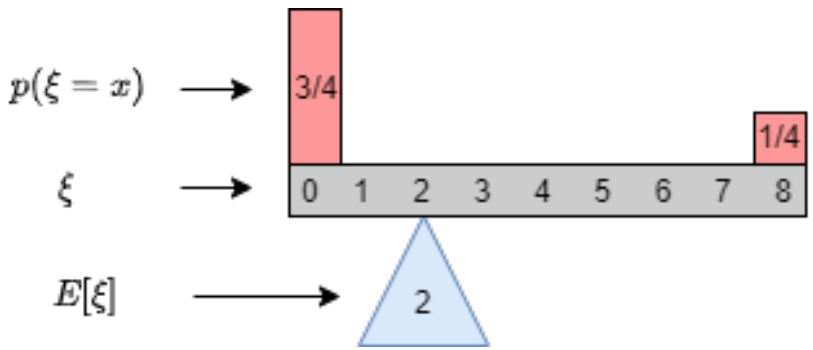
Stochastic Gradient Descent

Juan Bazerque

17 de noviembre de 2020

Variables aleatorias

- Variables aleatorias - modelan incertidumbre
  - Ejemplo - modelo de datos con ruido $X = \theta + \xi$
  - Ejemplo - imágenes como vectores aleatorios ξ
- Distribución
  - Normal - $\xi \sim \mathcal{N}(0, \sigma^2) \Rightarrow p(\xi = x) = 1/\sqrt{\pi\sigma^2}\exp(-(x-\mu)^2/\sigma)$
  - Uniforme - $\xi \sim \mathcal{U}(a, b) \Rightarrow p(\xi = x) = 1/(b-a)$ si $x \in (a, b)$ y cero sino
  - Bernoulli - $\xi \sim \mathcal{B}er(p) \Rightarrow p(\xi = 1) = p$ y $P(\xi = 0) = 1 - p$
- Esperanza $E[xi] = \sum_x xp(\xi = x)$ o $E[\xi] = \int_x xp(\xi = x)dx$
  - Ejemplo Bernoulli $E[\xi] = 0P(\xi = 0) + 1P(\xi = 1) = p$



Procesos estocásticos

- Proceso estocástico - secuencia de variables aleatorias $\{\xi_t\}_{t>=0}$
- Lo define su probabilidad conjunta $p(\xi_1 = x_1, \xi_2 = x_2, \xi_3 = x_3, ...)$
- Ley de los grandes números
  - Permite calcular la esperanza a partir de muestras
  - Sin conocer la distribución de probabilidades (model free)
  - Versión estándar asume ξ_t son i.i.d. ($E[\xi_t] = E[\xi]$)

$$\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T} \xi_t = E[\xi]$$

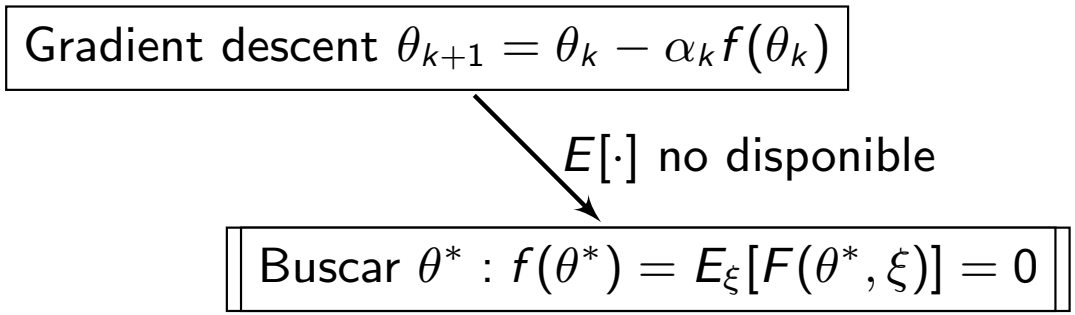
Aproximación Estocástica

- Teoría general de aproximación estocástica

$$\boxed{\text{Obtener } \theta^* : f(\theta^*) = E_\xi[F(\theta^*, \xi)] = 0}$$

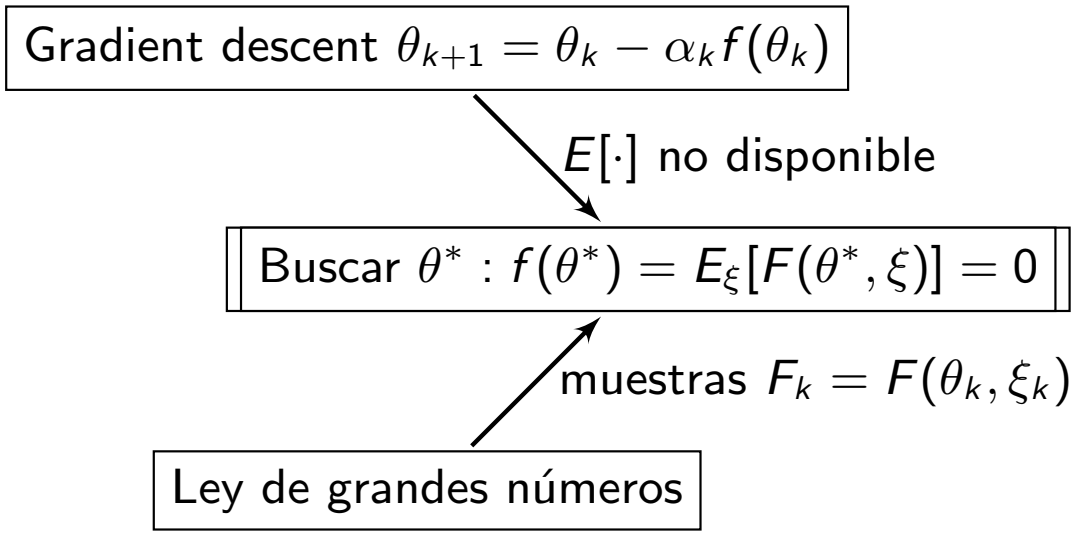
Aproximación Estocástica

- Teoría general de aproximación estocástica



Aproximación Estocástica

- Teoría general de aproximación estocástica



Aproximación Estocástica

- Teoría general de aproximación estocástica

$$\boxed{\text{Obtener } \theta^* : f(\theta^*) = E_\xi[F(\theta^*, \xi)] = 0}$$

- Idea: Combinar algoritmos determinísticos con la LGN
- Objetivo: Evitar calcular esperanzas y usar datos en cambio
- Permite el procesamiento recursivo de los datos
- Caso de interés: Descenso por gradiente estocástico (SGD)
- SGD = Stochastic Approximation + Gradient Descent

Ejemplo

- Neurona RELU para clasificación de imágenes

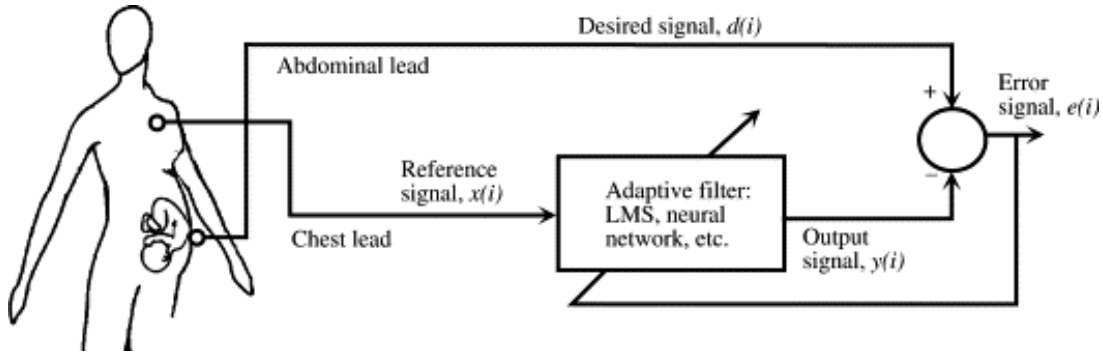



RELU

- Permite incorporar los datos uno a uno
- Generaliza a redes neuronales

Ejemplo

- Linear Mean Squares
- Se tiene un modelo lineal de los datos
- LMS ajusta los parámetros de dicho modelo
- Minimiza la diferencia entre los datos y el modelo lineal
- Procesa los datos recursivamente en tiempo real



Métodos de Aproximación Estocástica

- Familia de métodos iterativos para hallar raíces
- Usado en optimización donde los datos tienen ruido.
- También es útil para procesar grandes cantidades de datos
- Consideramos la esperanza de una función arbitraria $F : \mathbb{R}^d \times \Xi \to \mathbb{R}$

$$f(\theta) \triangleq \mathbb{E}_\xi \left[F(\theta, \xi) \right] = \int_\Xi F(\theta, \xi) dP(\xi)$$

- Parámetros $\theta \in \mathbb{R}^d$ y v.a. ξ con distribución en $\Xi \subseteq \mathbb{R}^m$
- El problema es hallar raíces de f, i.e., θ^* tales que $f(\theta^*) = 0$
- Formulación flexible que cubre muchos modelos
- E.g., ruido de medida: $F(\theta, \xi) = f(\theta) + \xi$
- Varias formas de resolver este problema
- Pondremos foco en Robbins-Monro y Martingalas

La idea detrás de Robbins-Monro

- Robbins-Monro es un simple algoritmo para hallar raíces
- Encuentra $\theta^\star$ tal que $f(\theta^\star) = 0$
- θ y $f(\theta)$ determinísticas $\Rightarrow$ descenso por gradiente

$$\theta_{k+1} = \theta_k - \alpha_k f(\theta_k) = \theta_k - \alpha_k \mathbb{E}_\xi \left[F(\theta_k, \xi) \right]$$

- Encontrando $f(\theta_k) = 0$ (gradiente nulo), estamos en un óptimo local
- Robbins-Monro descarta la esperanza $\mathbb{E}_\xi[\cdot]$

$$\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \xi_k)$$

- Solo una muestra $F_k \triangleq F(\theta_k, \xi_k)$ es necesaria por iteración
- También conocido como Stochastic Gradient Descent (SGD)

Convergencia del algoritmo Robbins-Monro

- Robbins-Monro usa muestras del gradiente $F_k = F(\theta_k, \xi_k)$

$$\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \xi_k)$$

- En vez de $\mathbb{E}_\xi [F(\theta_k, \xi)]$

$$\theta_{k+1} = \theta_k - \alpha_k \mathbb{E}_\xi [F(\theta_k, \xi)]$$

- ¿Aún así converge al óptimo?

Ejemplo: Mean Square Error

- Encontrar el error cuadrático medio

$$\theta^* = \arg\min_{\theta} \frac{1}{2}\mathbb{E}_{\xi}[(\theta - \xi)^2]$$

- Queremos encontrar la raíz del gradiente

$$\nabla_{\theta} \frac{1}{2}\mathbb{E}_{\xi}[(\theta - \xi)^2] = \theta - \mathbb{E}_{\xi}[\xi] = 0$$

- Anulando el gradiente obtenemos $\theta = \mathbb{E}_{\xi}[\xi]$
- Stochastic Gradient Descent (SGD) Se mueve en la dirección

$$\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \xi_k) = \theta_k - \alpha_k(\theta_k - \xi_k)$$

Ejemplo: Mean Square Error

- Calculemos la solución $\theta = \mathbb{E}_\xi[\xi]$ recursivamente como promedio

$$\begin{aligned}
\hat{\theta}_{k+1} &= \frac{1}{k+1}\sum_{i=0}^{k}\xi_i \\
&= \frac{1}{k+1}\sum_{i=0}^{k-1}\xi_i + \frac{1}{k+1}\xi_k \\
&= \underbrace{\frac{k}{k+1}}_{1-1/k+1}\hat{\theta}_k + \frac{1}{k+1}\xi_k \\
&= \hat{\theta}_k - \frac{1}{k+1}(\hat{\theta}_k - \xi_k)
\end{aligned}$$

- Luce muy similar a stochastic gradient descent...

Ejemplo: Mean Square Error

- Promedio de las muestras

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \frac{1}{k+1}(\hat{\theta}_k - \xi_k)$$

- Stochastic Gradient Descent (SGD)

$$\theta_{k+1} = \theta_k - \alpha_k(\theta_k - \xi_k) = \theta_k - \frac{1}{k+1}(\theta_k - \xi_k)$$

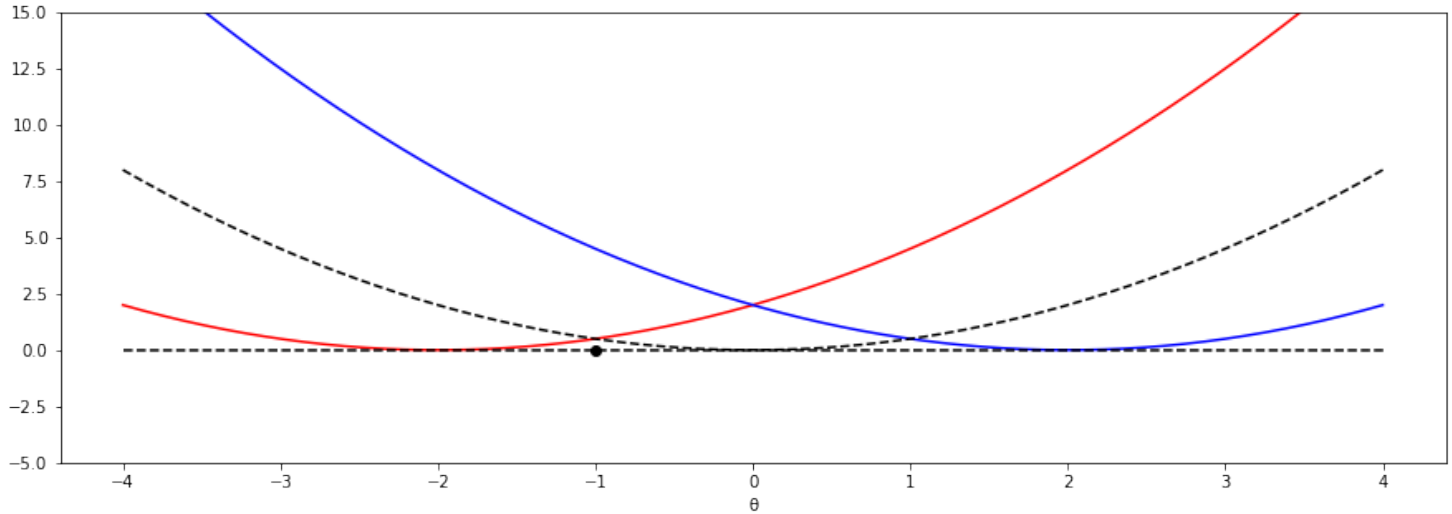
- Gradient Descent

$$\theta_{k+1} = \theta_k - \alpha_k \mathbb{E}_\xi[\theta_k - \xi] = \theta_k - \frac{1}{k+1}\mathbb{E}_\xi[\theta_k - \xi]$$

- SGD descarta la esperanza y aprende de los datos uno a uno
- Acumulador - minimiza y promedia simultáneamente

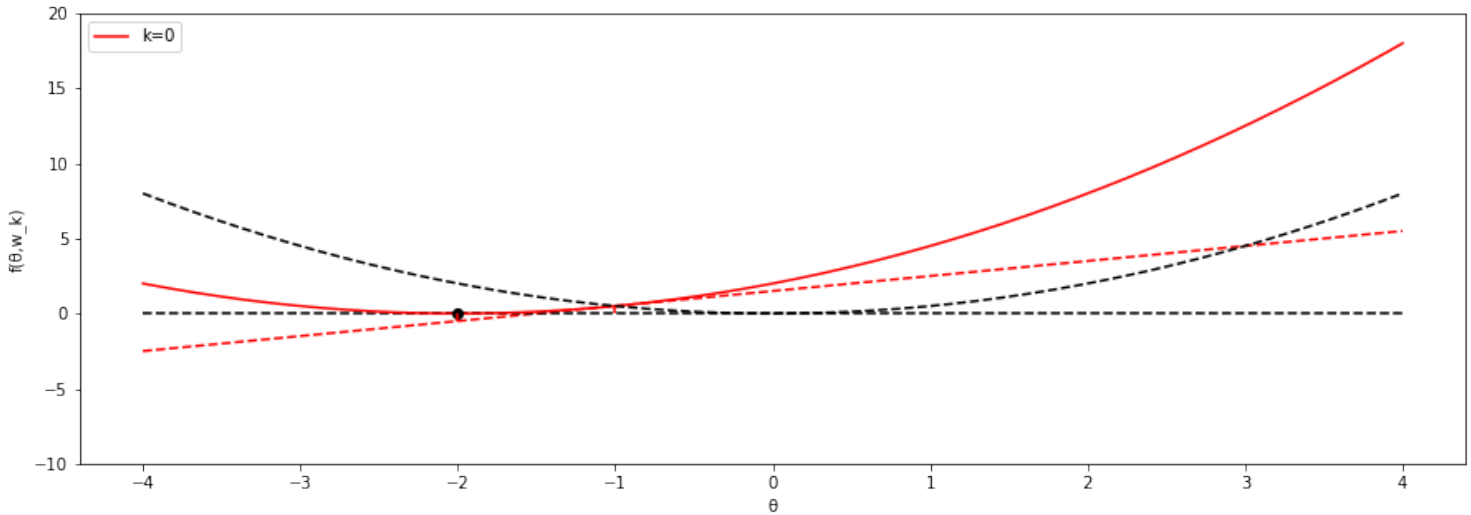
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



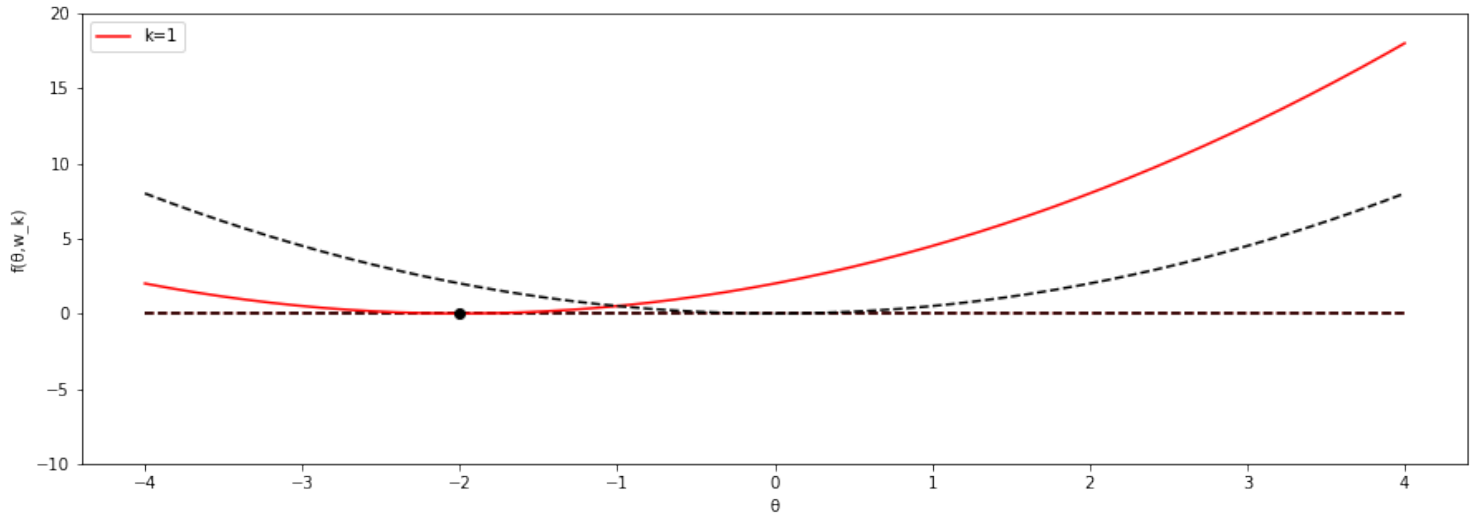
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



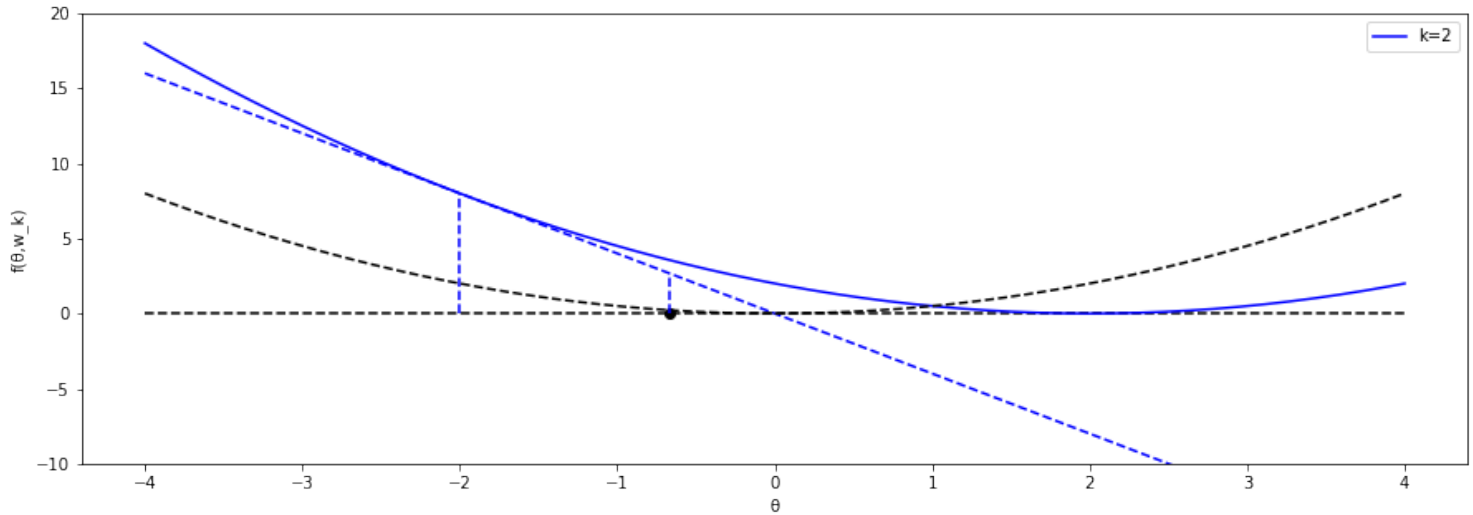
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



