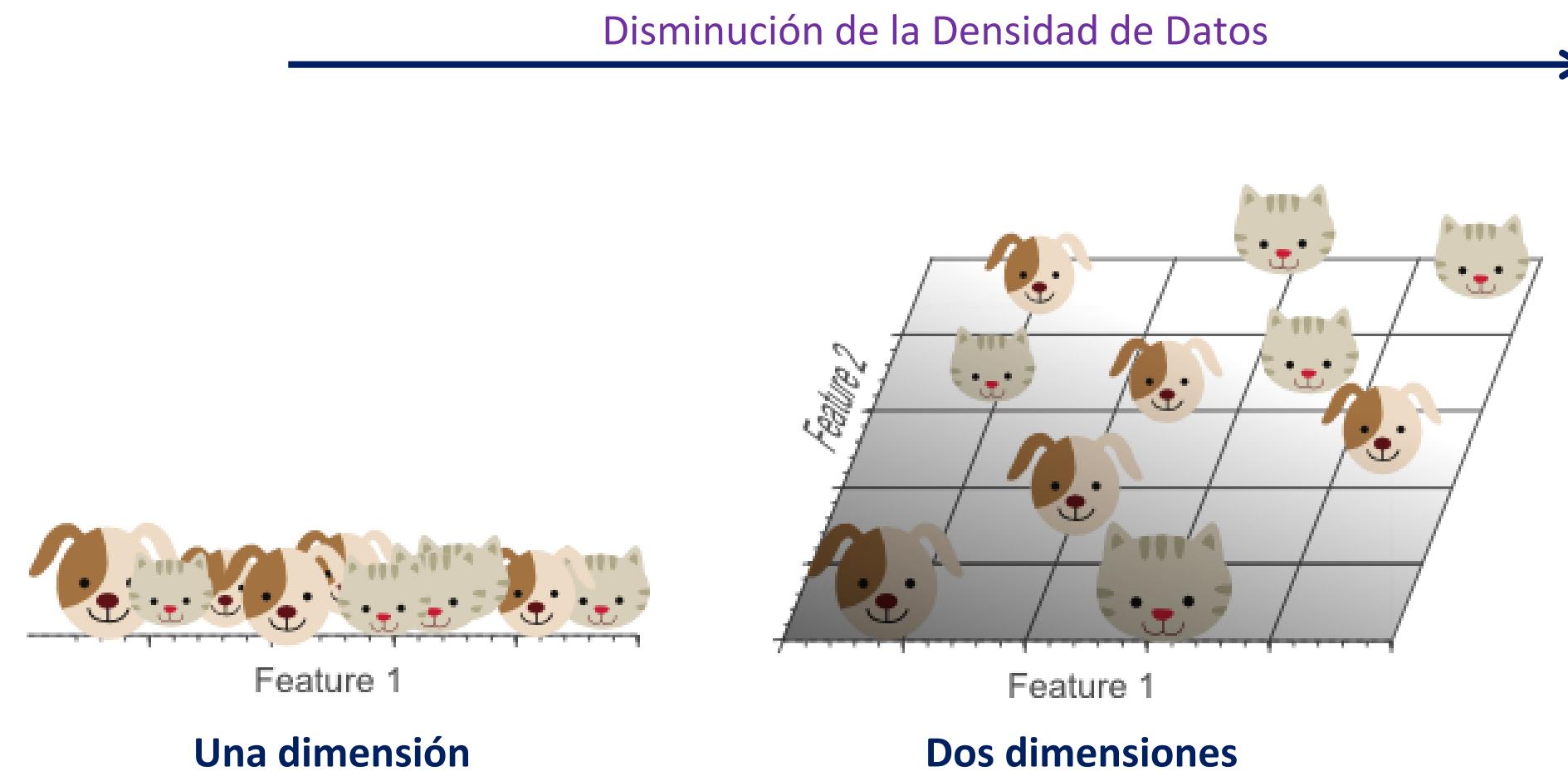
A medida que se incrementa el número de características, el rendimiento del modelo mejora hasta un punto óptimo. Más allá de este, el exceso de dimensiones introduce ruido y dispersión en los datos, provocando una caída en la capacidad de generalización.



Curse of dimensionality

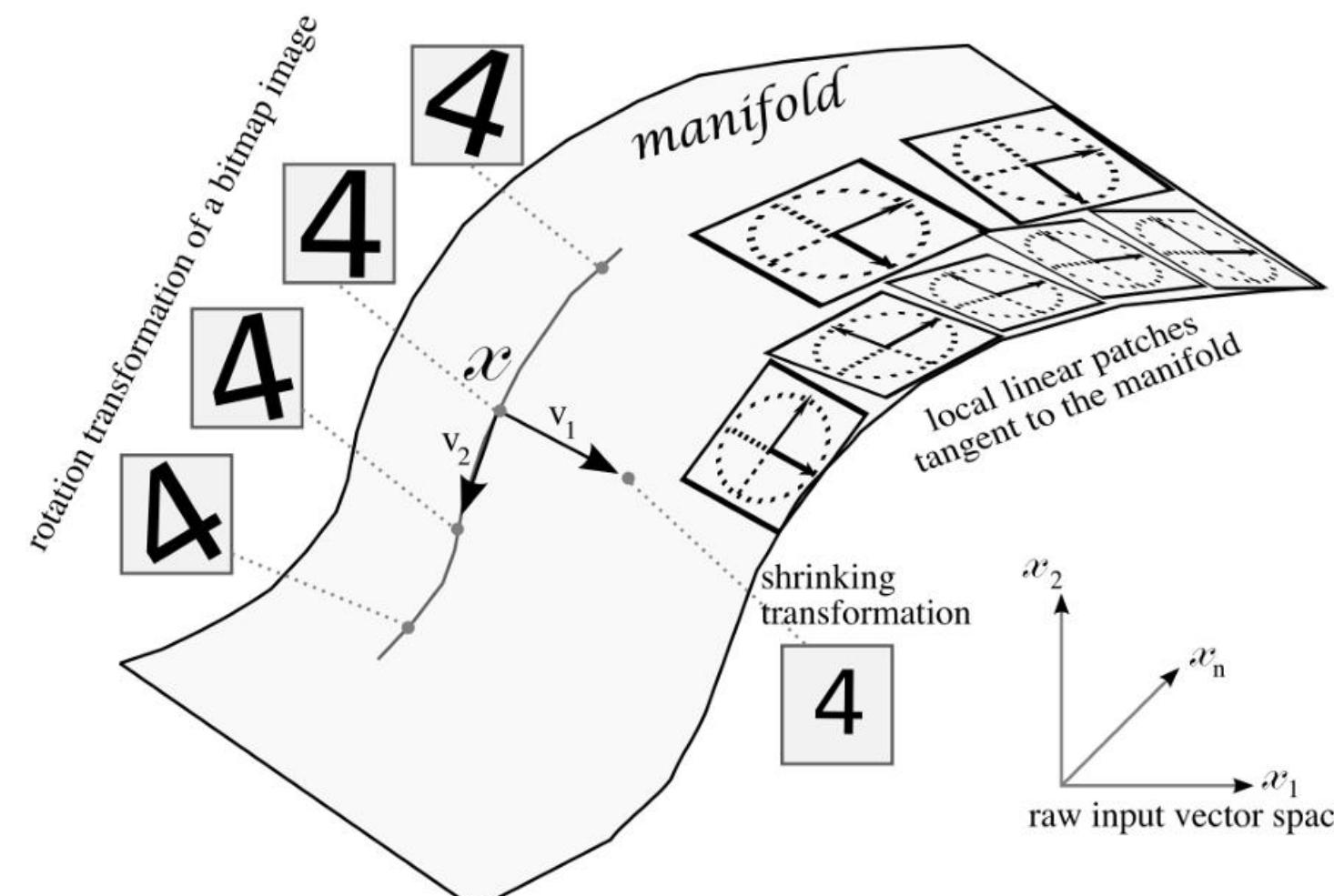
Sparse data space

Al aumentar las dimensiones, los datos se dispersan en el espacio y su densidad disminuye.