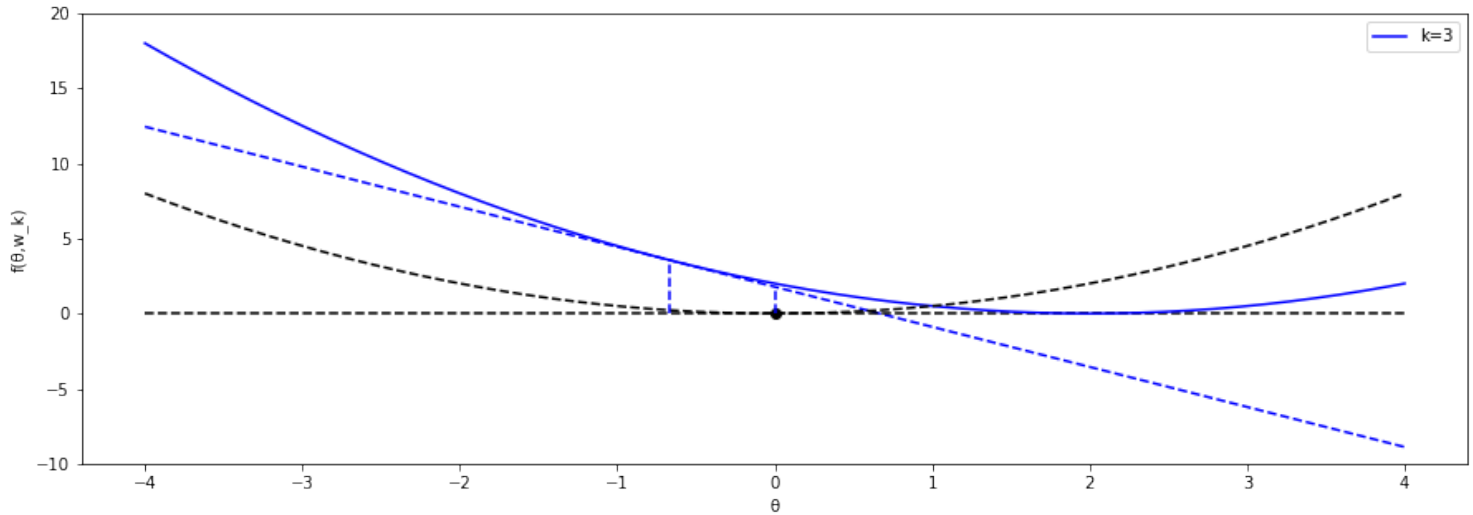
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



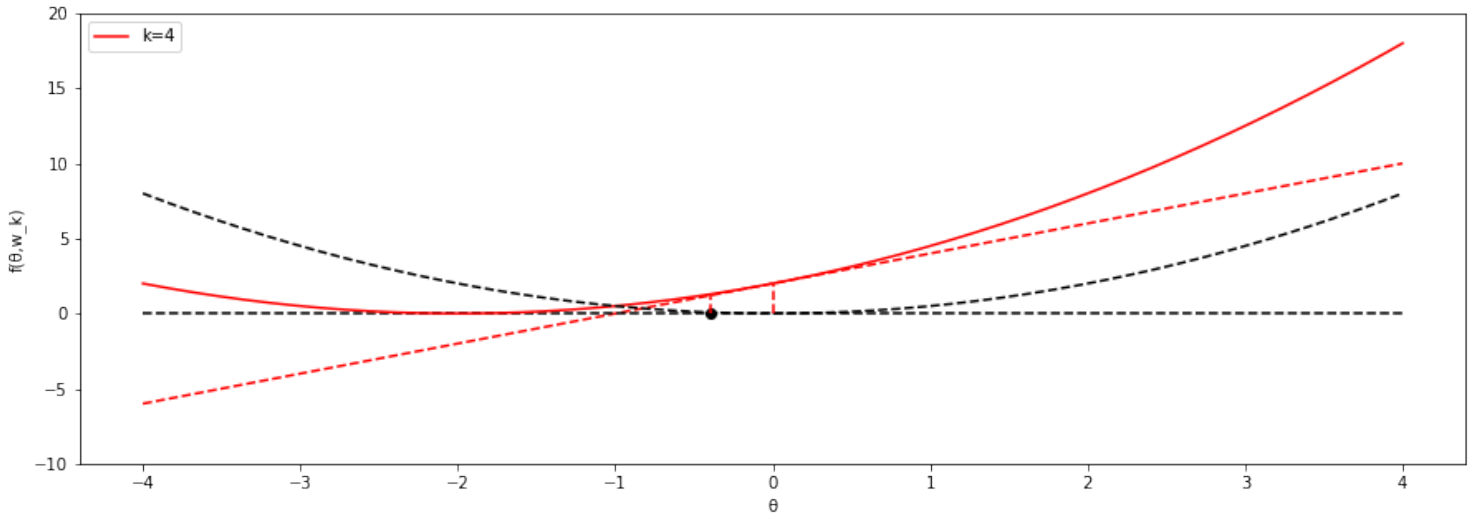
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



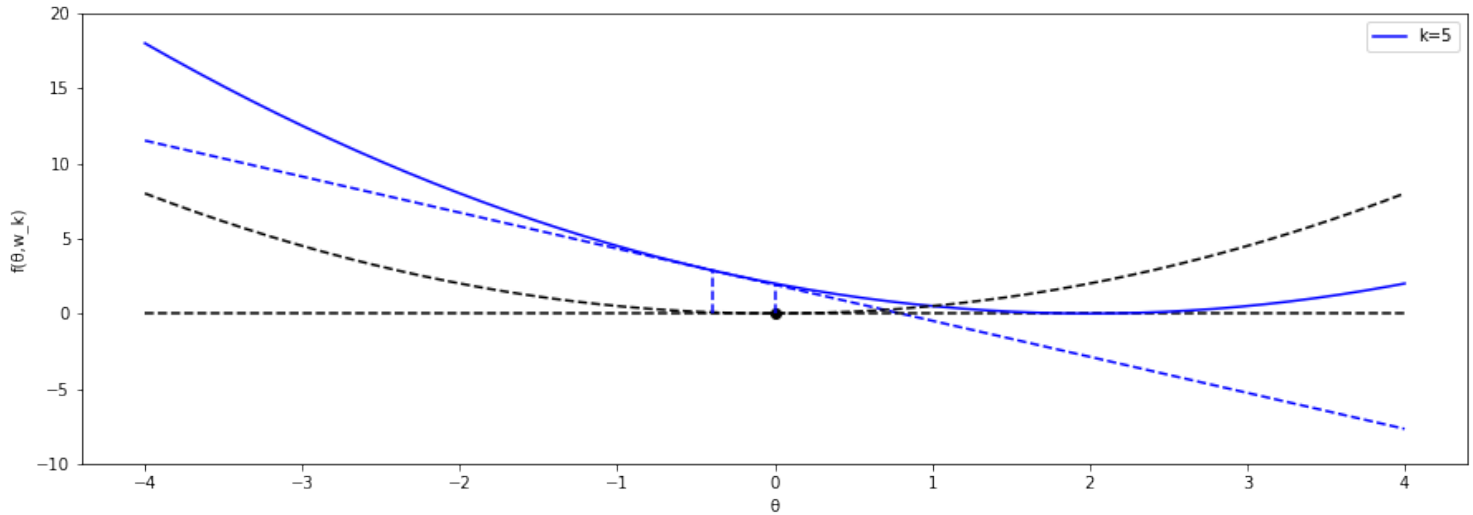
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



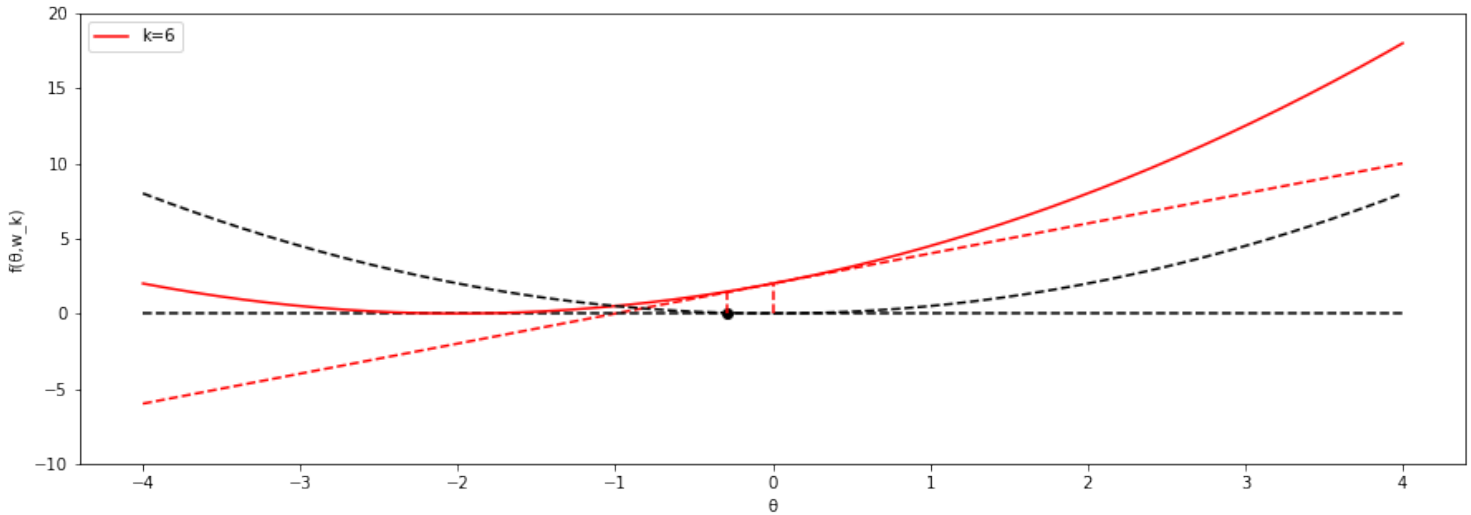
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



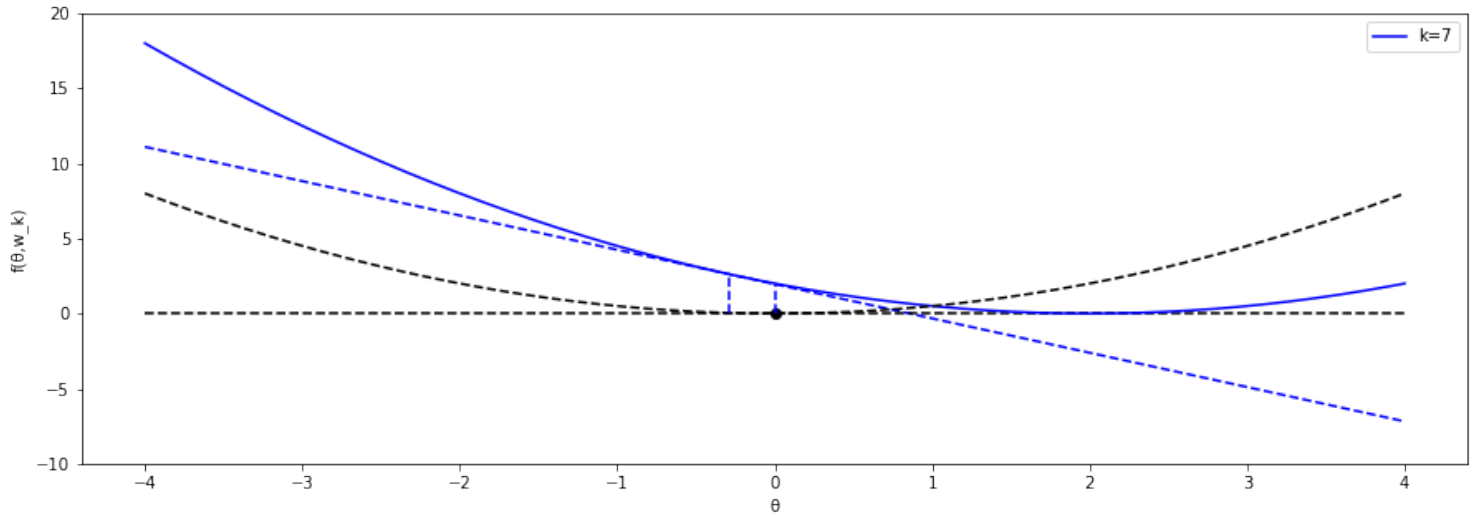
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



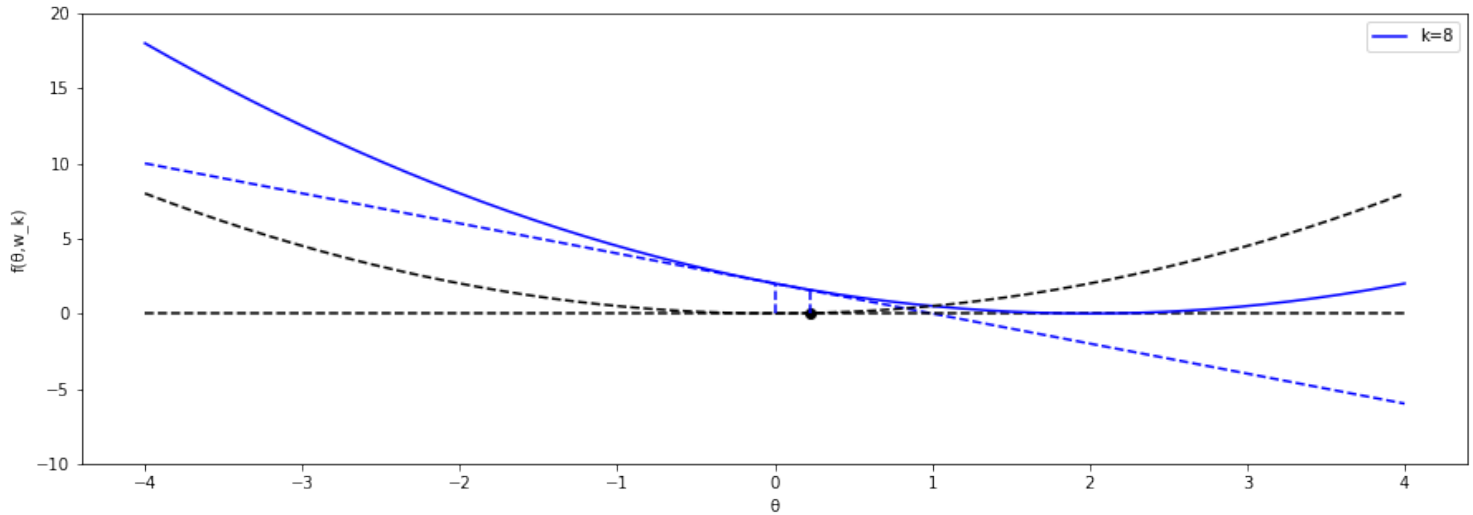
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



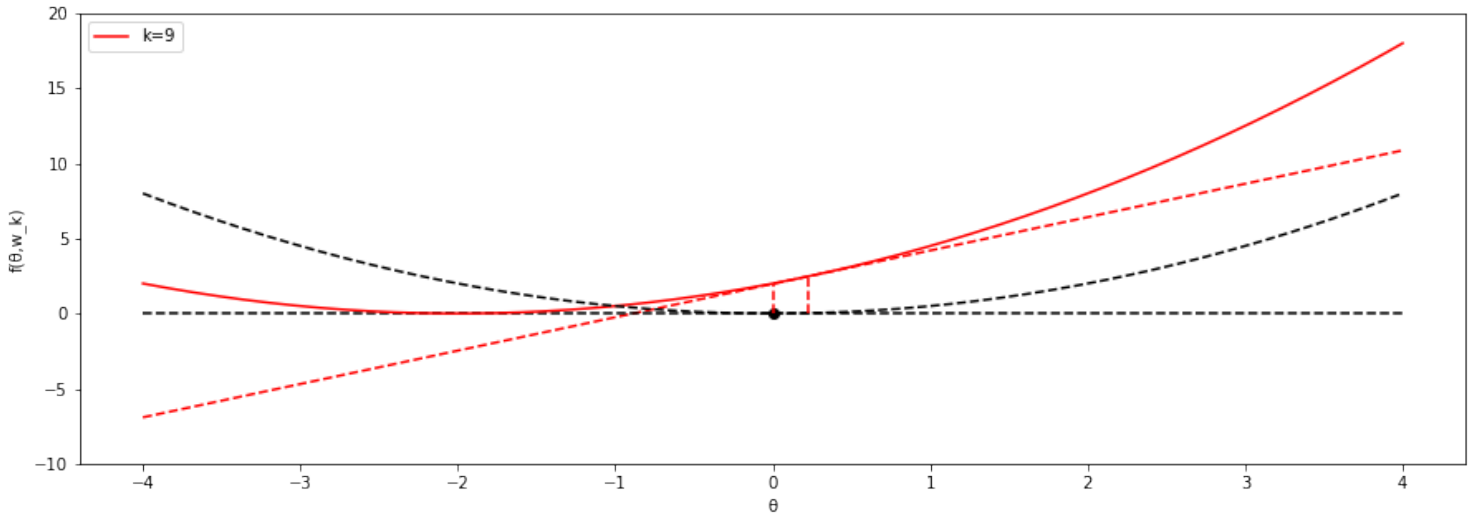
Aproximación estocástica en acción

- $\min_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



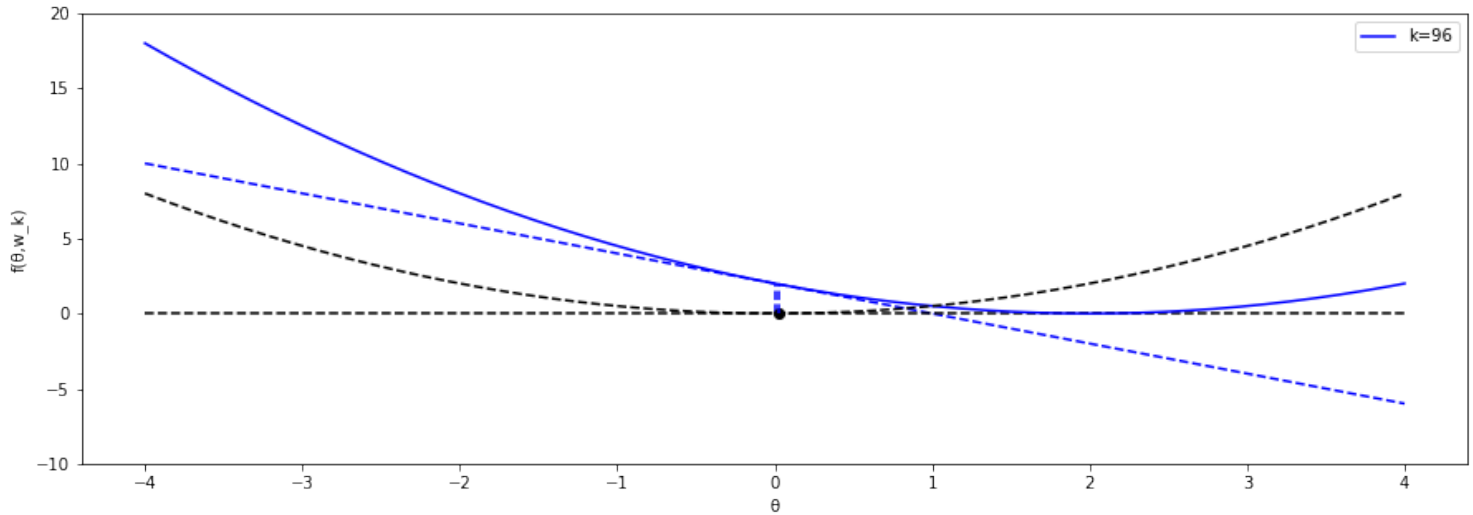
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



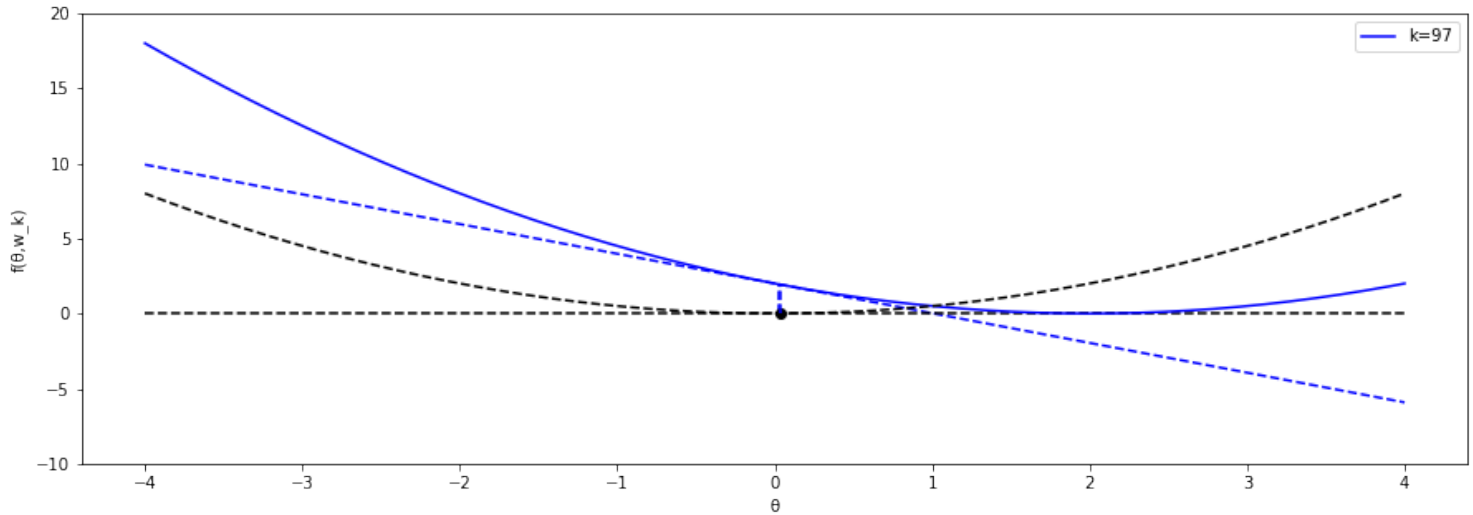
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



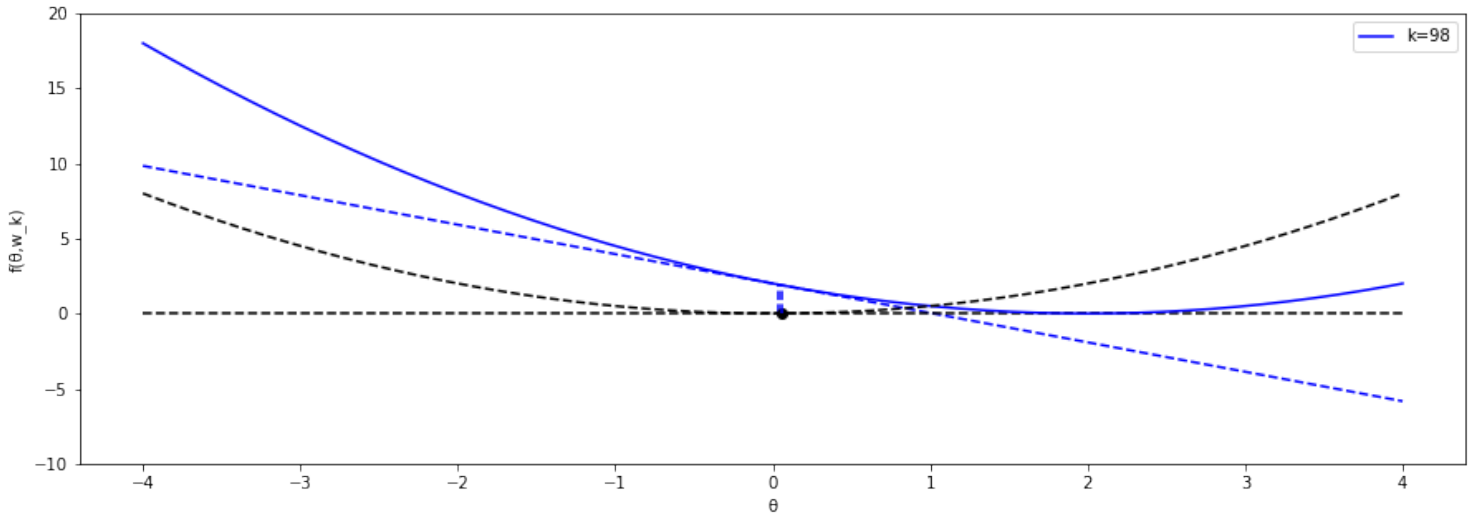
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



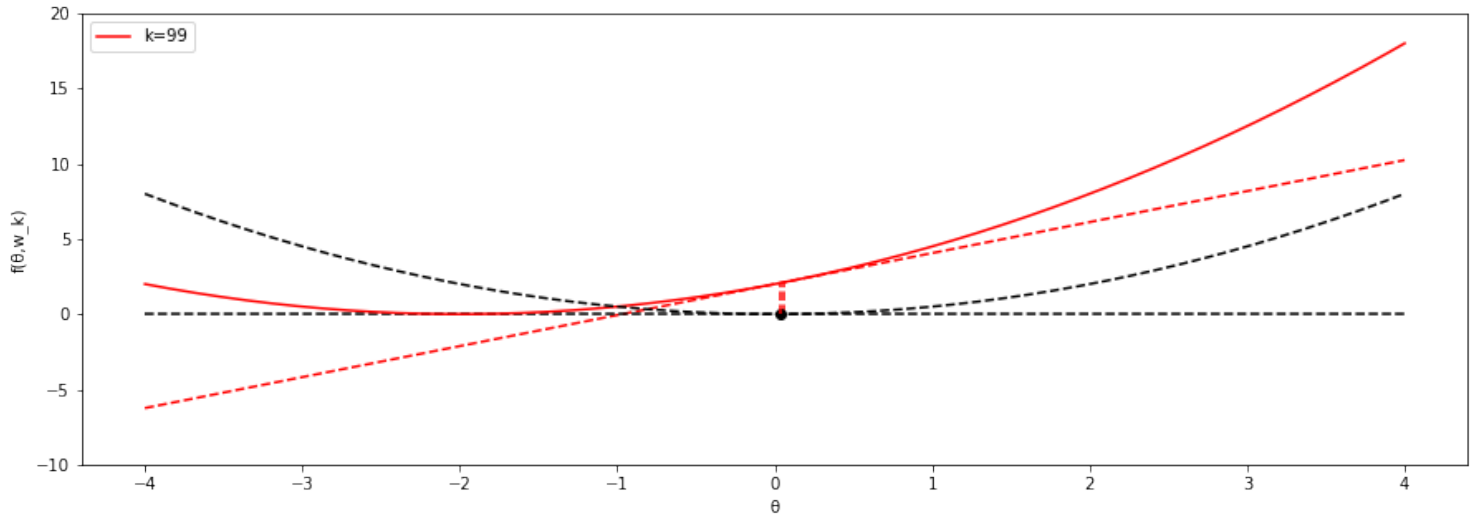
Aproximación estocástica en acción

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



Aproximación estocástica en acción

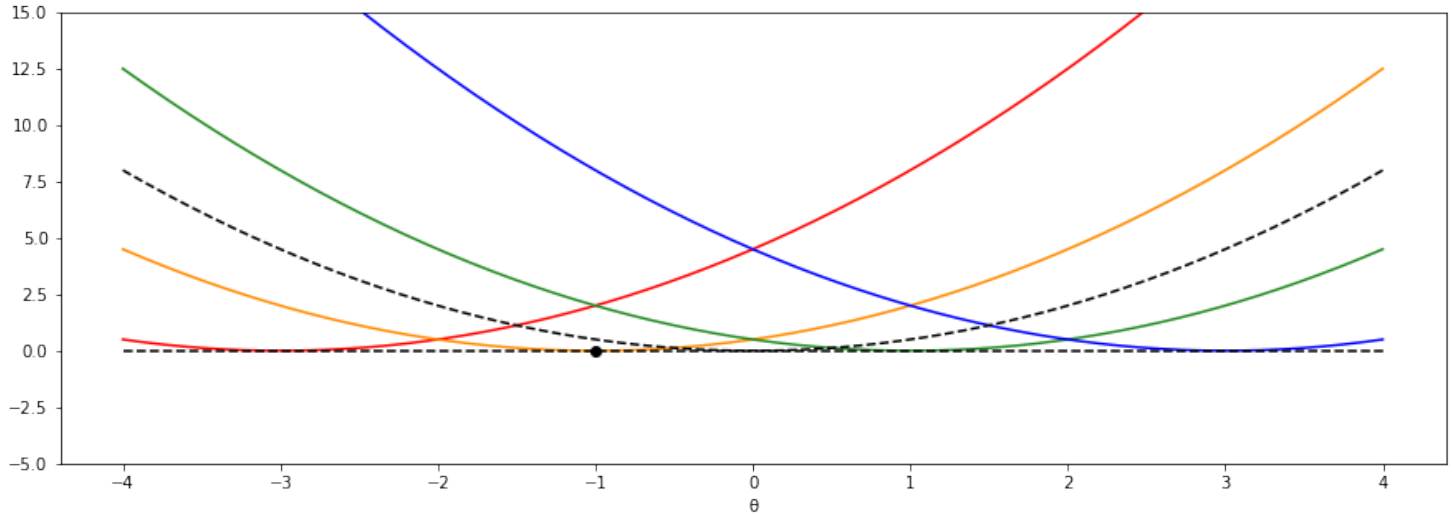
- $\min_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-2, 2\}$ con prob. $1/2$
- $E_\xi[(\theta - \xi)^2] = \frac{1}{2}(\theta - 2)^2 + \frac{1}{2}(\theta + 2)^2 = \theta^2 + 2 \Rightarrow \theta* = 0$



- Stochastic gradient descent minimiza acumulando al mismo tiempo la esperanza

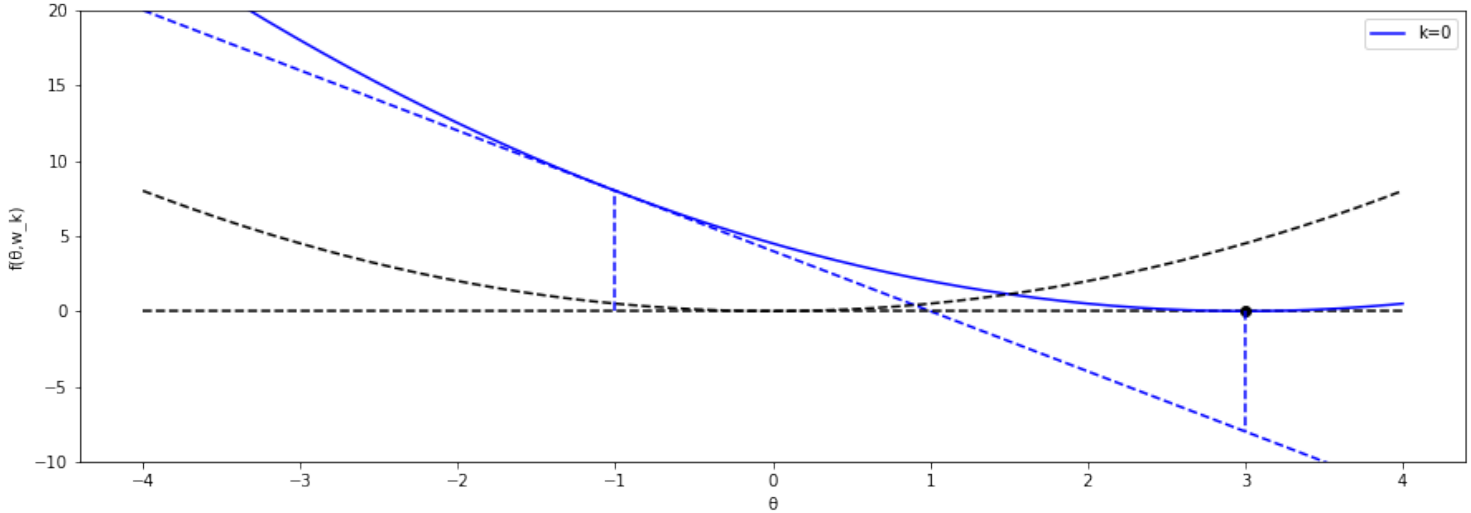
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



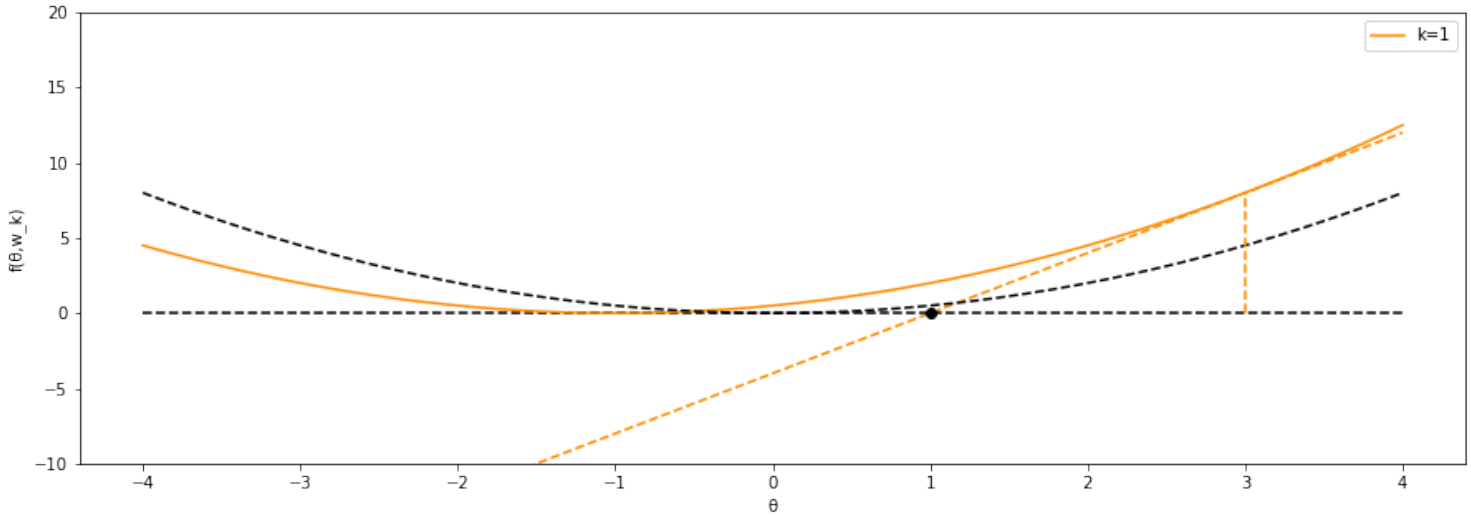
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



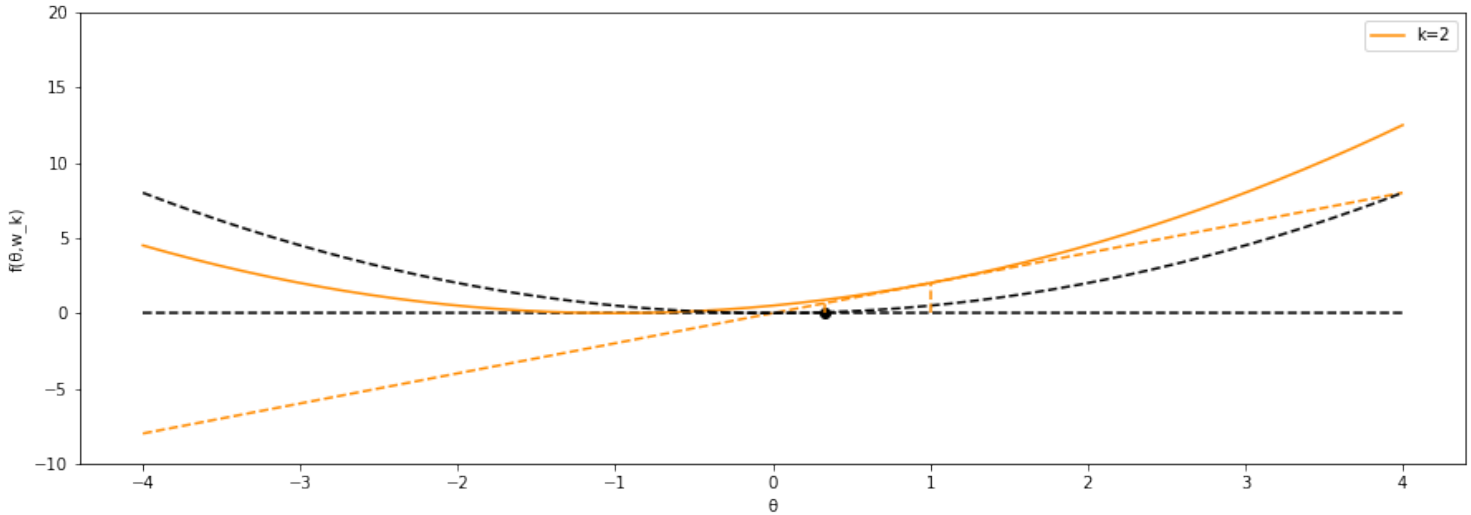
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



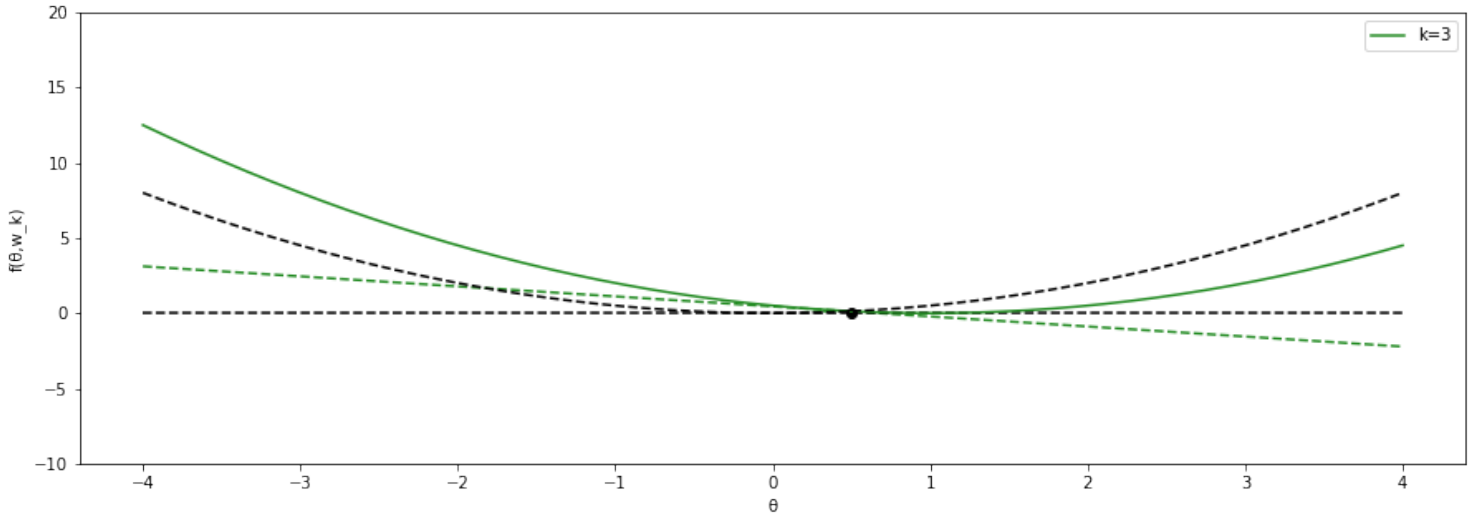
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



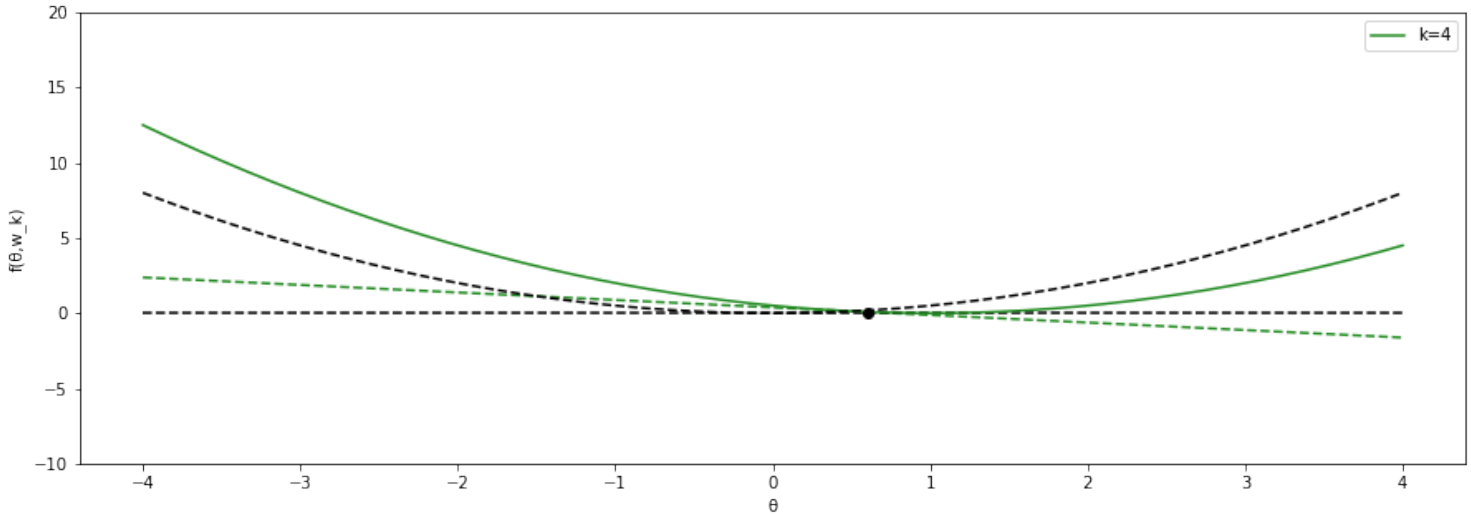
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



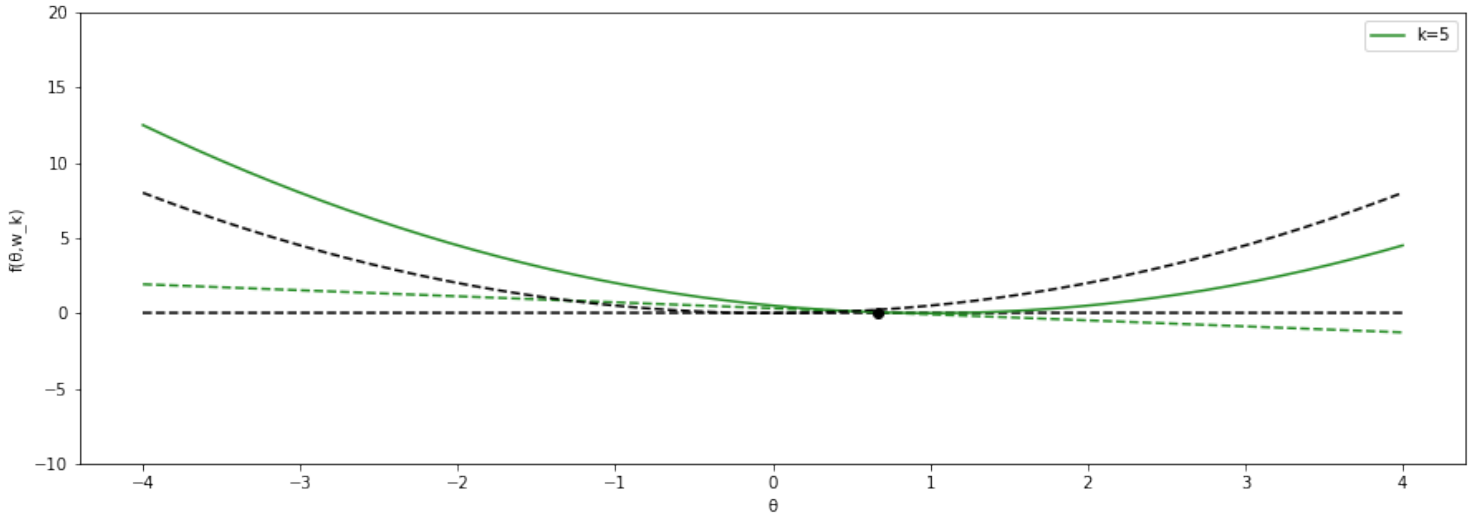
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