Manifold hypothesis

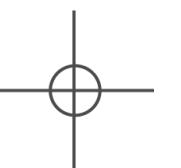
En aplicaciones de machine learning se suele partir de datos en un espacio de alta dimensión \mathbb{R}^D . Manifold Hypothesis plantea que, aunque los datos se describan con muchas features (D dimensiones), estos en realidad “viven” en (o muy cerca de) un subconjunto de dimensión intrínseca mucho menor que D . Dicho subconjunto se denomina manifold.



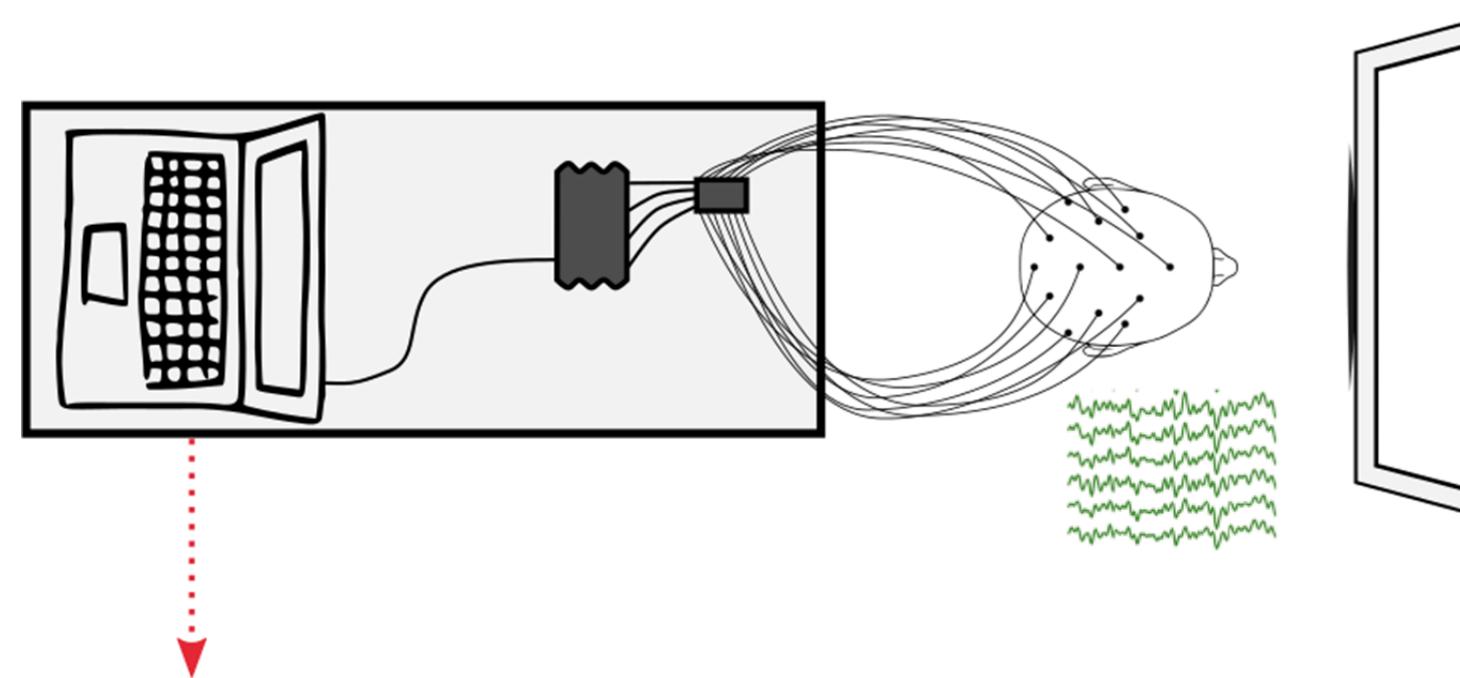
Existe un manifold $M \subset \mathbb{R}^D$ de dimensión $d \ll D$ tal que:

$$\mathcal{D}(\{x \in \mathbb{R}^D : d(x, M) \leq \epsilon\}) \approx 1,$$

para algún $\epsilon > 0$ pequeño y donde $d(x, M)$ es la distancia mínima de x a M . Así, casi todos los datos se encuentran dentro de un “tubo” de radio ϵ alrededor de M .

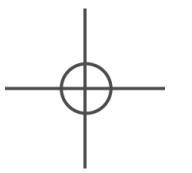
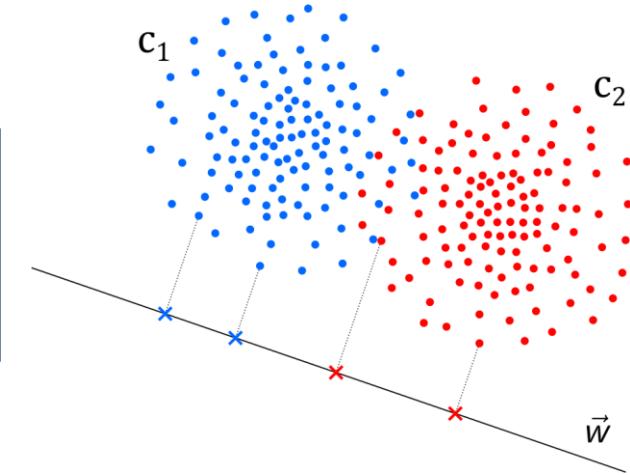
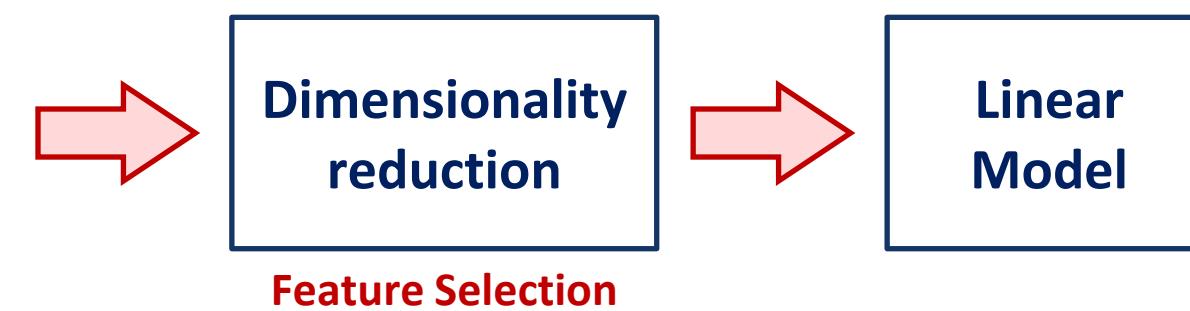