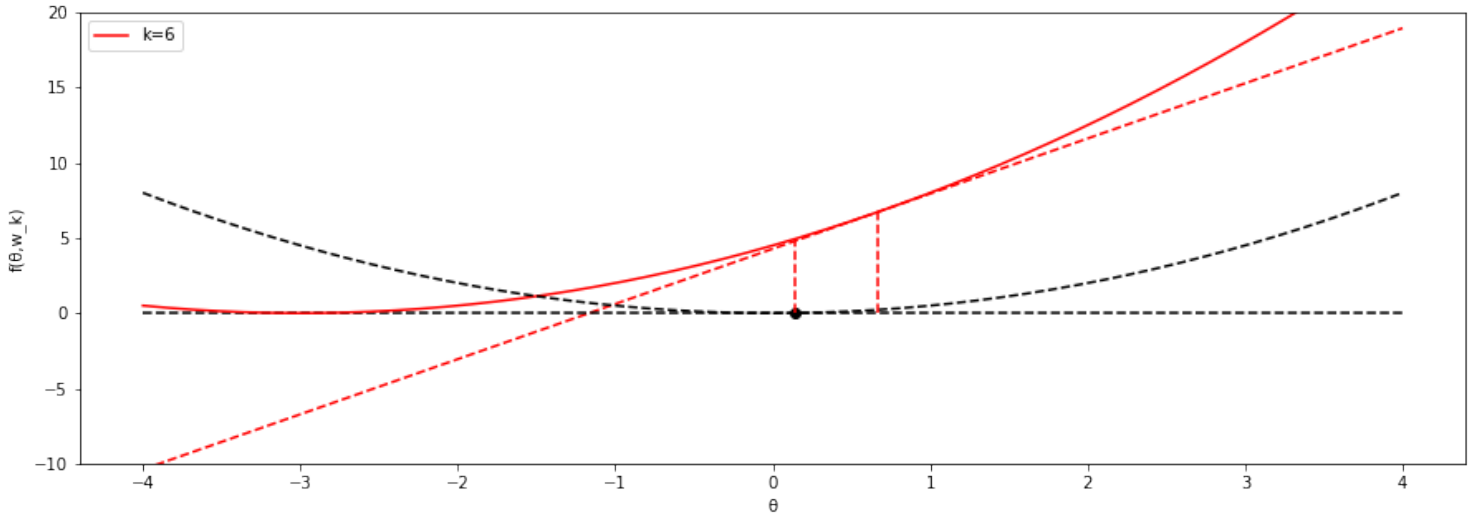
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



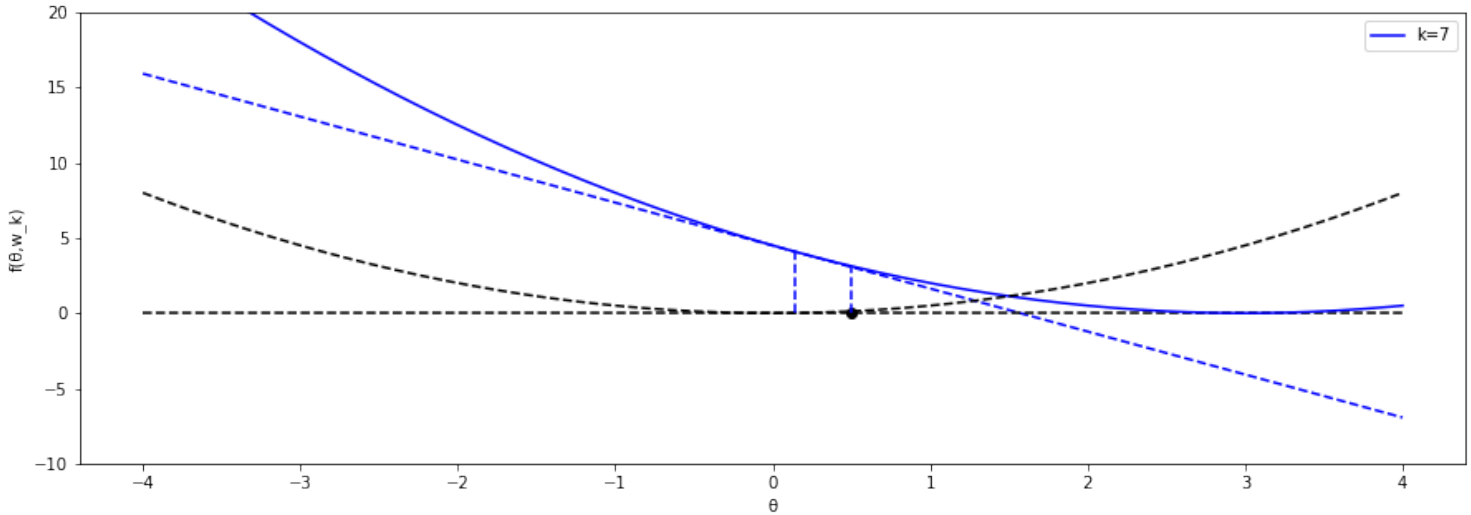
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



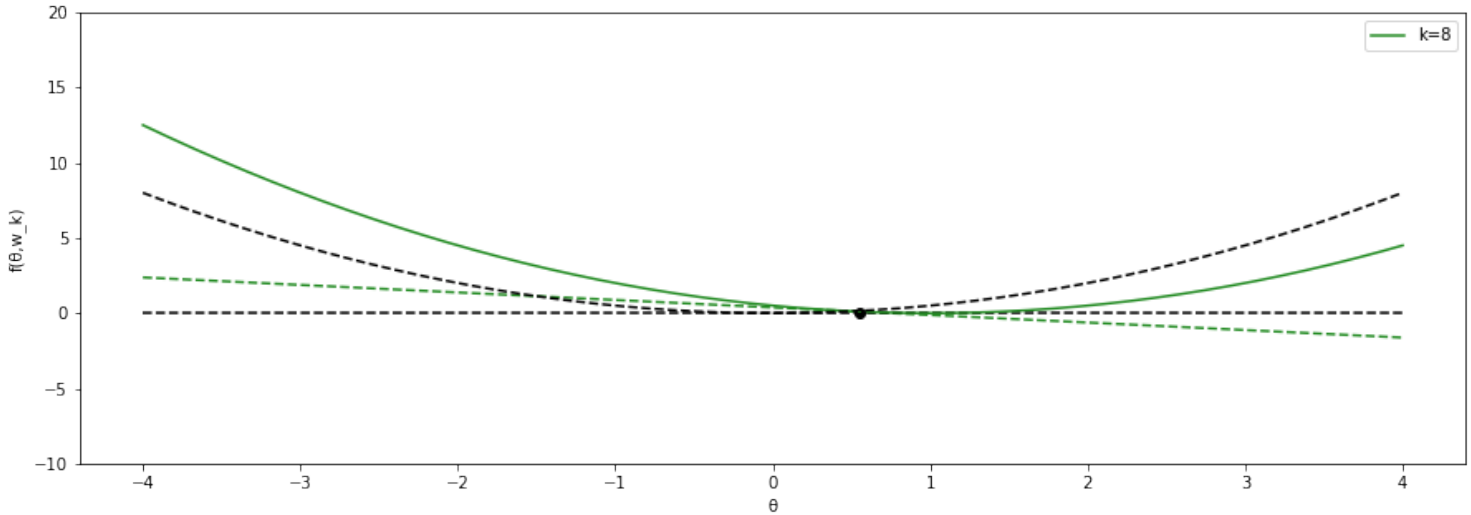
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



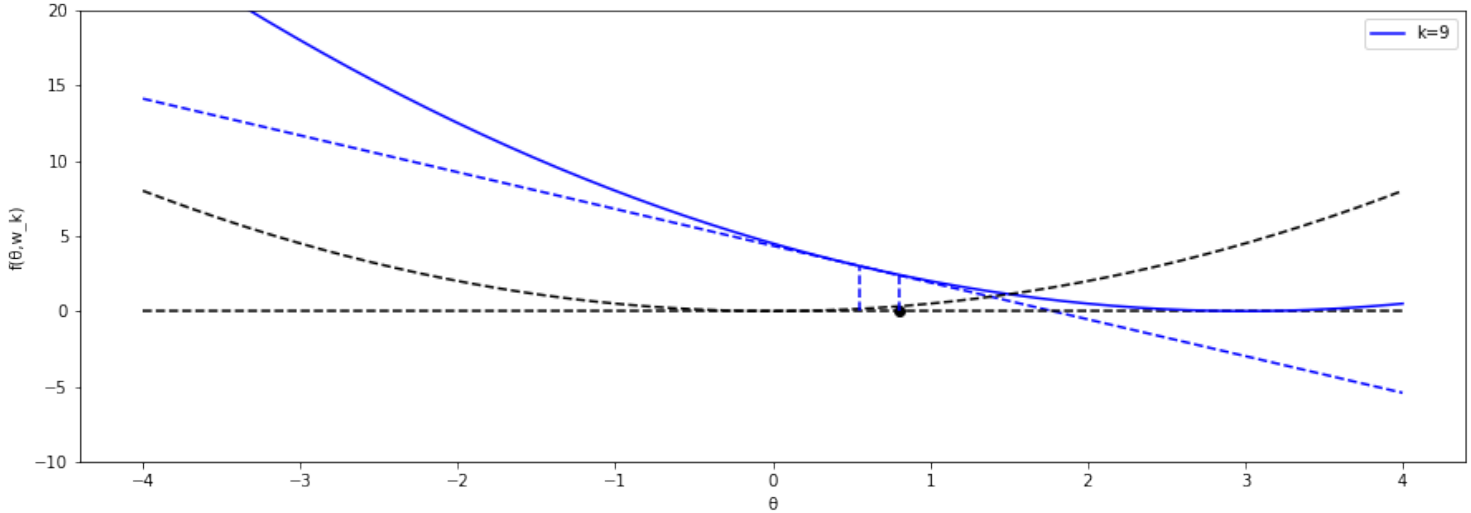
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



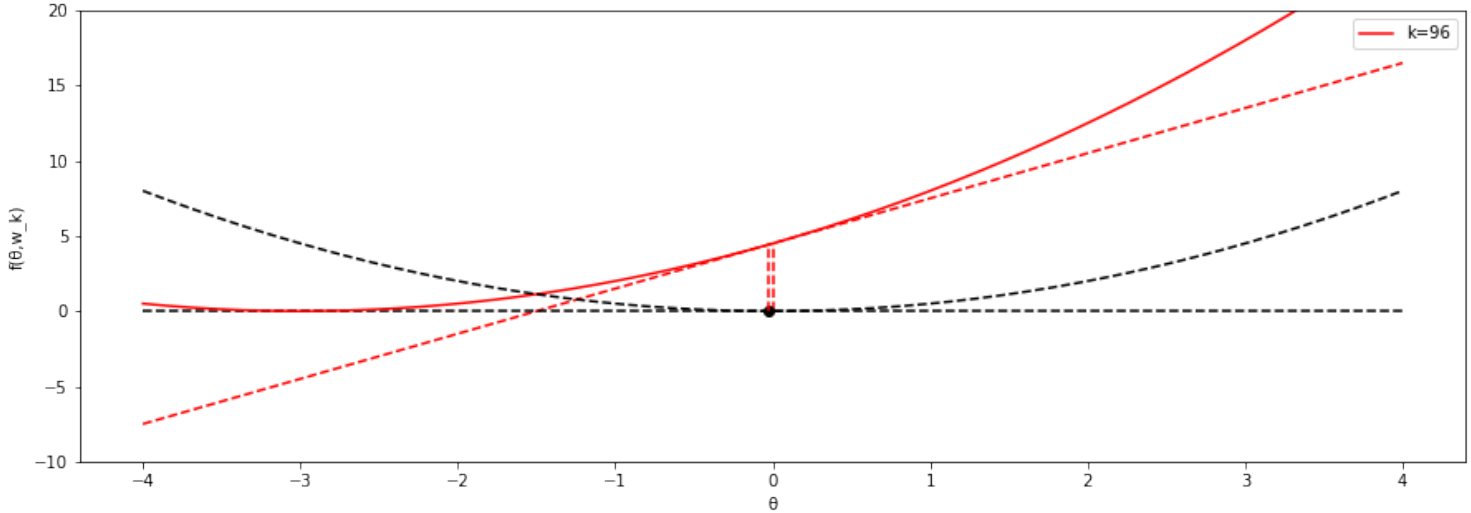
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



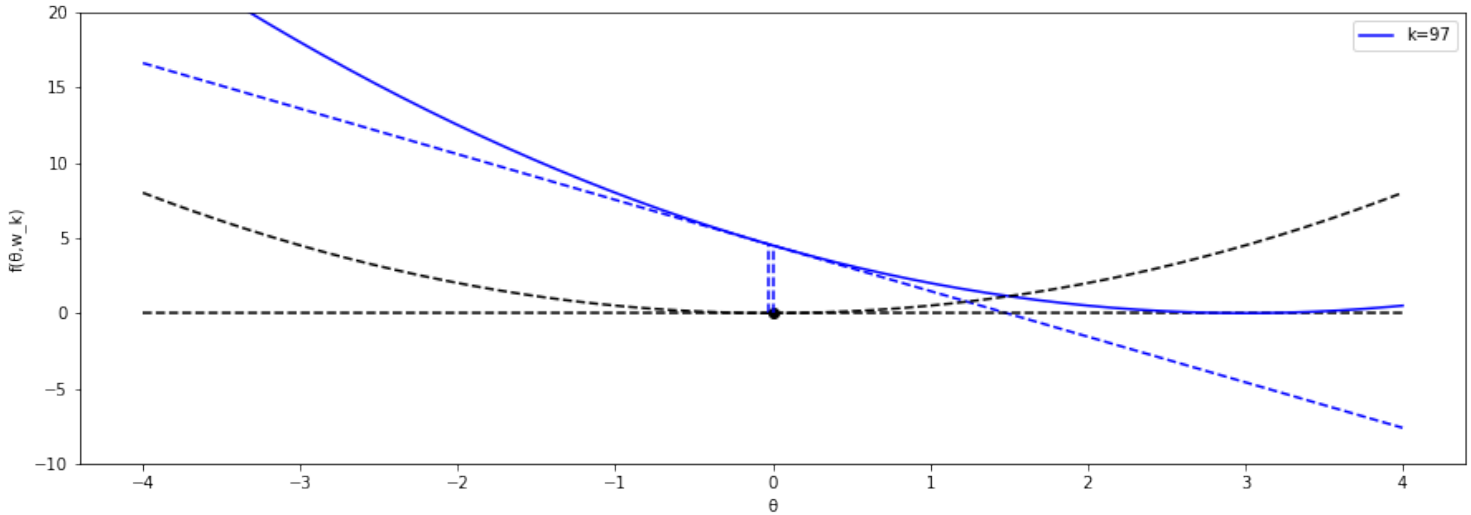
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



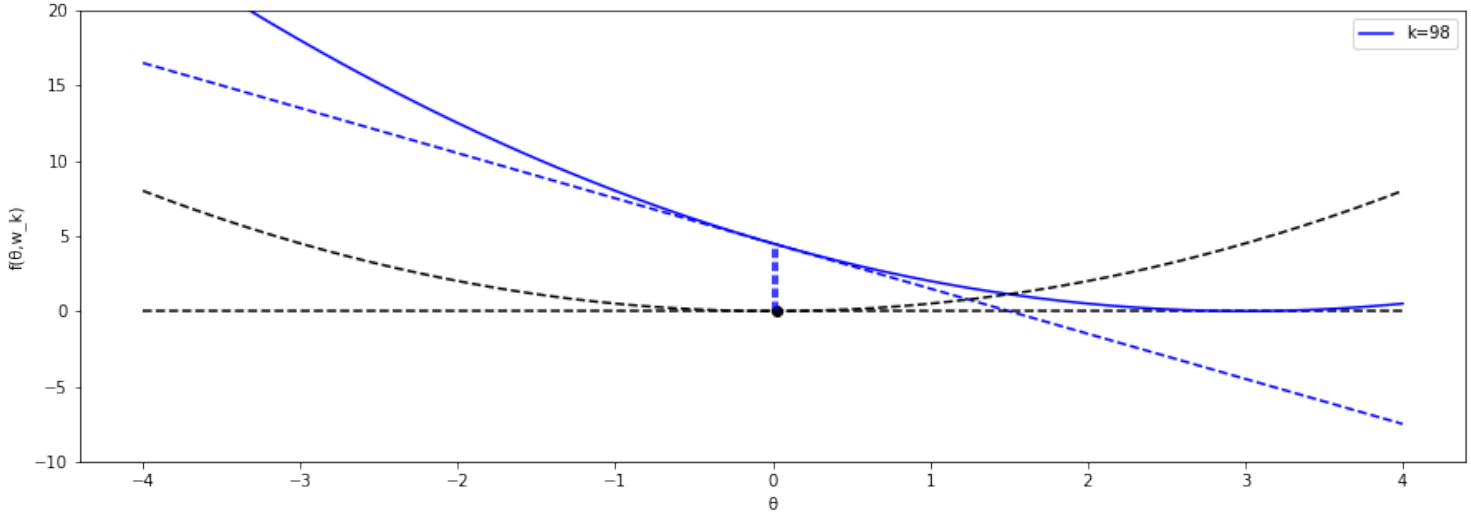
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



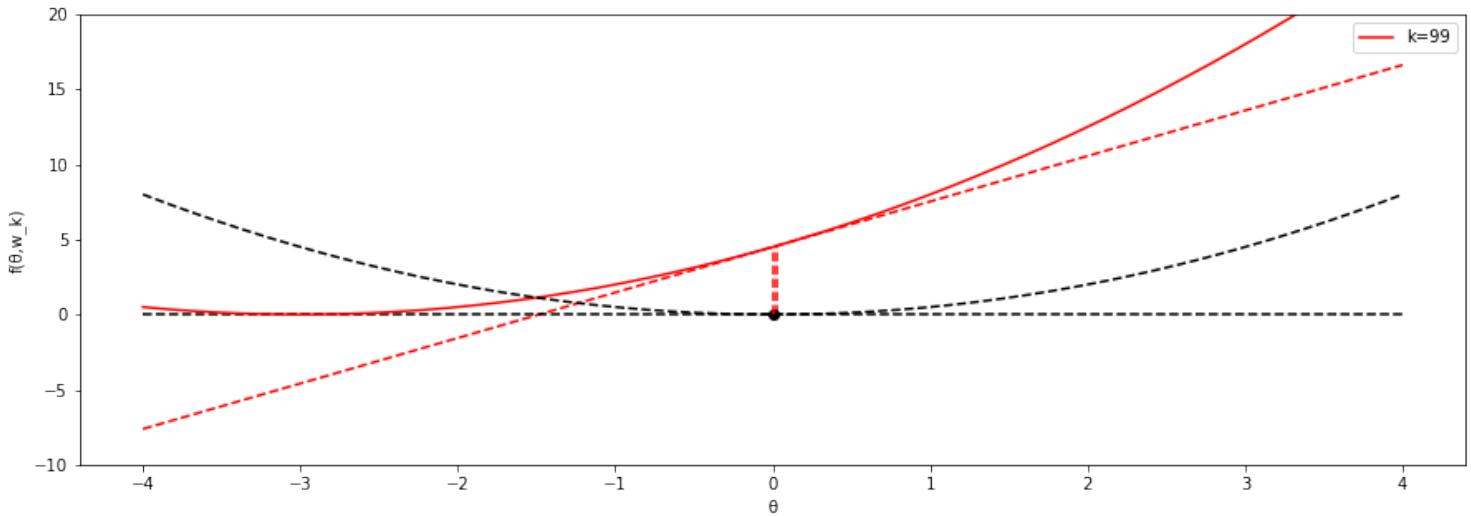
SGD no binario

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



SGD no binario

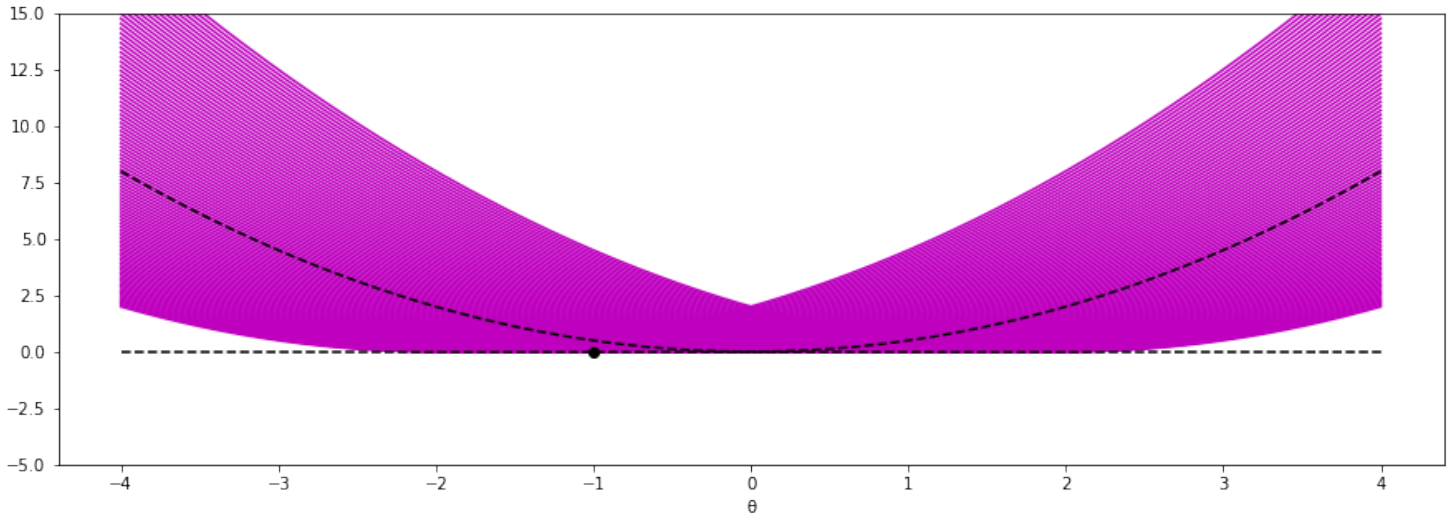
- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \in \{-3, , -1, 1, 3\}$ con prob. $1/4$
- $E_\xi[(\theta - \xi)^2] = \theta^2 + 5 \Rightarrow \theta* = 0$