Modeling



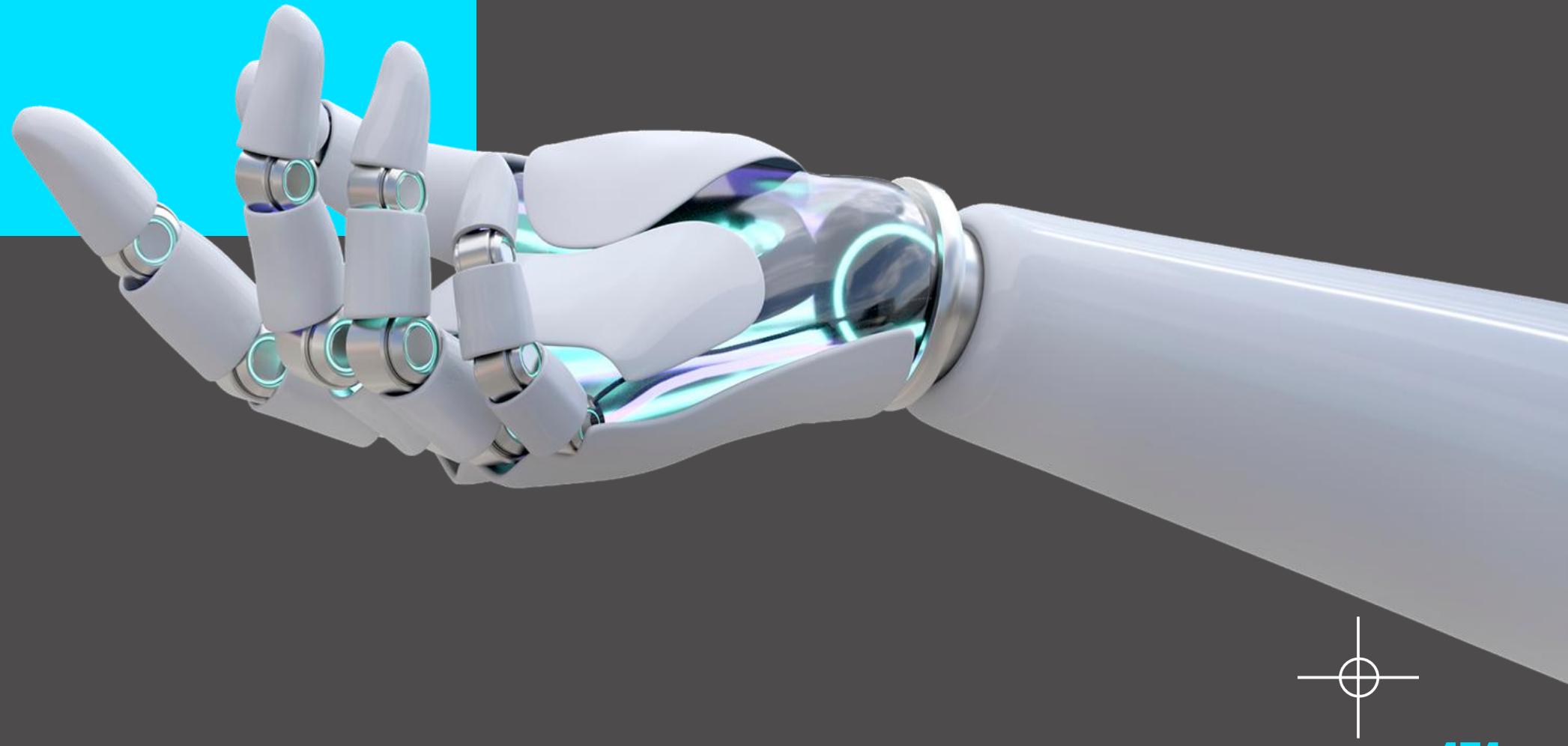
Feature extraction

- Raw data
- Standardized moment
- Hjorth parameters
- Discrete Fourier coefficients
- Discrete wavelet coefficients

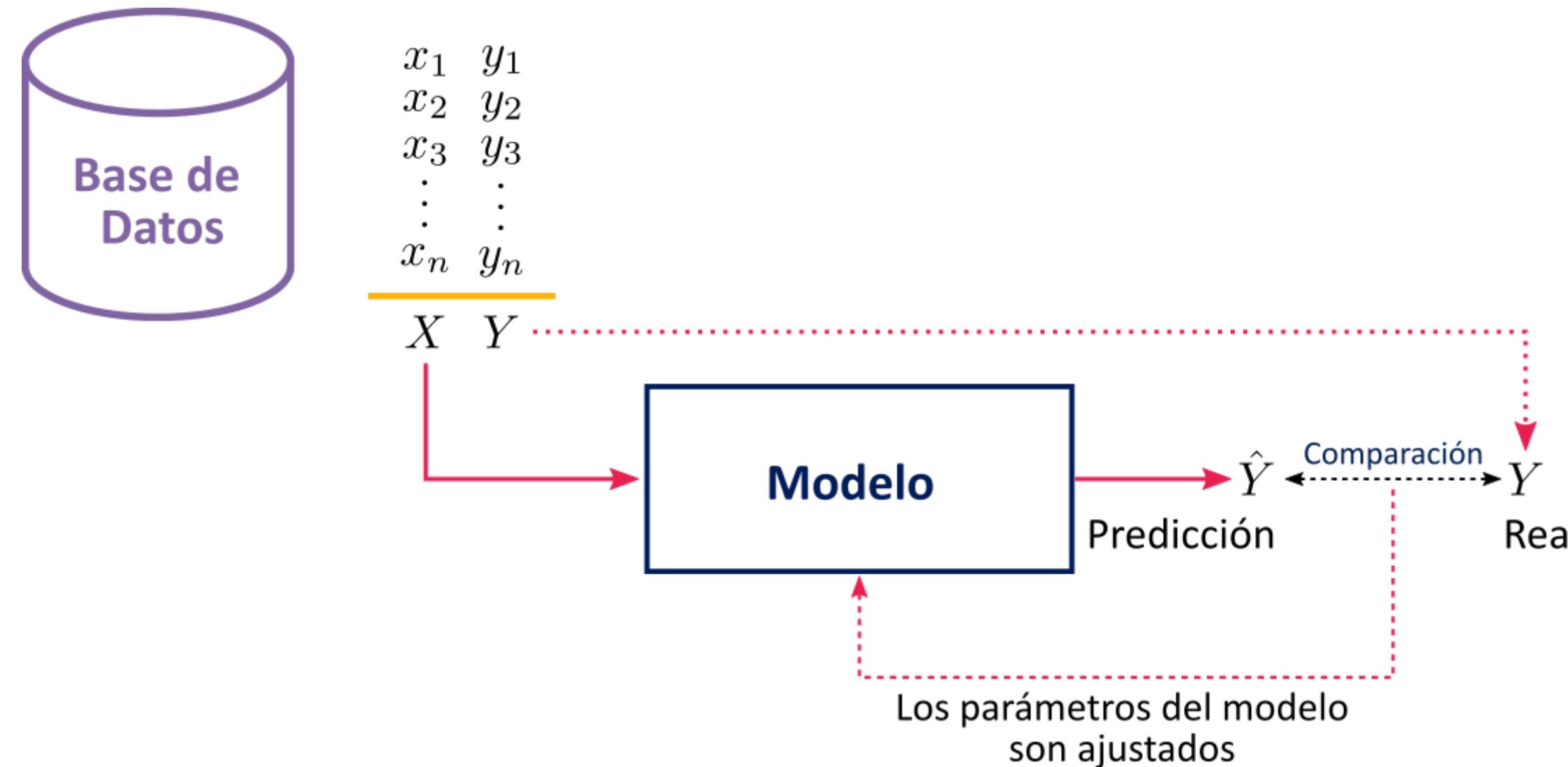


Evaluación de modelos

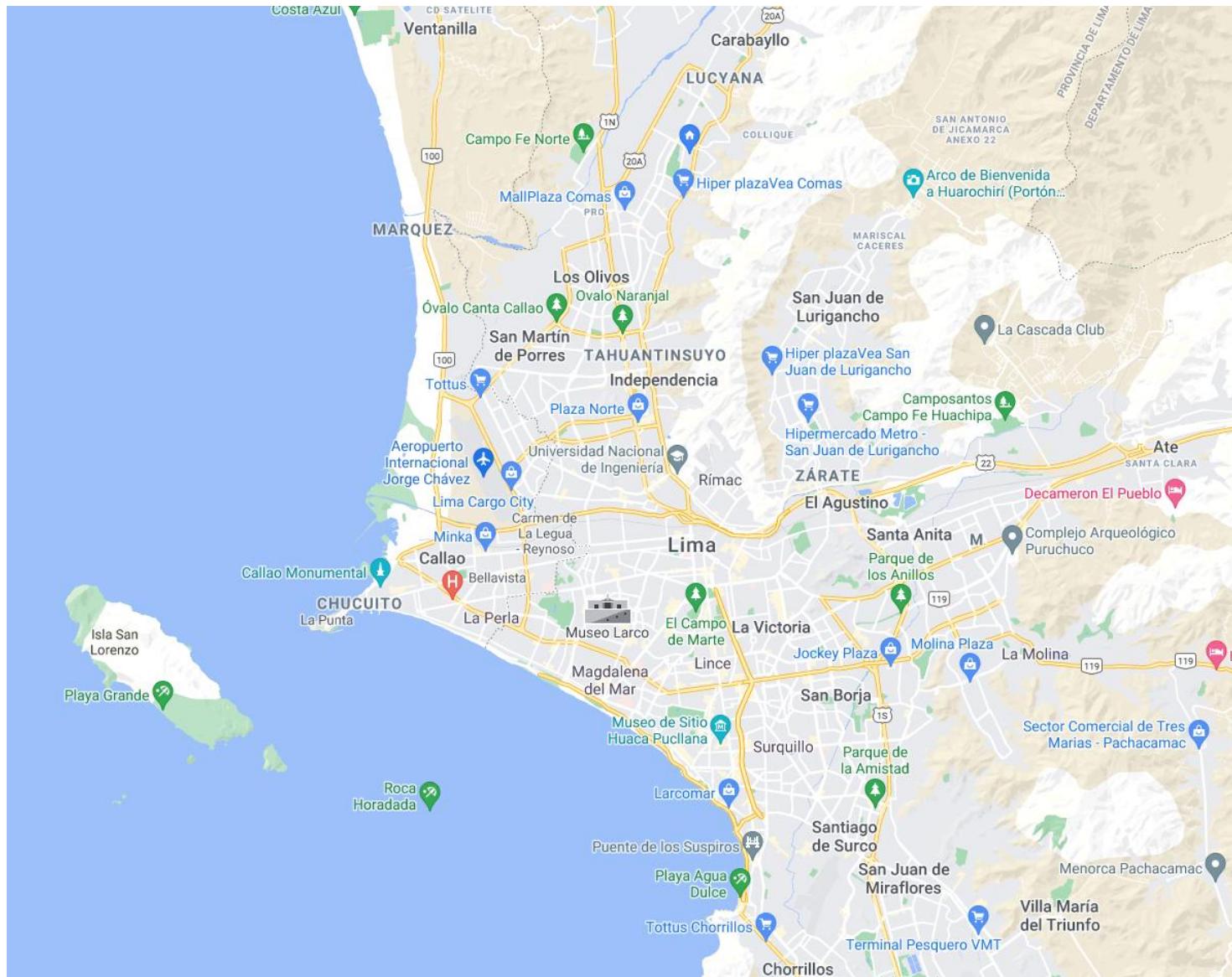
Holdout method, Cross-validation



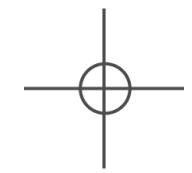
Training



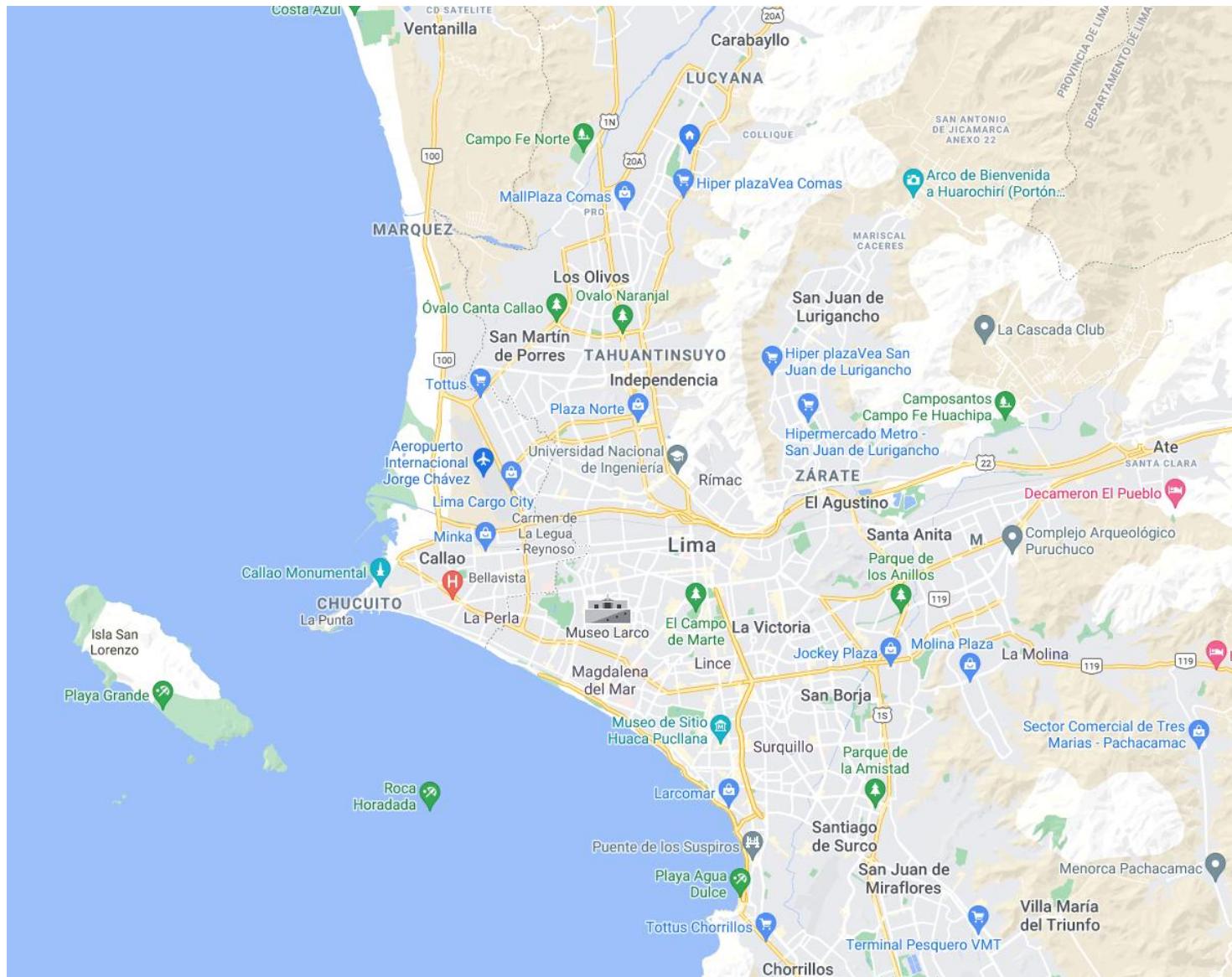
Training



Realizamos un estudio de mercado en varios distritos de la capital