- Stochastic gradient descent funciona con variables no binarias

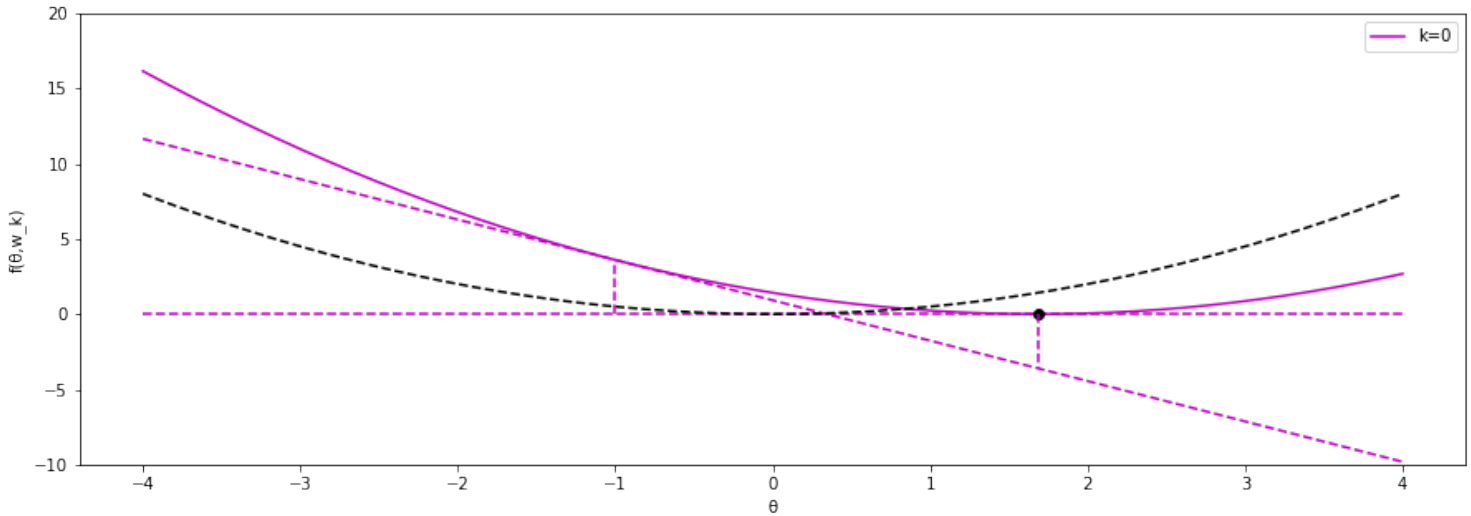
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



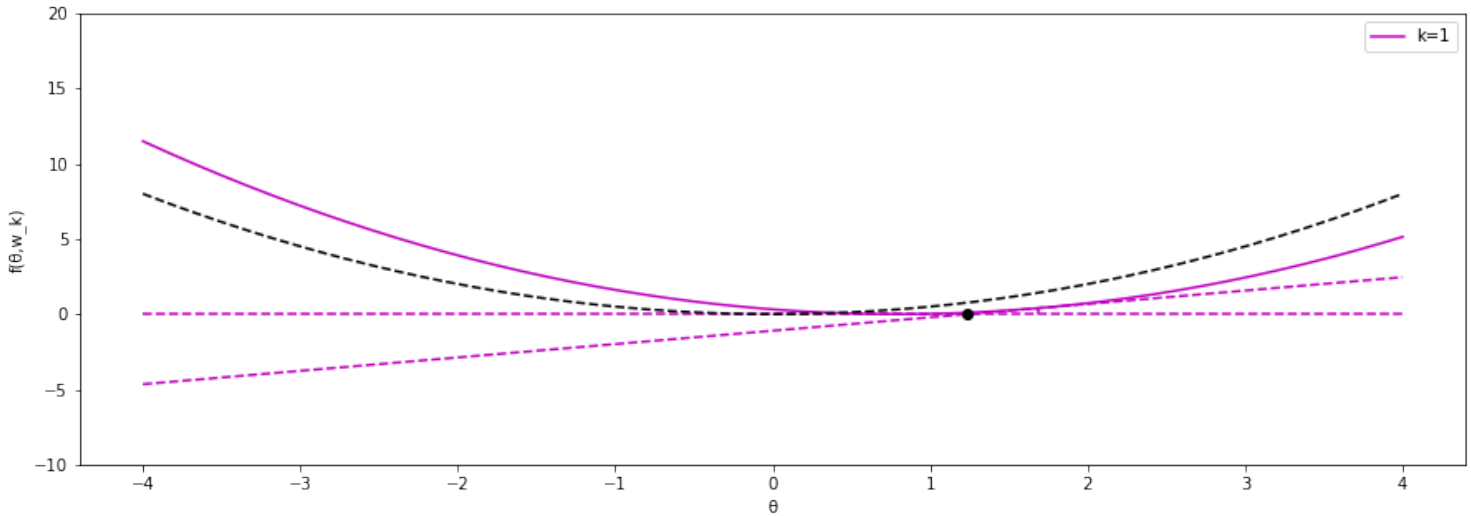
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



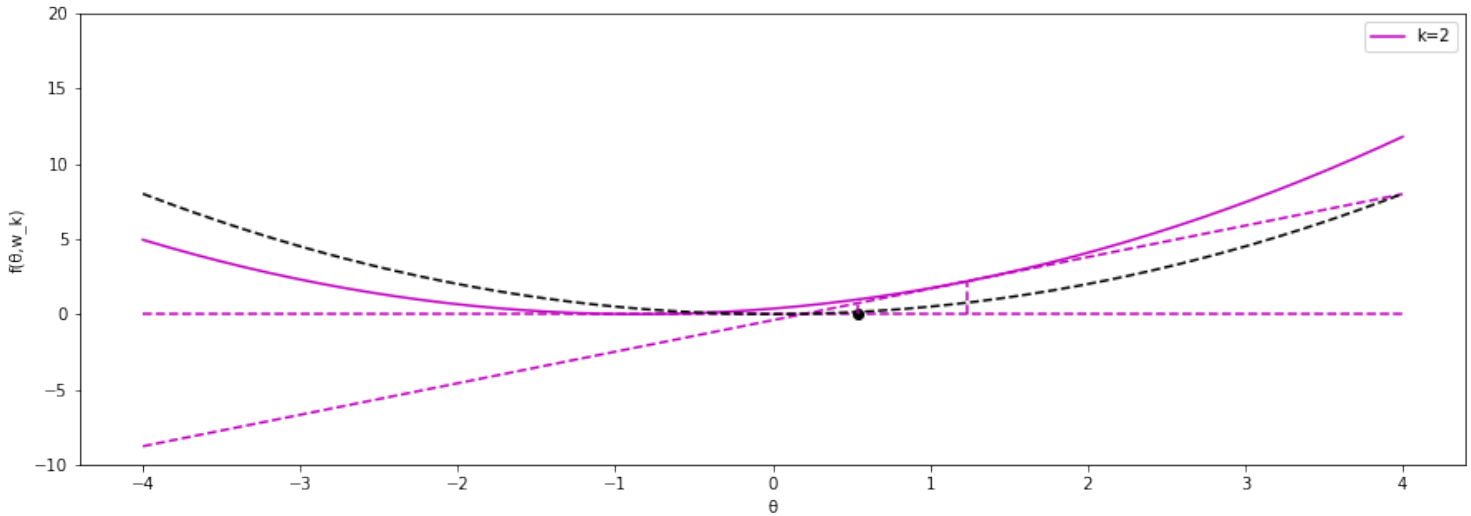
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



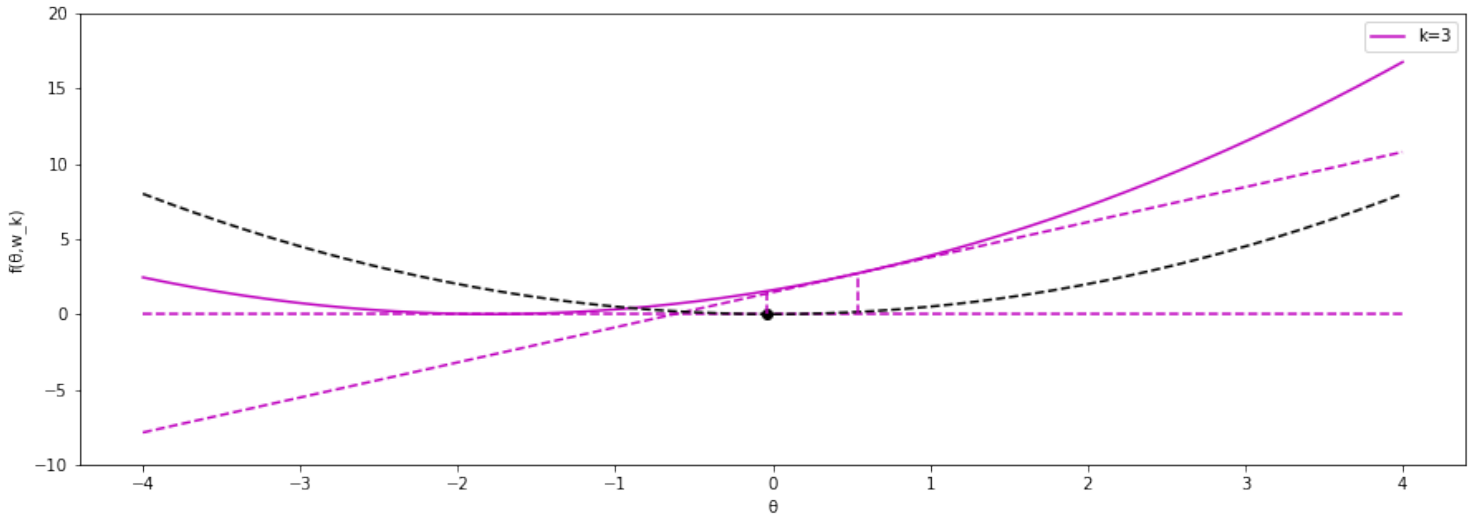
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



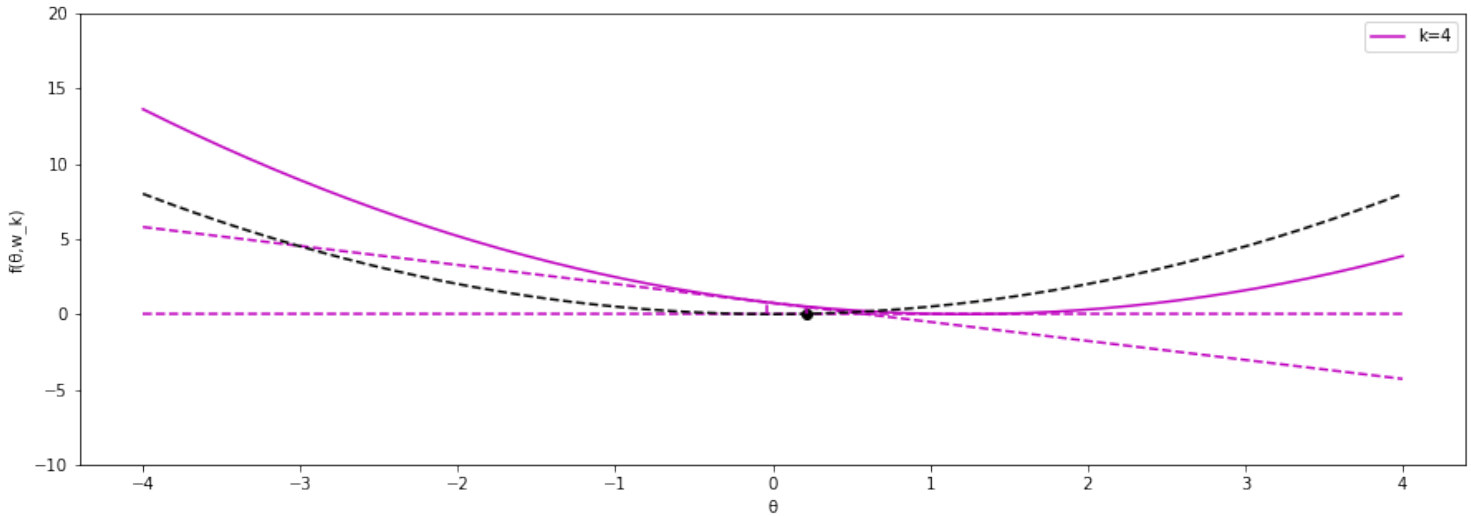
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



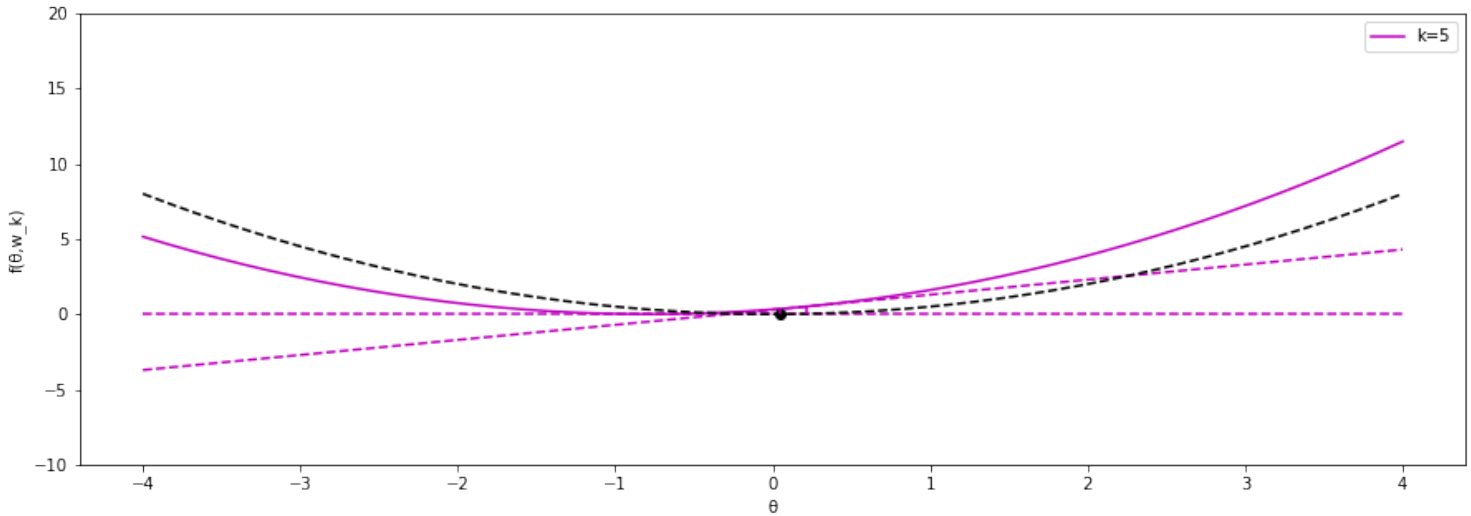
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



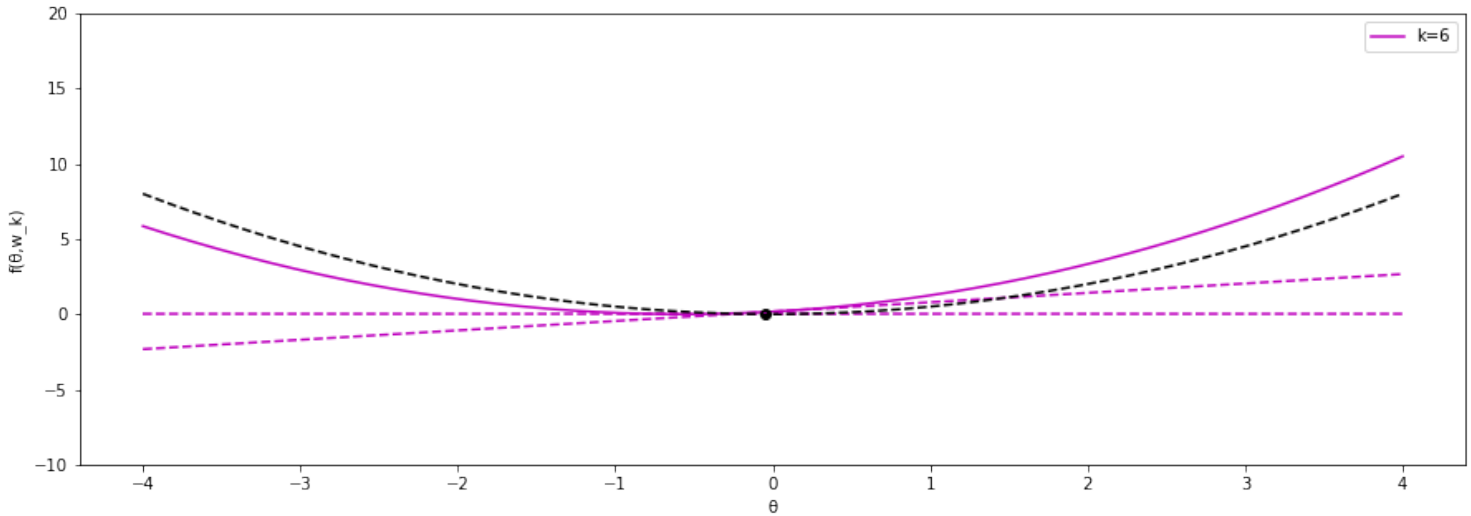
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



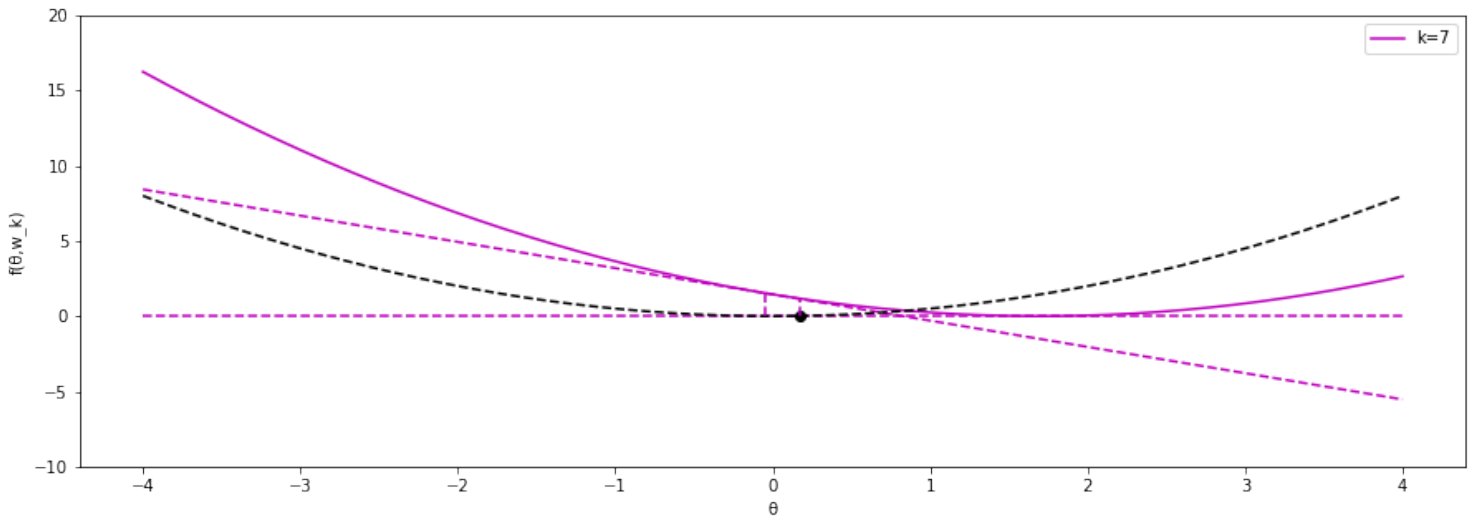
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



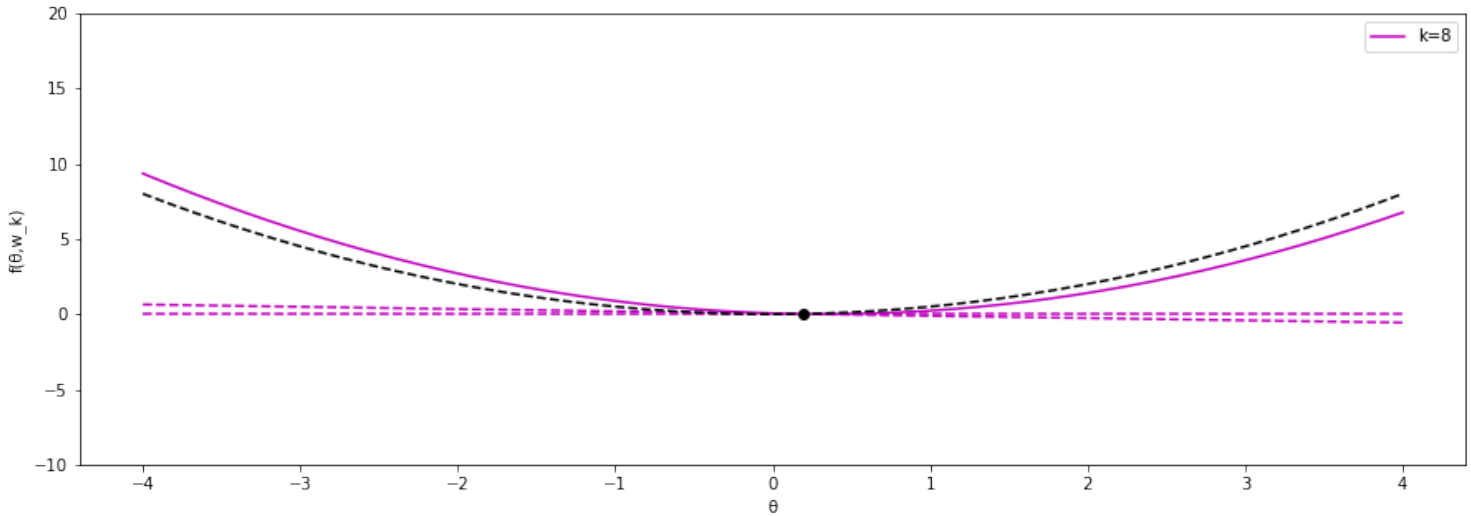
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



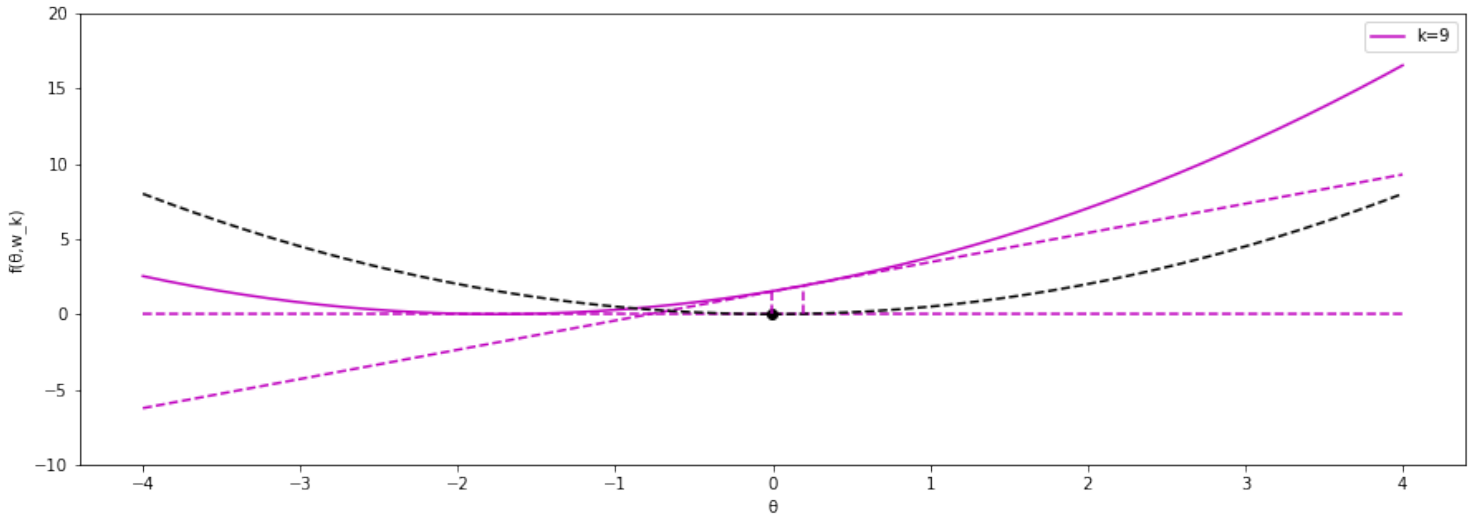
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



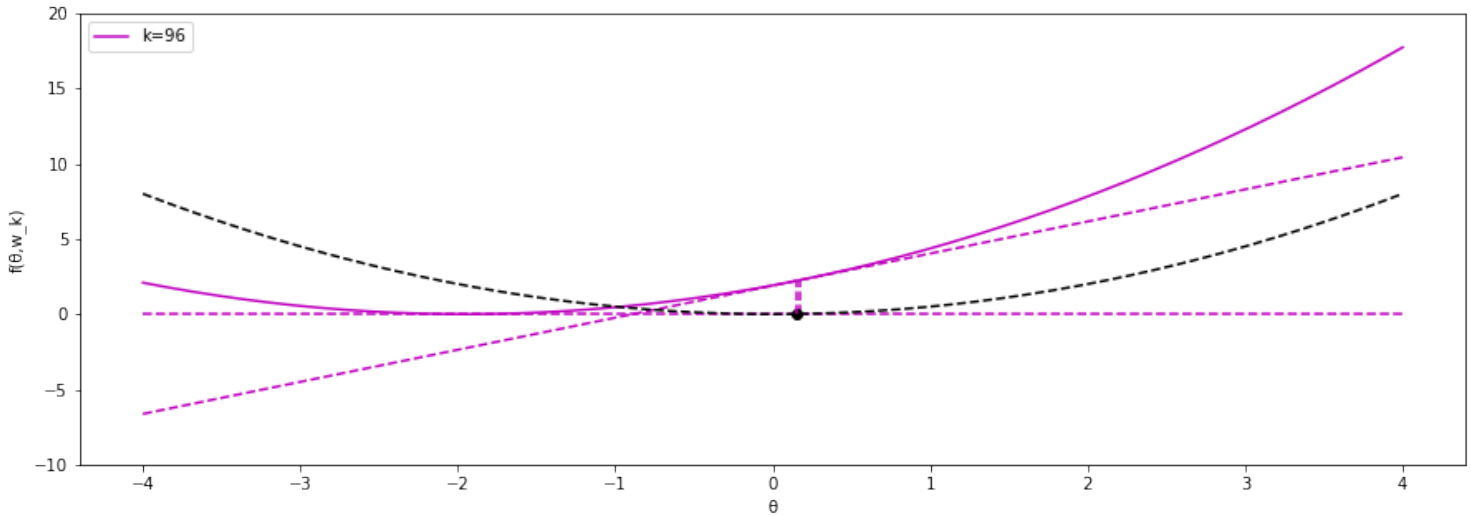
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



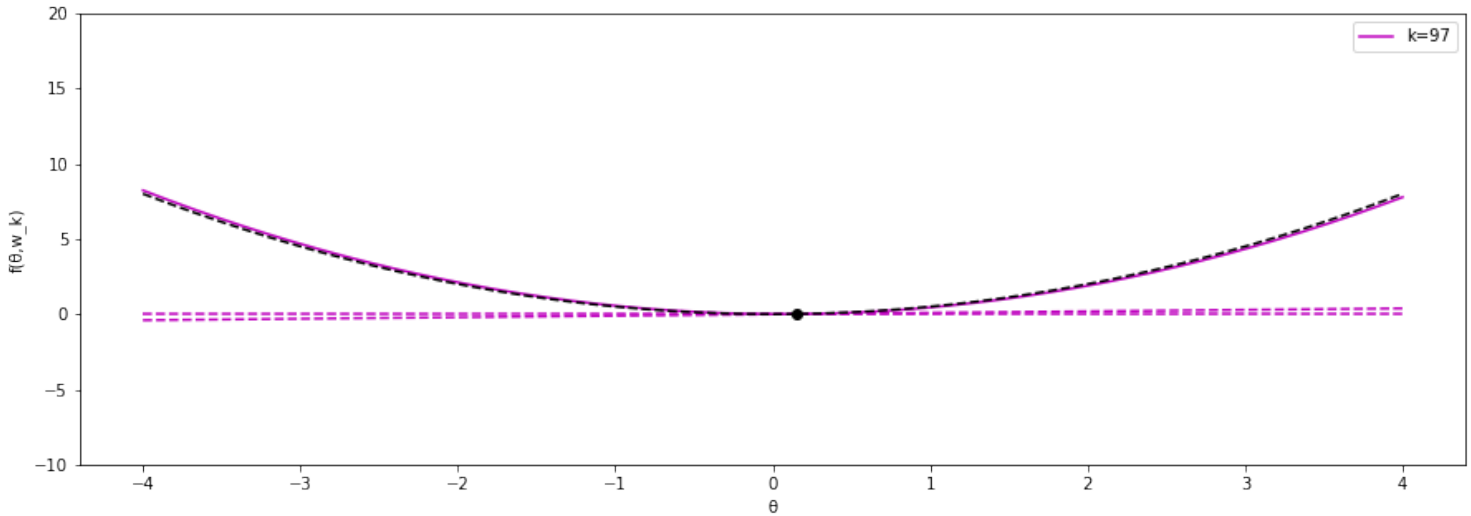
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



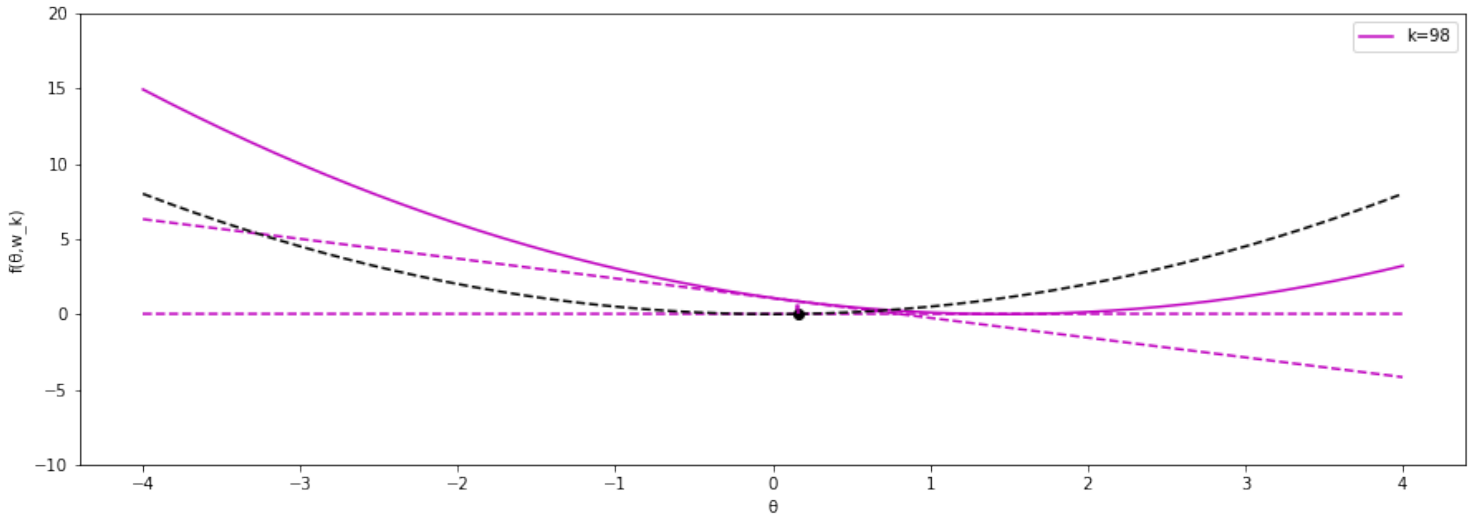
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



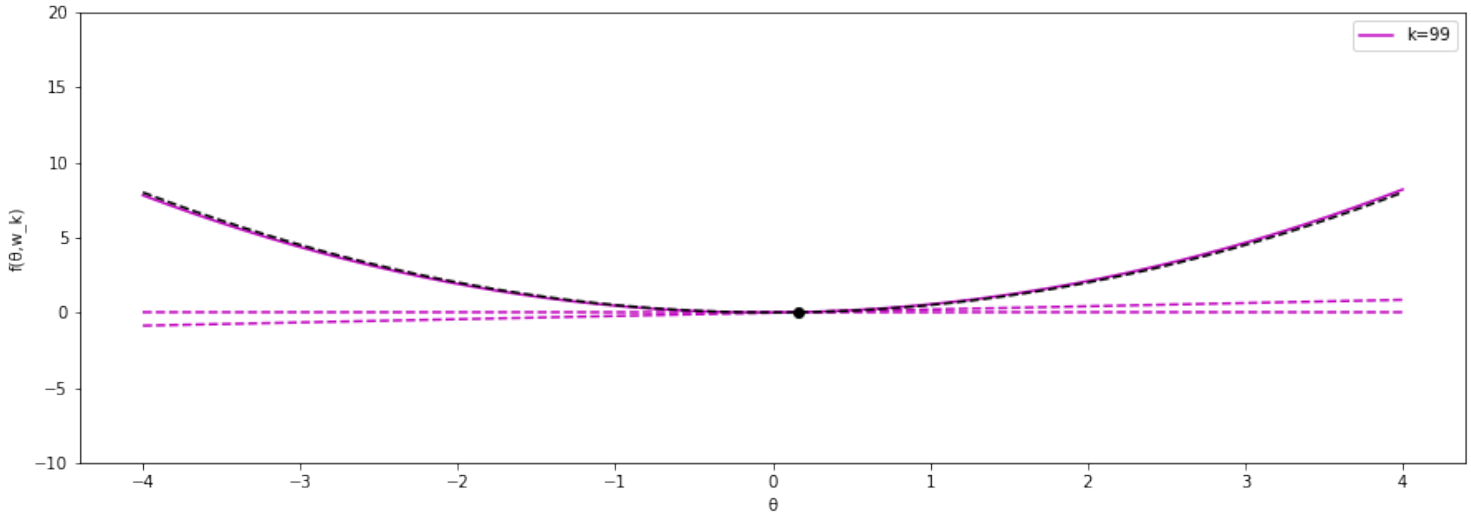
SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



SGD con variables continuas

- $\text{mín}_\theta \frac{1}{2} E_\xi[(\theta - \xi)^2]$ donde $\xi \sim \mathcal{U}[-2, 2]$
- $E_\xi[(\theta - \xi)^2] = \int_{-2}^{2} \frac{1}{4}(\theta - \xi)^2 d\xi = \theta^2 + \frac{4}{3} \Rightarrow \theta* = 0$



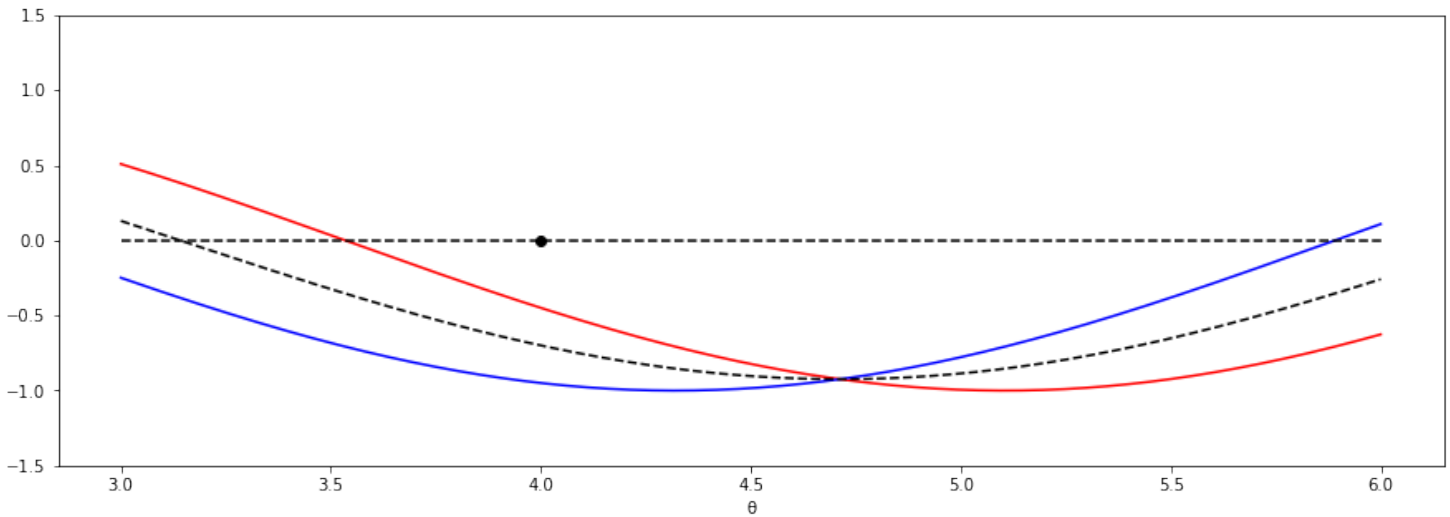
- Stochastic gradient descent funciona con variables continuas

Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

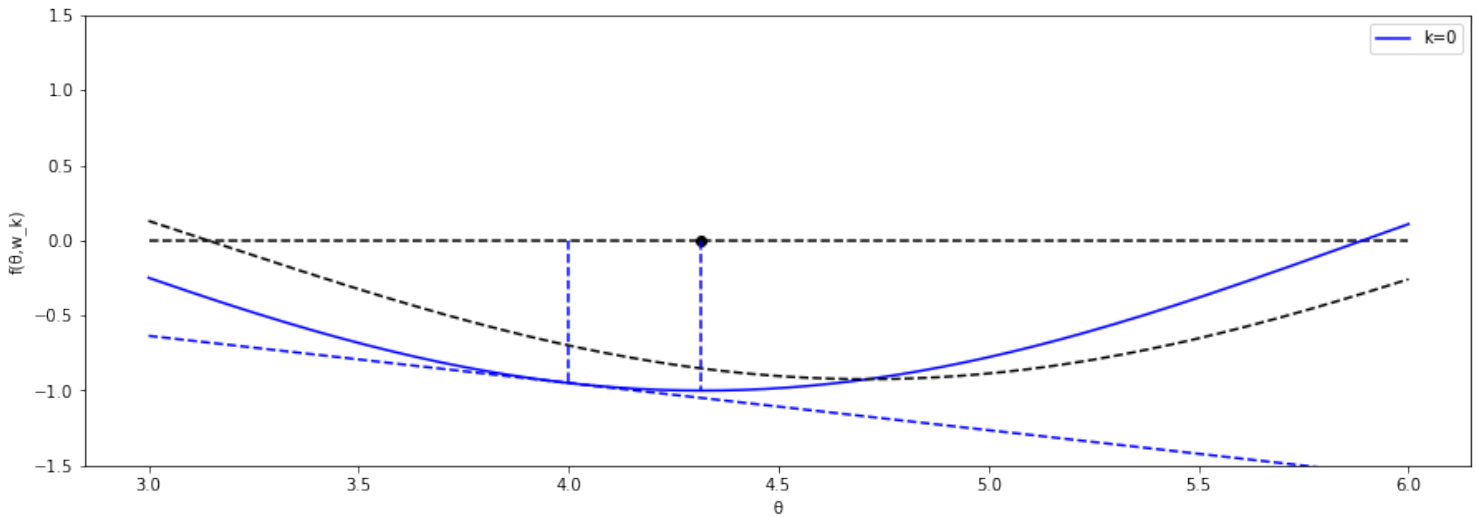


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

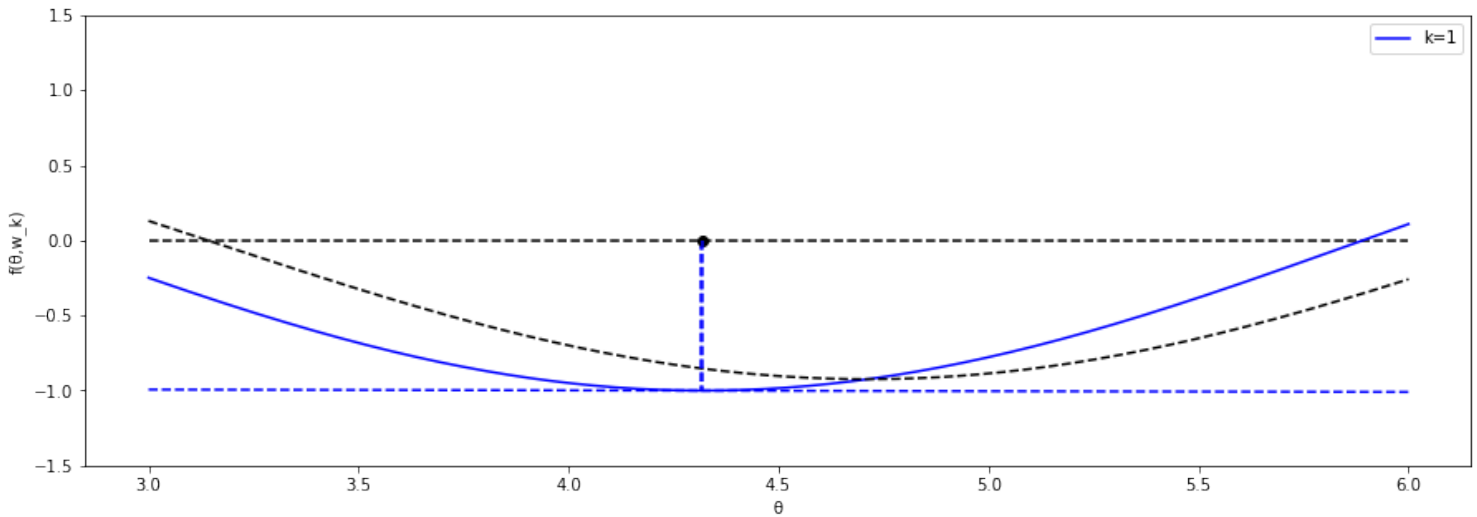


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

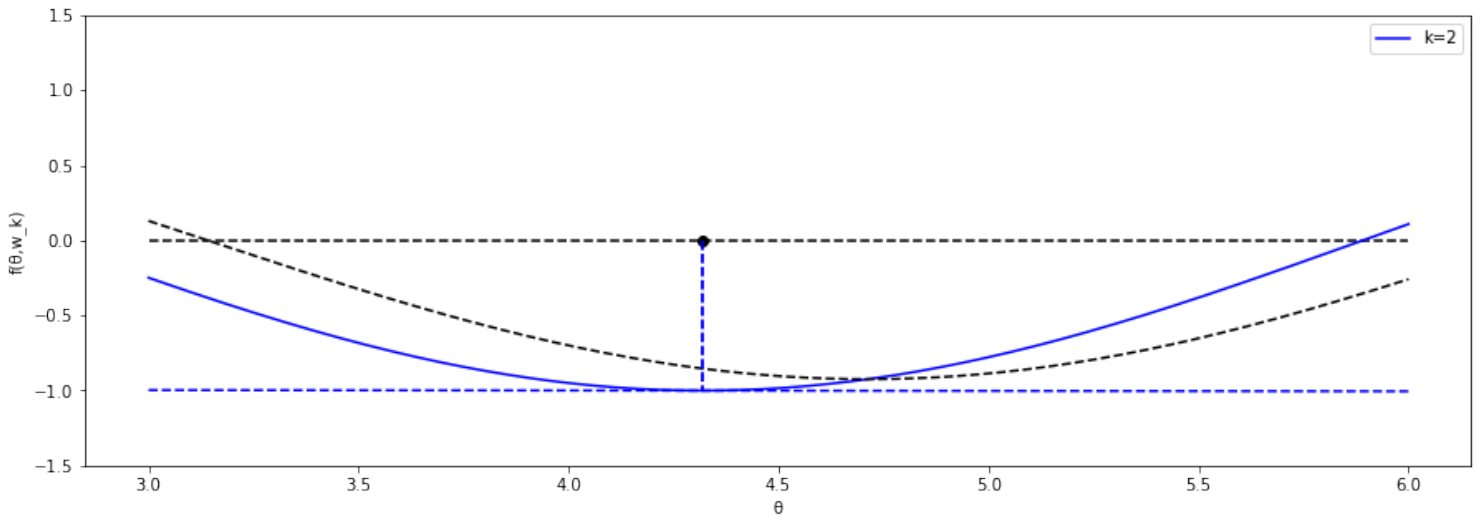


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

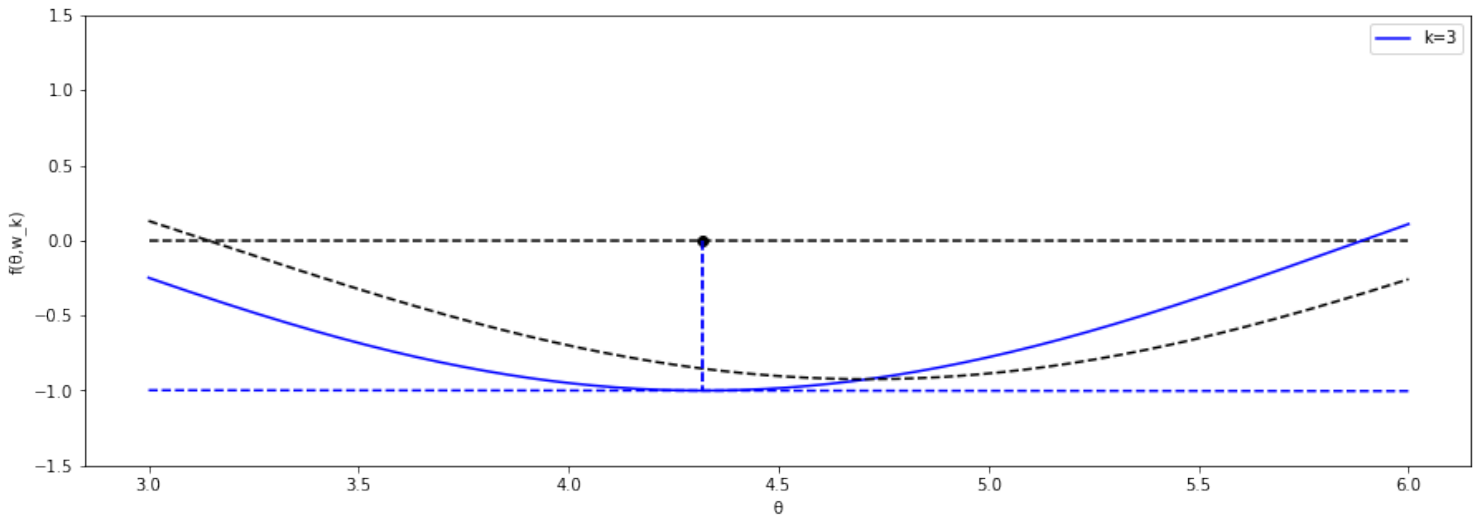


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k \frac{\pi}{8})}_{\varphi_k}$