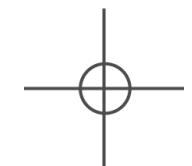


Training

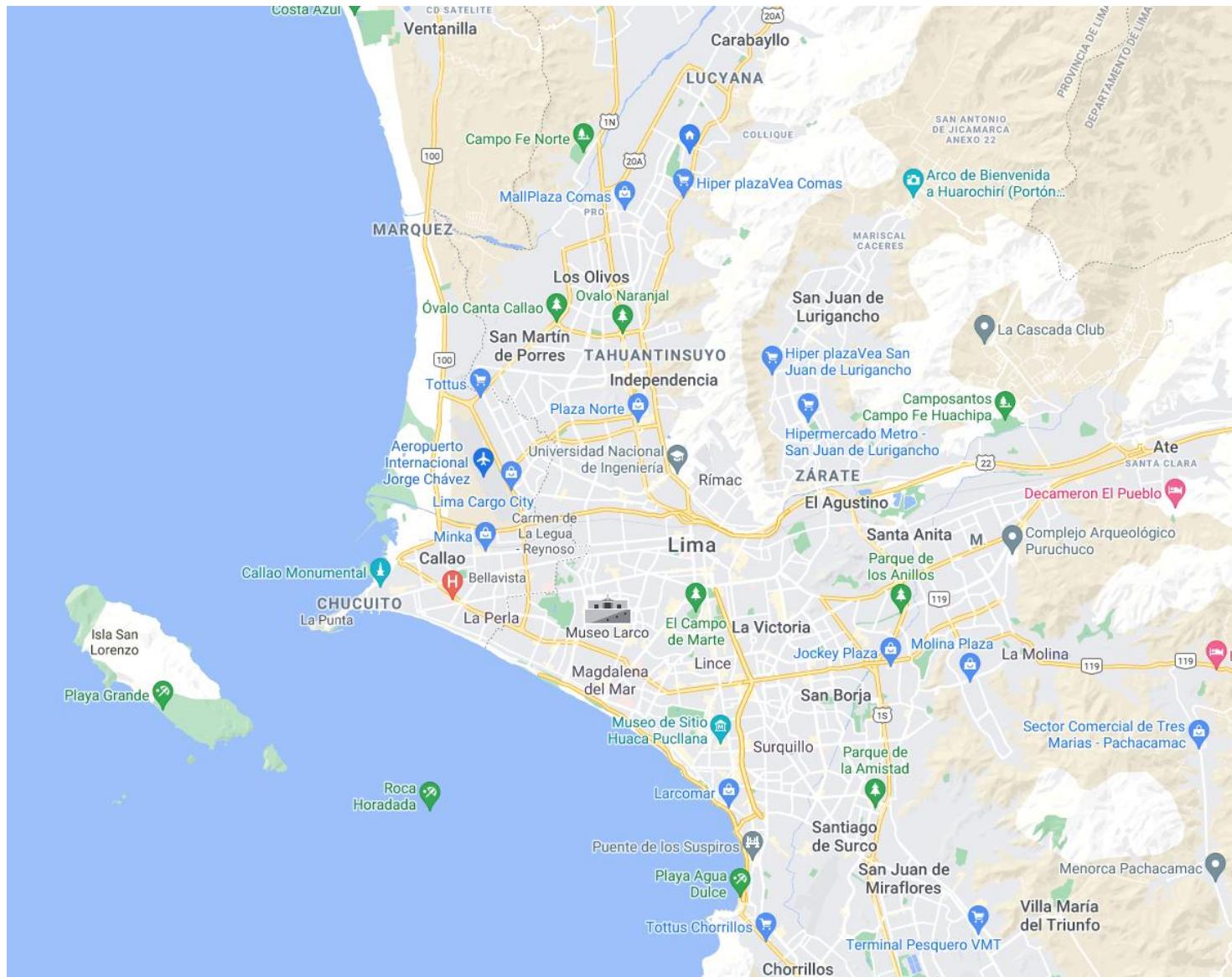


Datos recolectados

Carabayllo	10
Comas	10
Los Olivos	10
Centro de Lima	10
Independencia	10
La Victoria	10
Santa Anita	10
La Molina	10
Surco	10
Miraflores	10