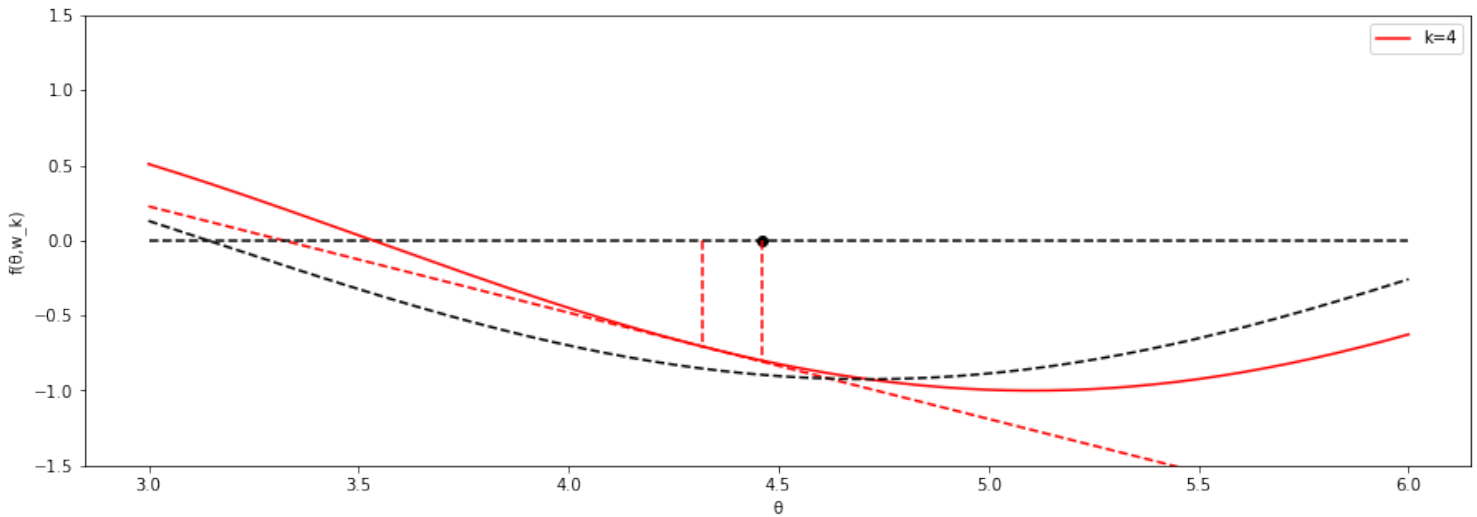


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

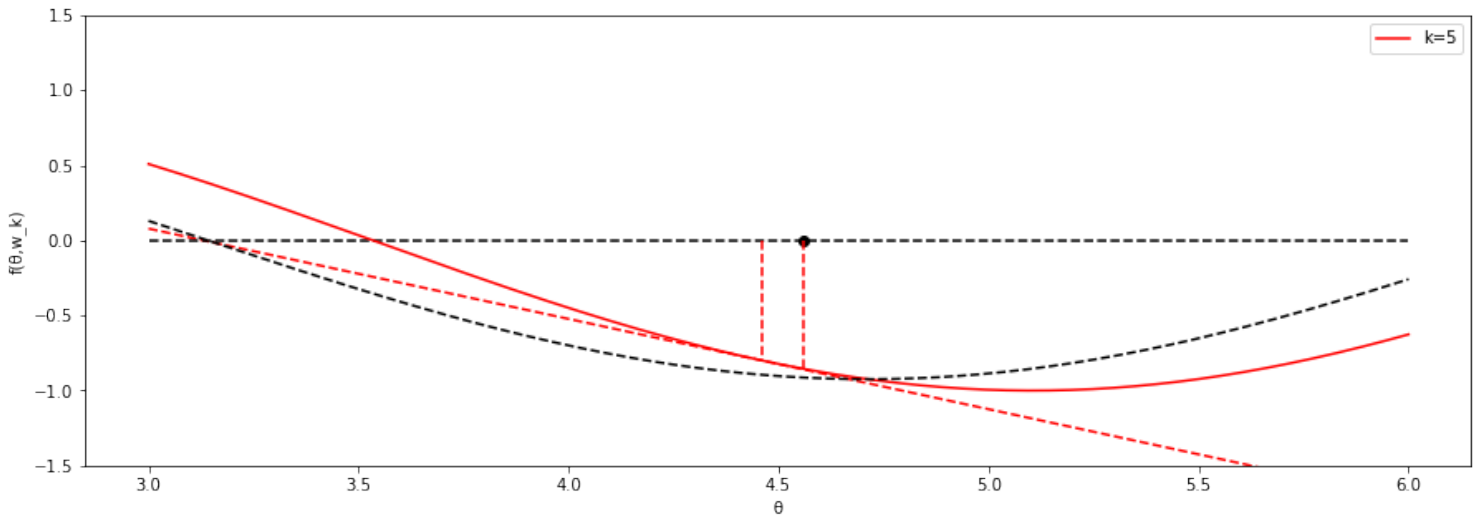


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

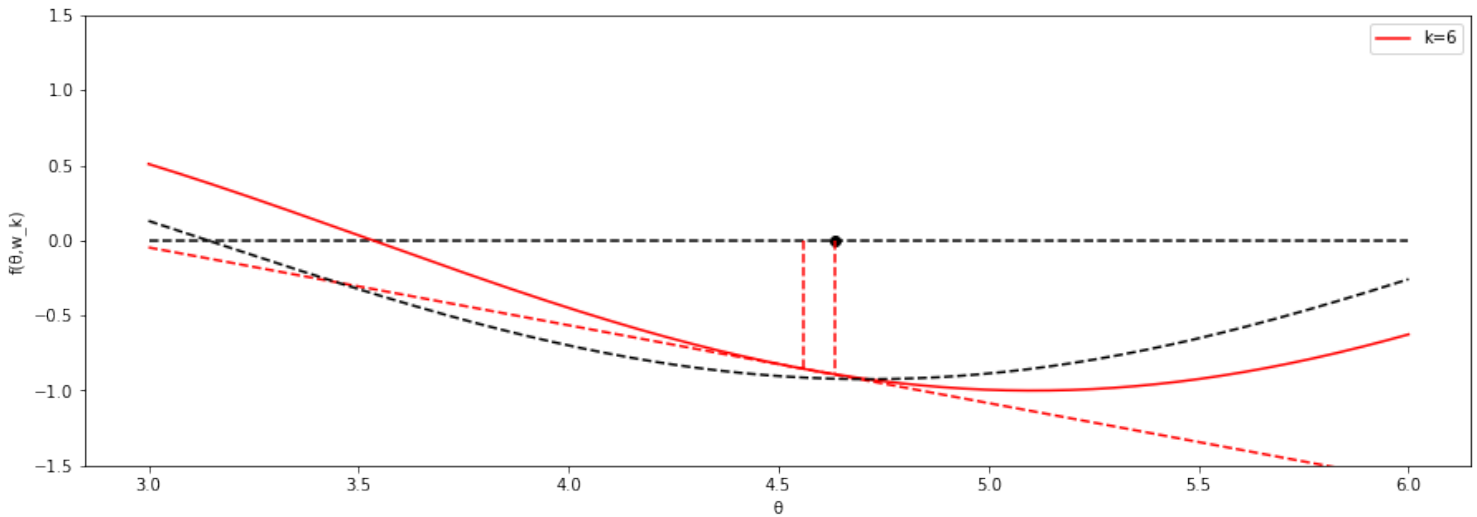


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

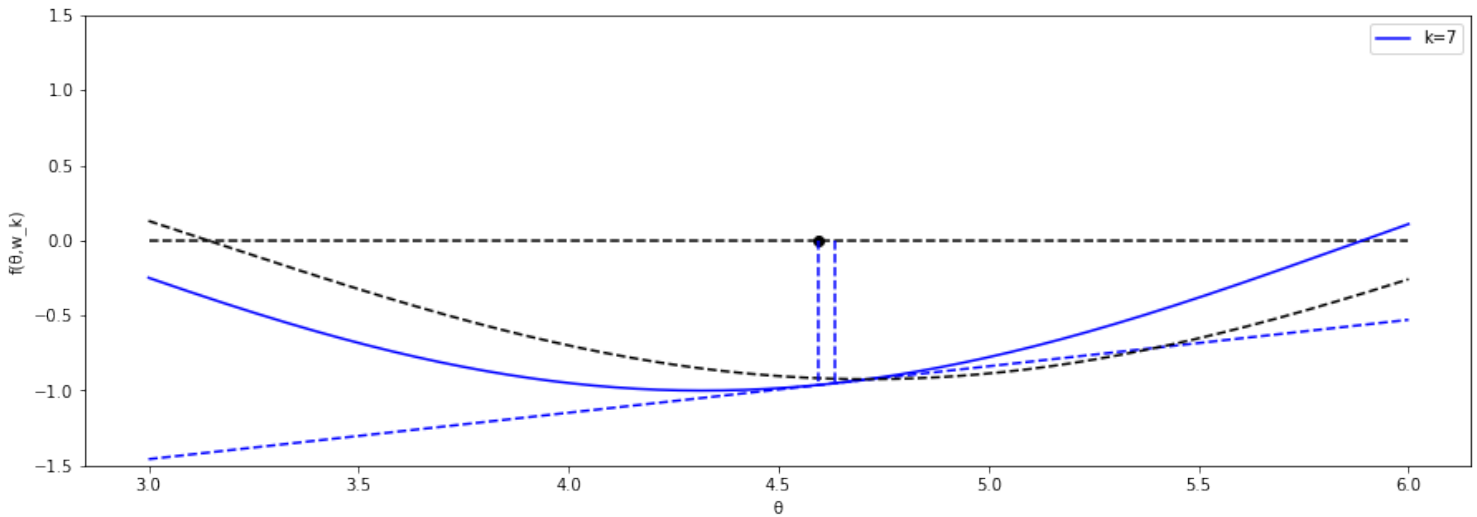


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

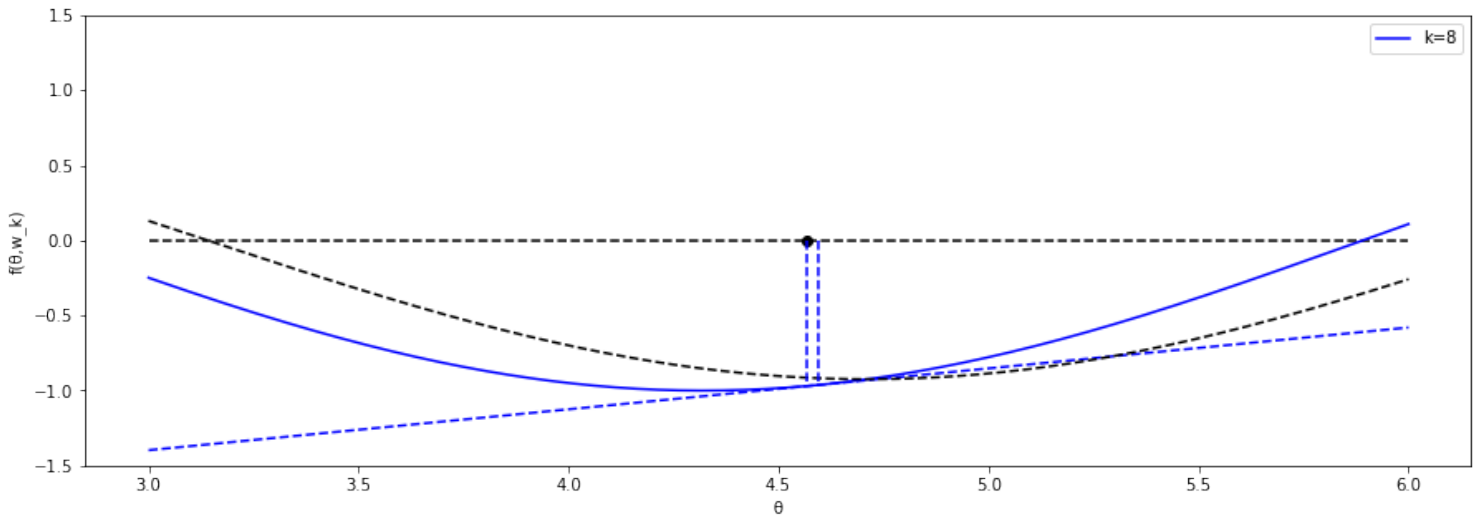


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

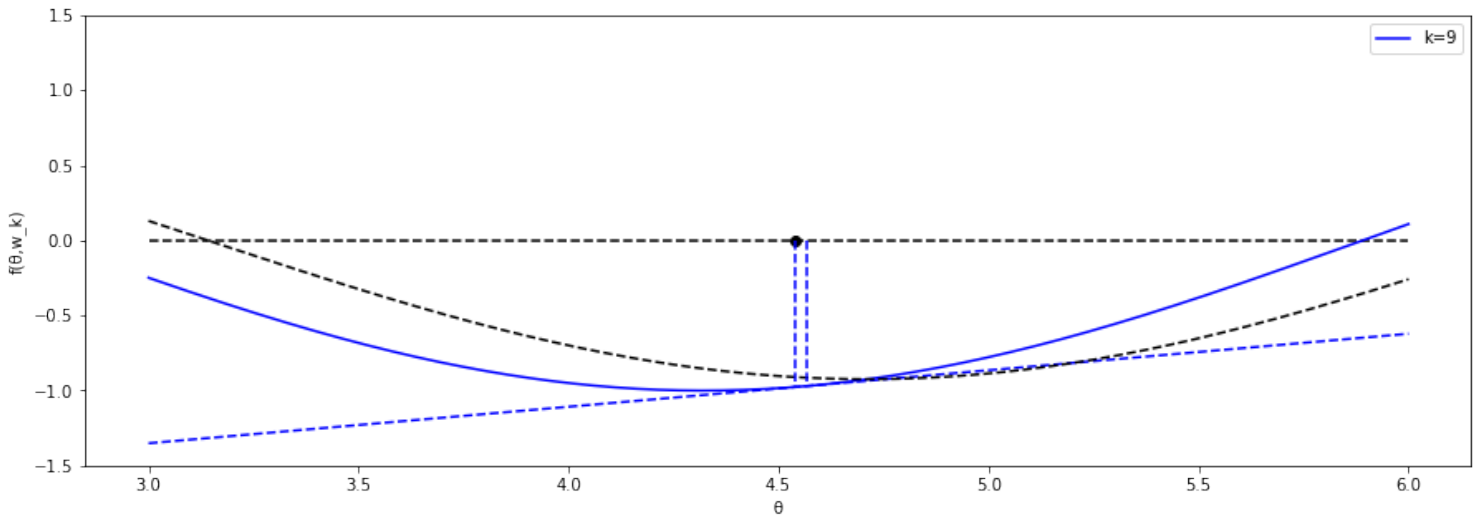


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

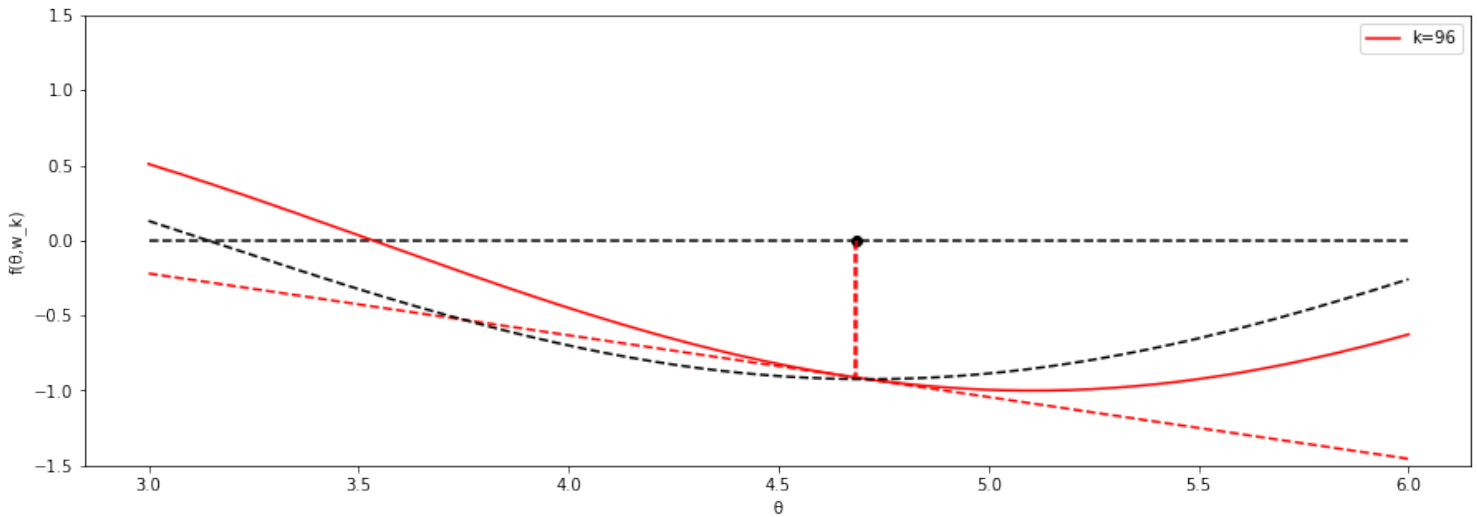


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k \frac{\pi}{8})}_{\varphi_k}$

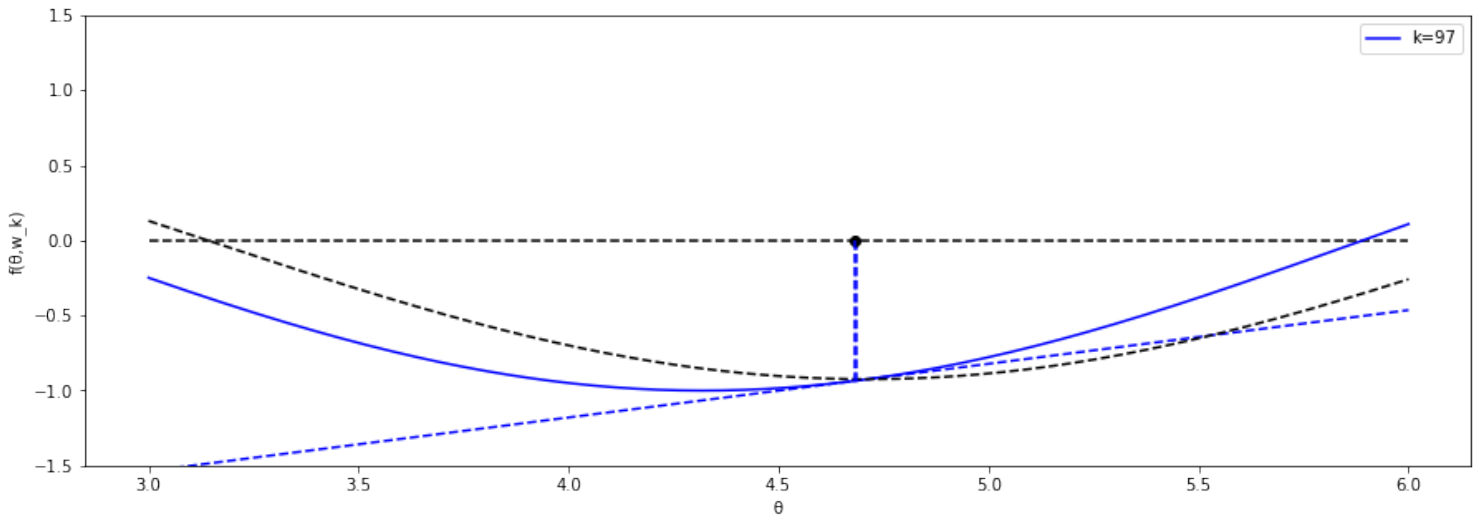


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

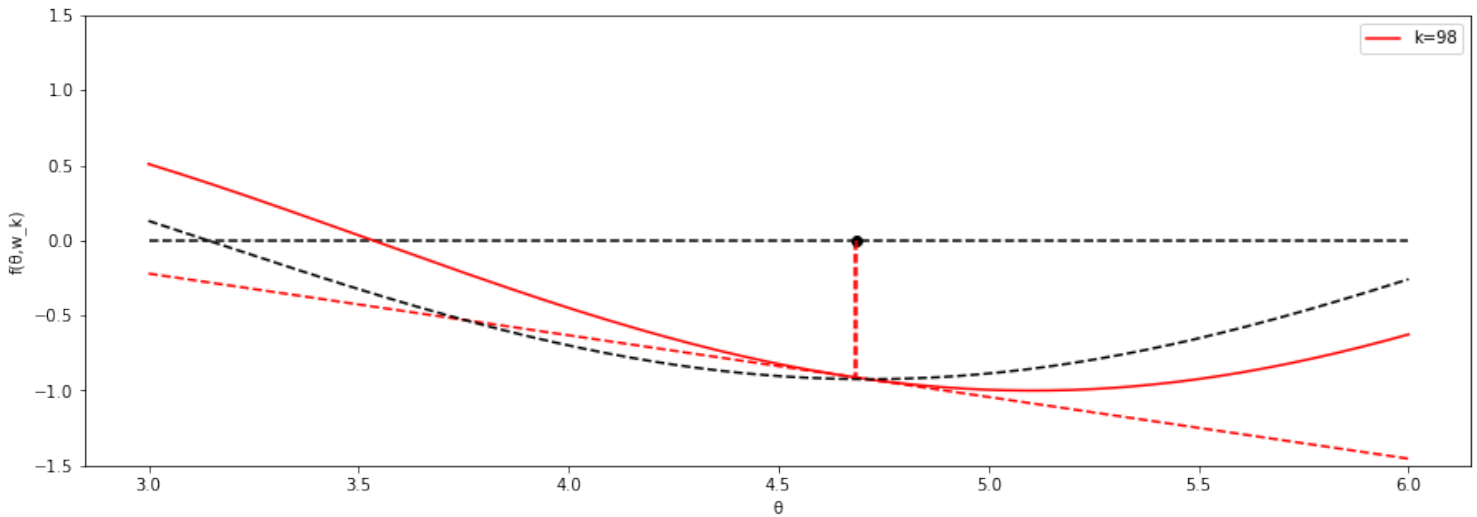


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k\frac{\pi}{8})}_{\varphi_k}$