Training



Datos recolectados

Carabayllo



Comas



Los Olivos



Centro de Lima



Independencia



La Victoria



Santa Anita



La Molina



Surco



Miraflores



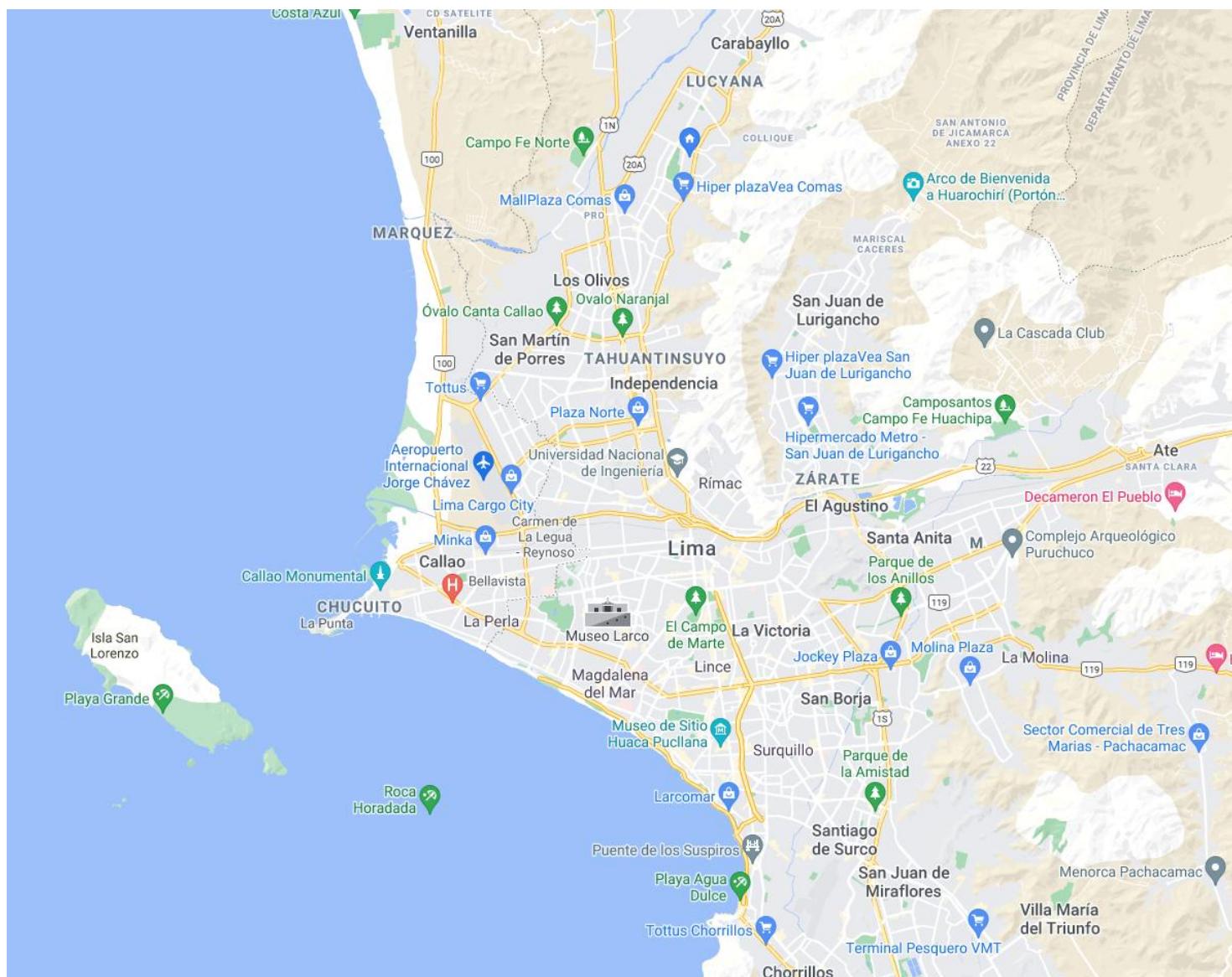
Entrenamiento

Evaluación

Precisión
del modelo: **65%**



Training



Datos recolectados

Carabayllo



Comas



Los Olivos



Centro de Lima



Independencia



La Victoria



Santa Anita



La Molina



Surco



Miraflores



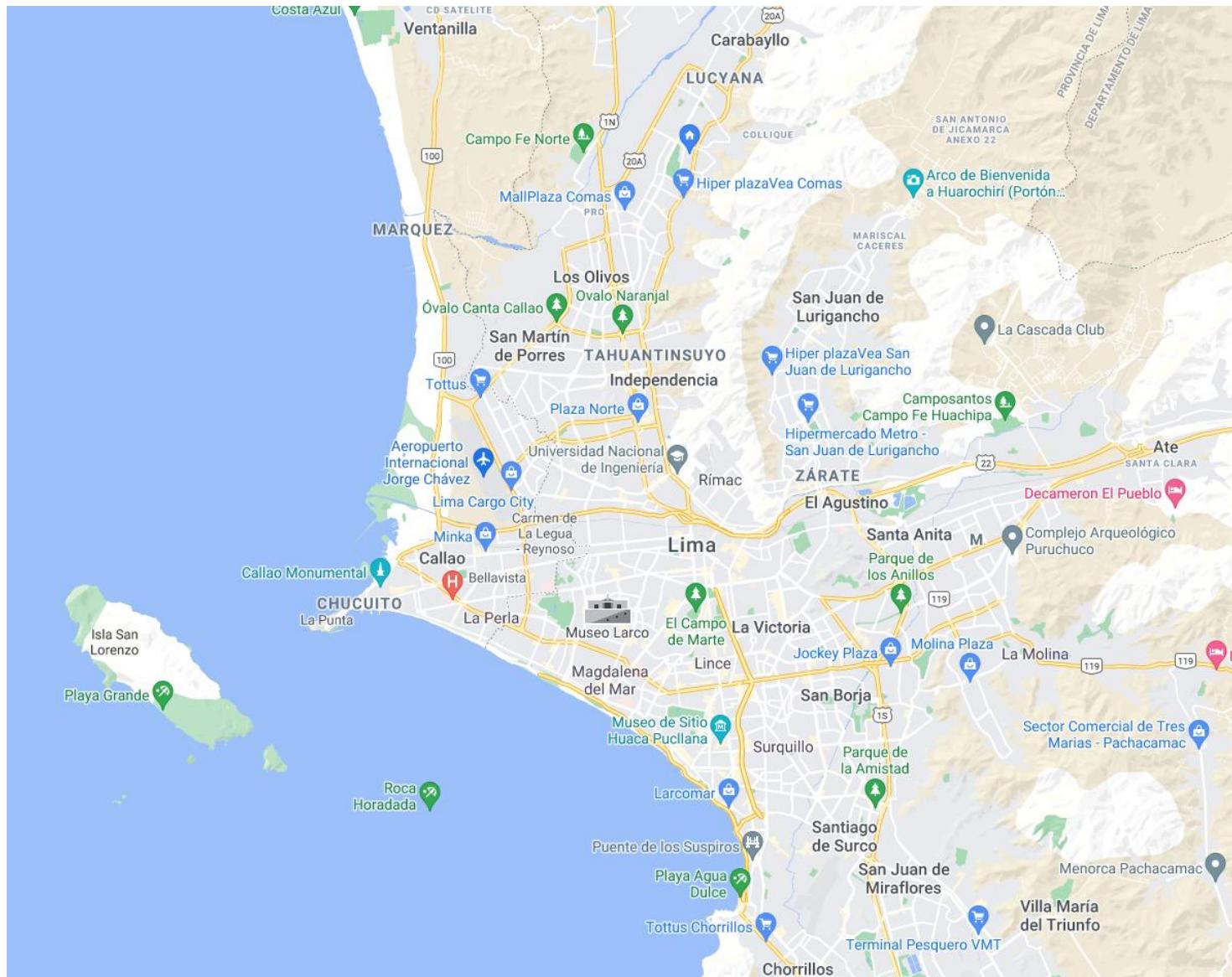
Evaluación

Entrenamiento

Precisión
del modelo: **73%**



Training



Datos recolectados

Carabayllo



Comas



Los Olivos



Centro de Lima



Independencia



La Victoria



Santa Anita



La Molina



Surco



Miraflores

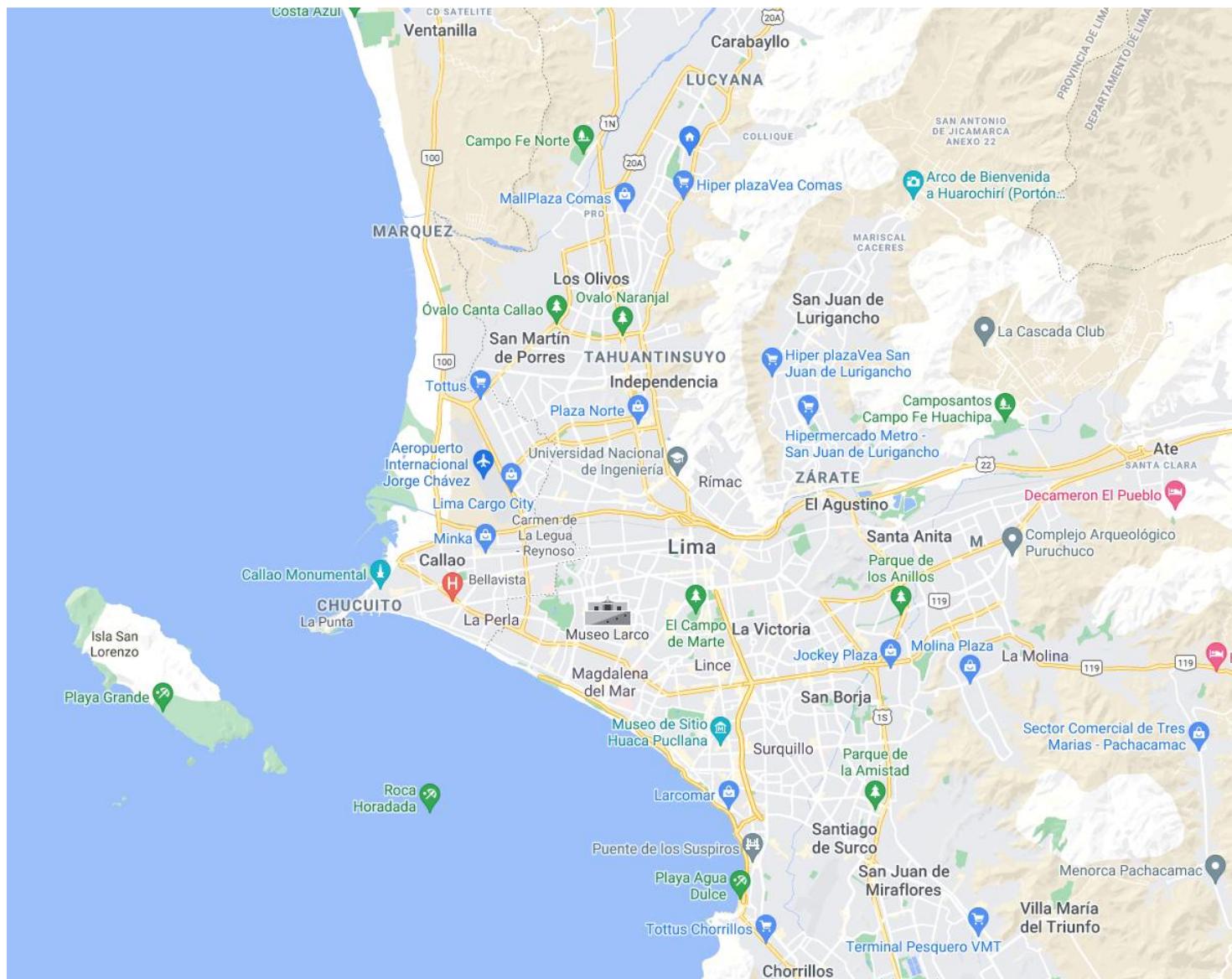


Entrenamiento

Evaluación



Training



Datos recolectados

Carabayllo



Comas



Los Olivos



Centro de Lima



Independencia



La Victoria



Santa Anita



La Molina



Surco



Miraflores



Entrenamiento

Evaluación

Precisión
del modelo: **92%**



Holdout method

Partición del dataset D en tres subconjuntos disjuntos:



Train: Ajustar parámetros del modelo θ (OLS, GD).

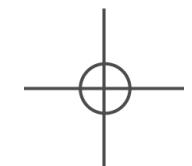
Validation: Selección de modelo/hiperparámetros (grado polinomial, λ , etc)

Test: Estimación final de desempeño (generalización).

Se usa una sola vez al final; no se usa para decidir hiperparámetros.

Buenas prácticas:

- Mezclar y fijar semilla (reproducibilidad).
- Si hay estructura temporal/grupos, no mezclar indiscriminadamente.
- Todo preprocessamiento (scaling, imputación, selección de features) se ajusta solo con Train y se aplica a Val/Test.

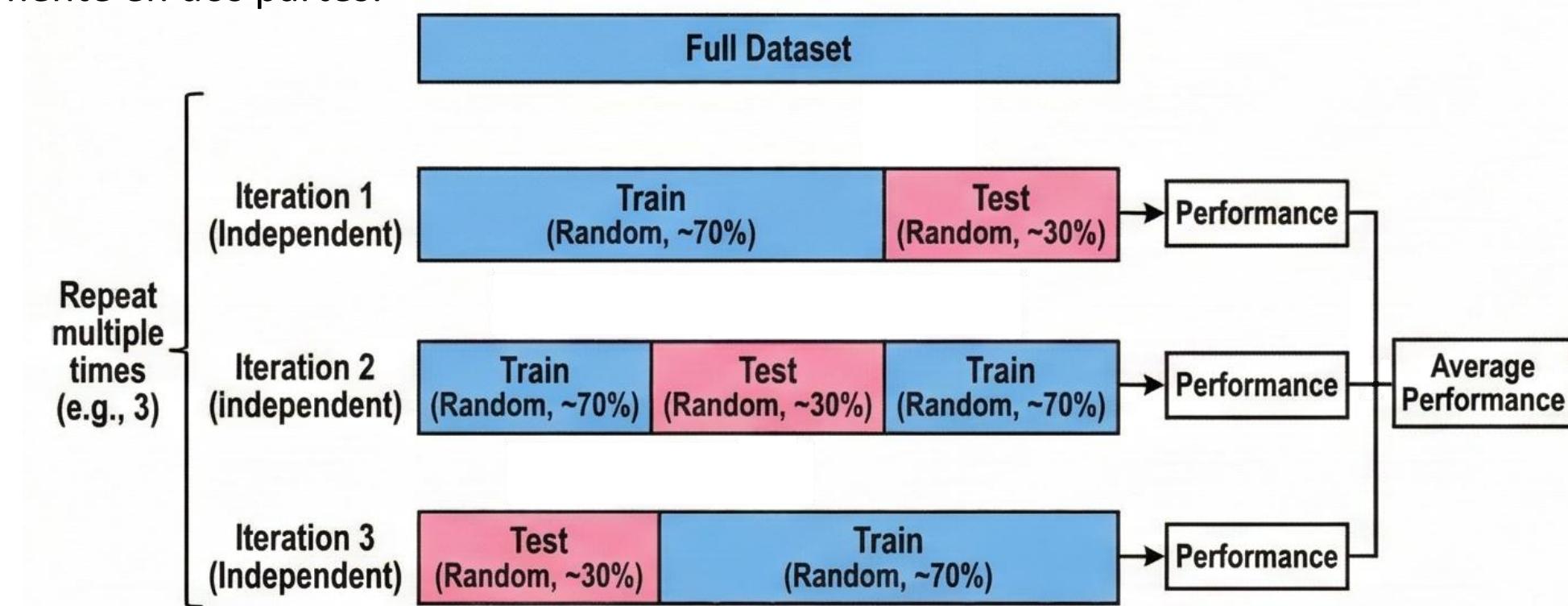


Cross-Validation

Random subsampling (repeated hold-out)

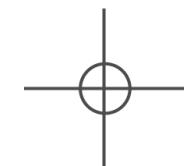
En esta técnica, el conjunto de datos se divide aleatoriamente en dos partes: training y validation, múltiples veces.

1. Dividimos tus datos (por ejemplo 70% entrenamiento, 30% prueba).
2. Entrenamos el modelo y evaluamos su rendimiento.
3. Volvemos a barajar todos los datos y repetimos el paso 1 y 2 varias veces.
4. El resultado final es el promedio de las evaluaciones.



Ventajas: Tenemos control total sobre cuántas veces repites el proceso, independientemente del tamaño de tus datos.

Desventajas: Algunas muestras pueden no ser seleccionadas nunca para ser evaluada, mientras que otras pueden ser seleccionadas múltiples veces, lo que puede introducir sesgo.



Cross-Validation

K-Fold Cross-Validation

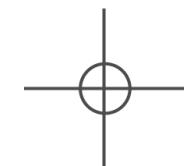
Es una de las técnicas más utilizadas por su equilibrio entre costo computacional y precisión estadística.

1. Dividimos el conjunto de datos aleatoriamente en K grupos (folds) de tamaño similar.
2. El proceso se repite K veces.
3. En cada iteración, usamos un grupo diferente como conjunto de evaluación y los $K - 1$ restantes como entrenamiento.
4. El rendimiento final es el promedio de los K resultados.



Ventajas: Garantiza que cada observación se use para entrenamiento y exactamente una vez para evaluación.

Desventajas: Computacionalmente más costoso que Hold-Out (una sola división), ya que entrenamos el modelo K veces.

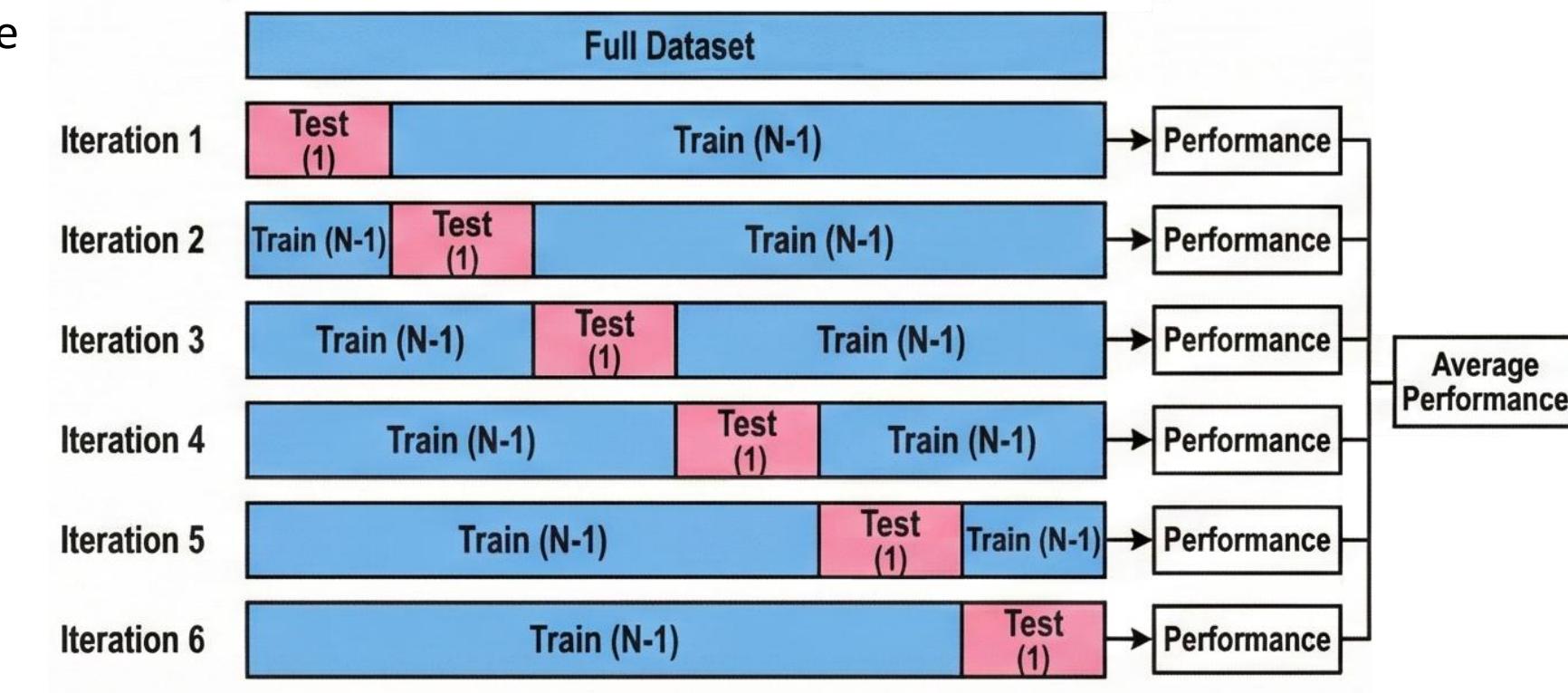


Cross-Validation

Leave-One-Out Cross-Validation (LOOCV)

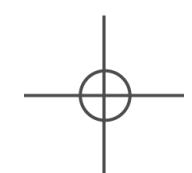
Es un caso extremo de K-Fold donde K es igual al número total de observaciones (N).

1. Tomamos una sola observación como conjunto de prueba.
2. Usamos todas las demás ($N - 1$) observaciones para entrenar.
3. Repetimos N veces, de modo que cada dato individual haya sido usado como prueba una vez.



Ventajas: No hay aleatoriedad en la división; los resultados son perfectamente reproducibles. Aprovecha al máximo los datos de entrenamiento

Desventajas: Extremadamente costoso computacionalmente si el dataset es grande (tienes que entrenar el modelo miles de veces). Además, puede tener una varianza alta en la estimación del error.

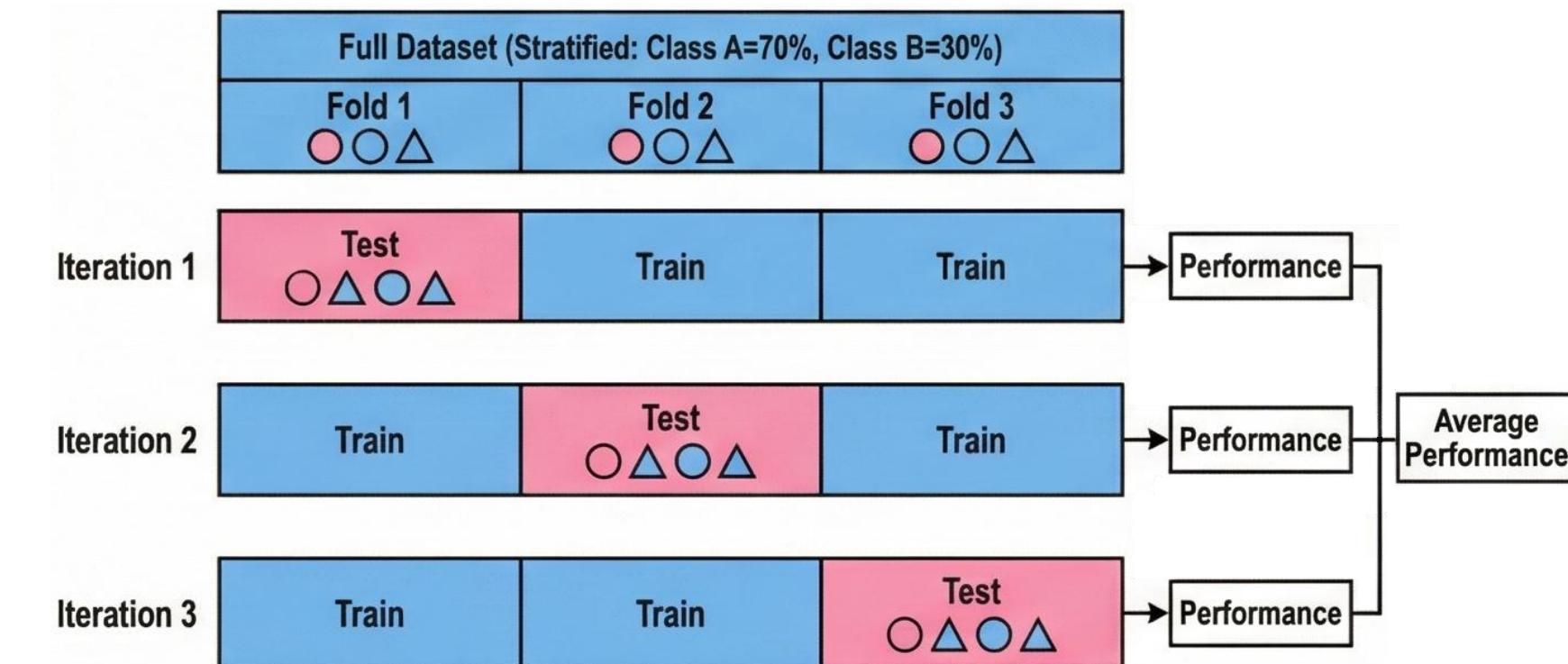


Cross-Validation

Stratified K-Fold Cross-Validation

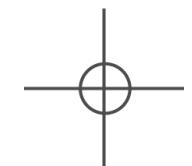
Es una variación del K-Fold diseñada para conjuntos de datos desbalanceados (donde una clase tiene muchas más muestras que otra).

Funciona igual que el K-Fold estándar, pero al dividir los K grupos, se asegura de mantener la misma proporción de clases que en el conjunto de datos original.



Ventajas: Evita que un fold de prueba quede sin muestras de la clase minoritaria, lo que daría una evaluación engañosa del modelo.

Desventajas: No es suficiente cuando hay correlación entre muestras (Por ejemplo: múltiples registros por individuo)





UTEC Posgrado