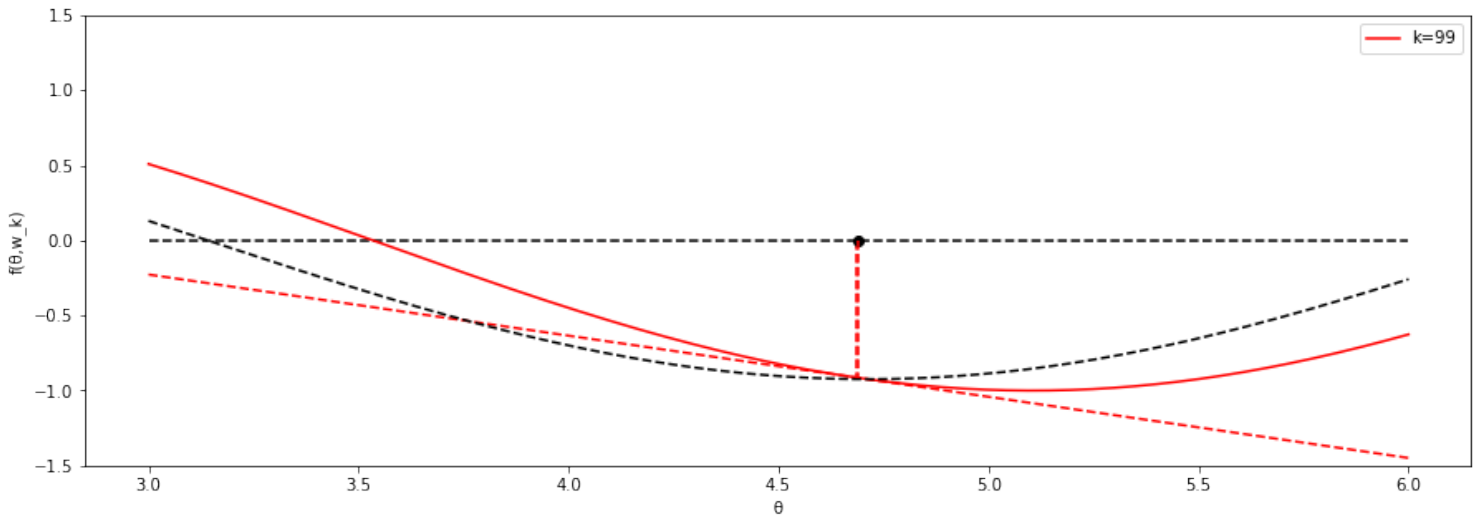


Caso no lineal

- $\min_\theta E_\xi[\sin(\theta + \xi\frac{\pi}{8})]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$

$$E_\xi[\sin(\theta+\xi\frac{\pi}{8})] = \frac{1}{2}\sin(\theta+\frac{\pi}{8})+\frac{1}{2}\sin(\theta-\frac{\pi}{8}) = \sin(\theta)\cos(\frac{\pi}{8}) \Rightarrow \theta^* = \frac{3\pi}{2}$$

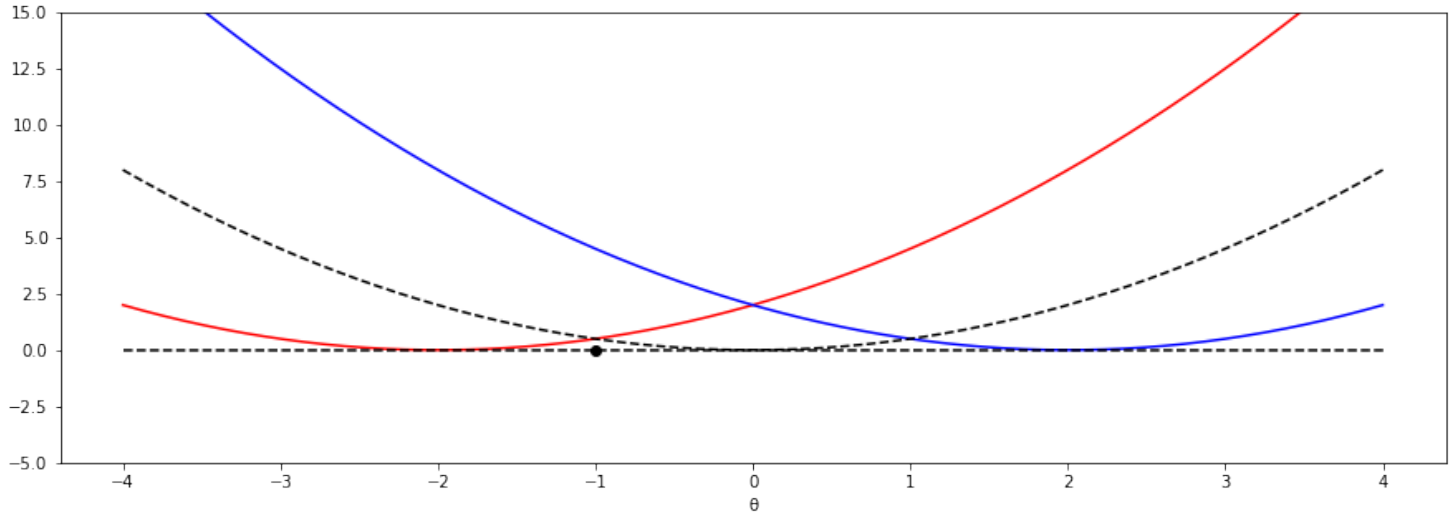
- SGD se mueve según $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \underbrace{\cos(\theta_k + \xi_k \frac{\pi}{8})}_{\varphi_k}$



- Nada especial del caso lineal

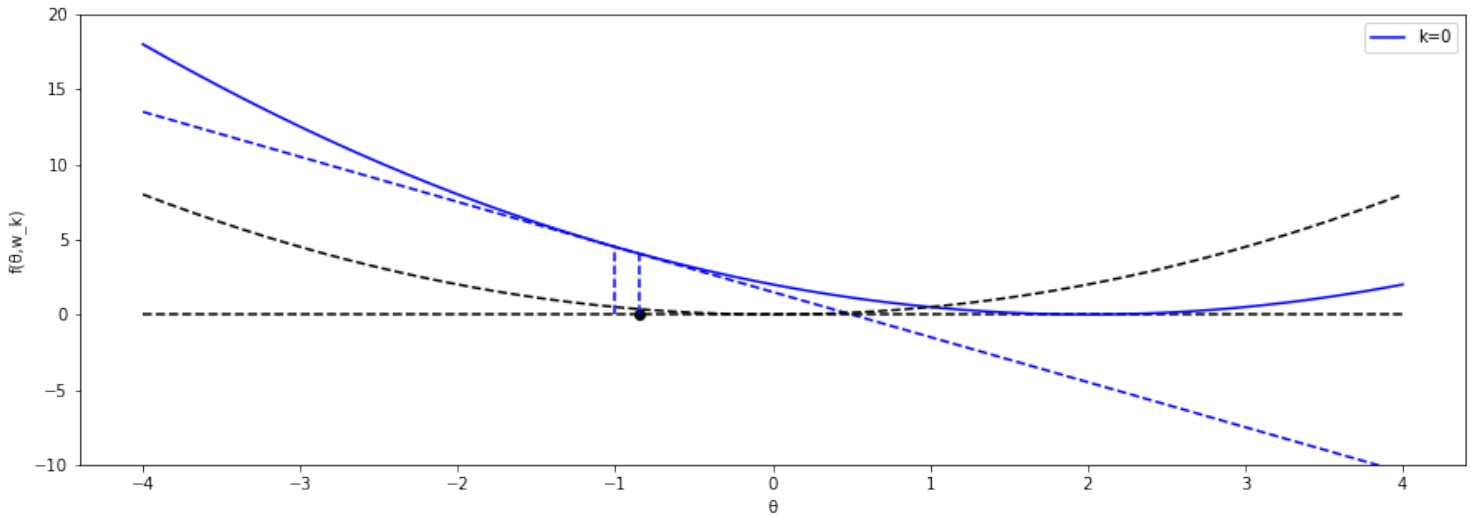
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



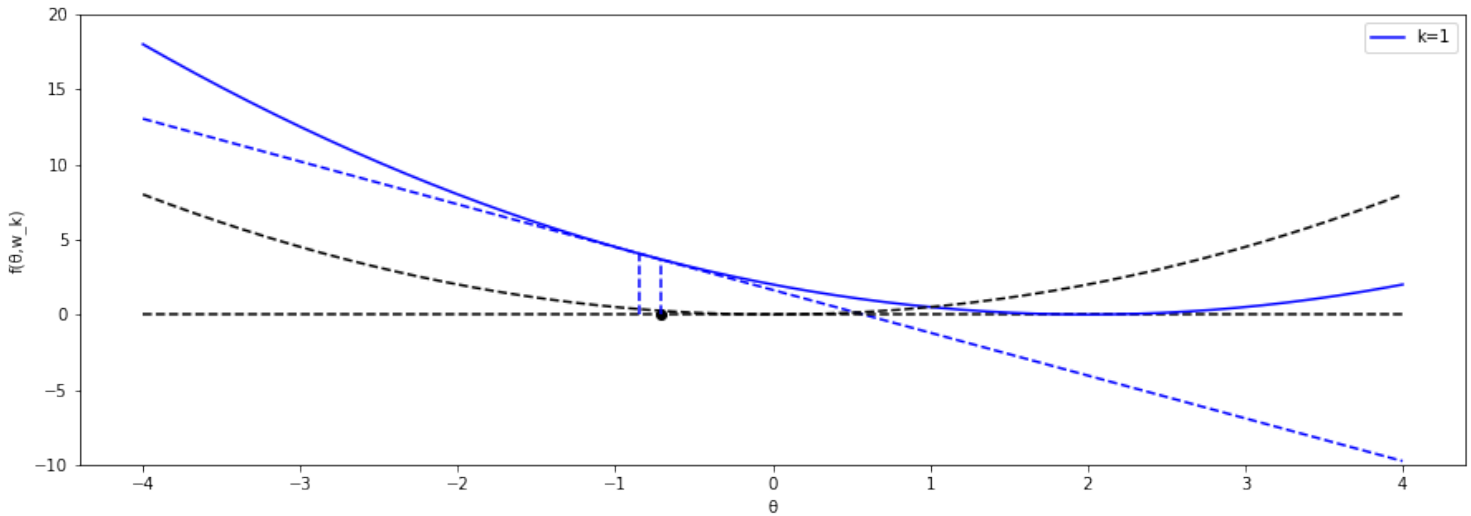
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



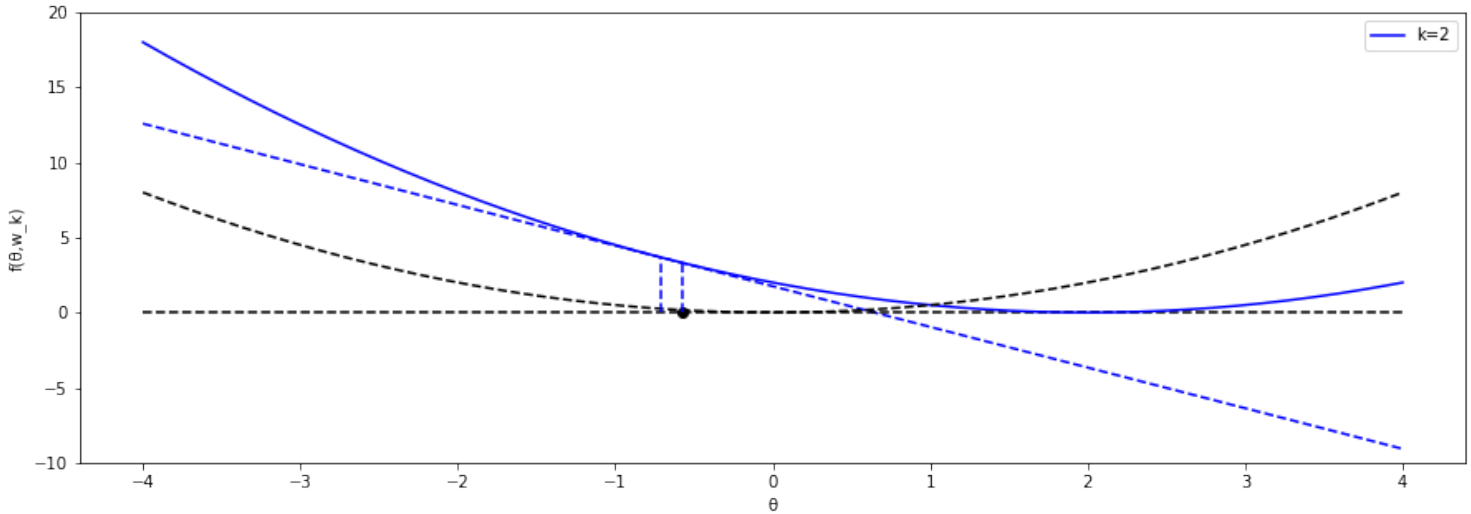
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



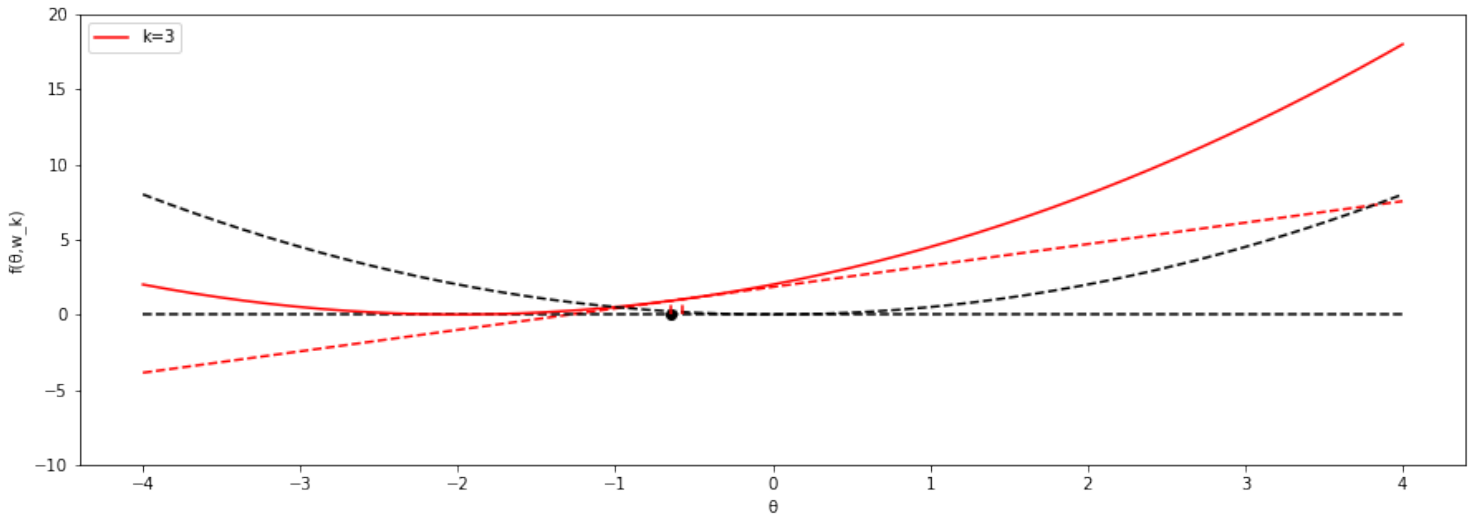
Efecto del paso

- Volviendo a mín$_\theta$ $\frac{1}{2}E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



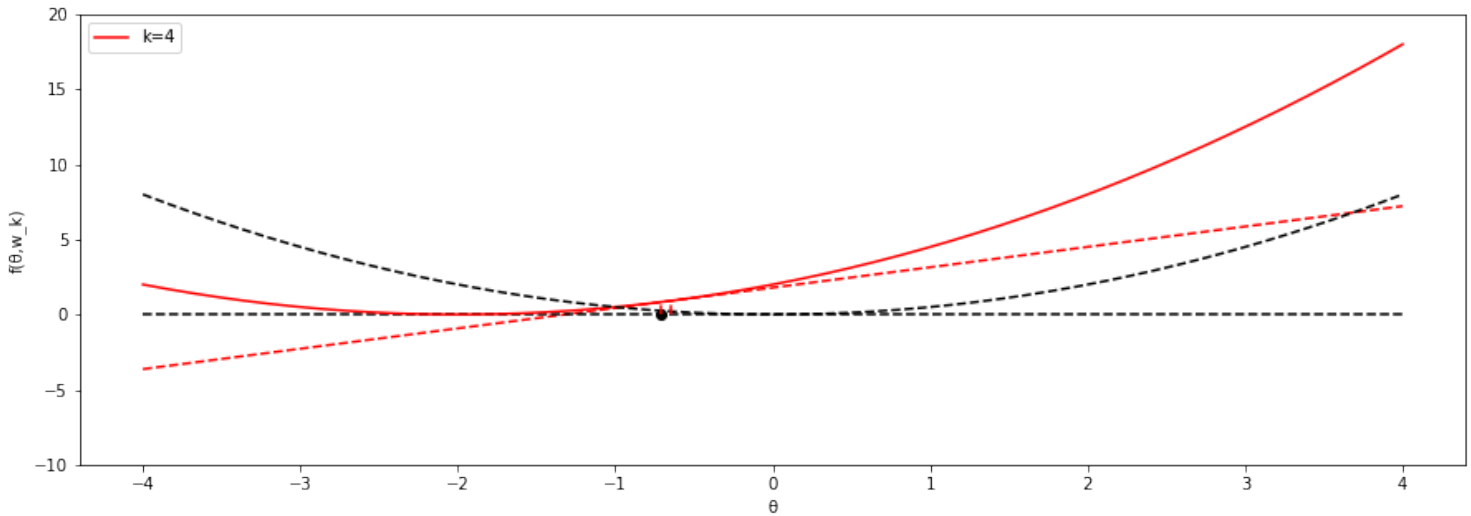
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



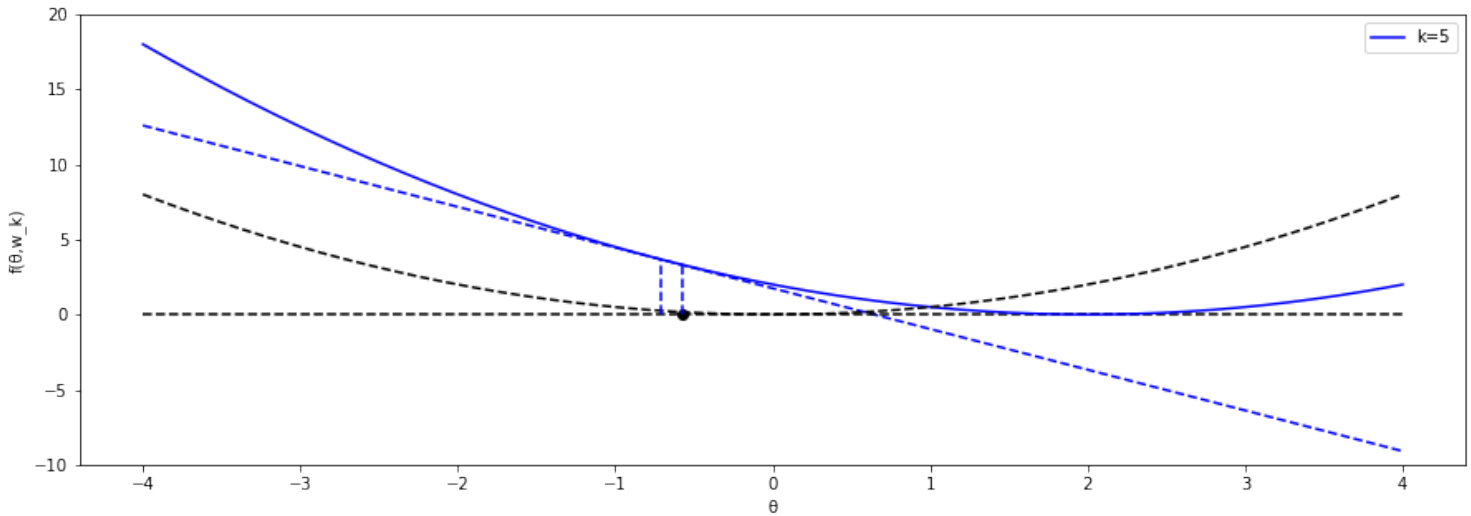
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



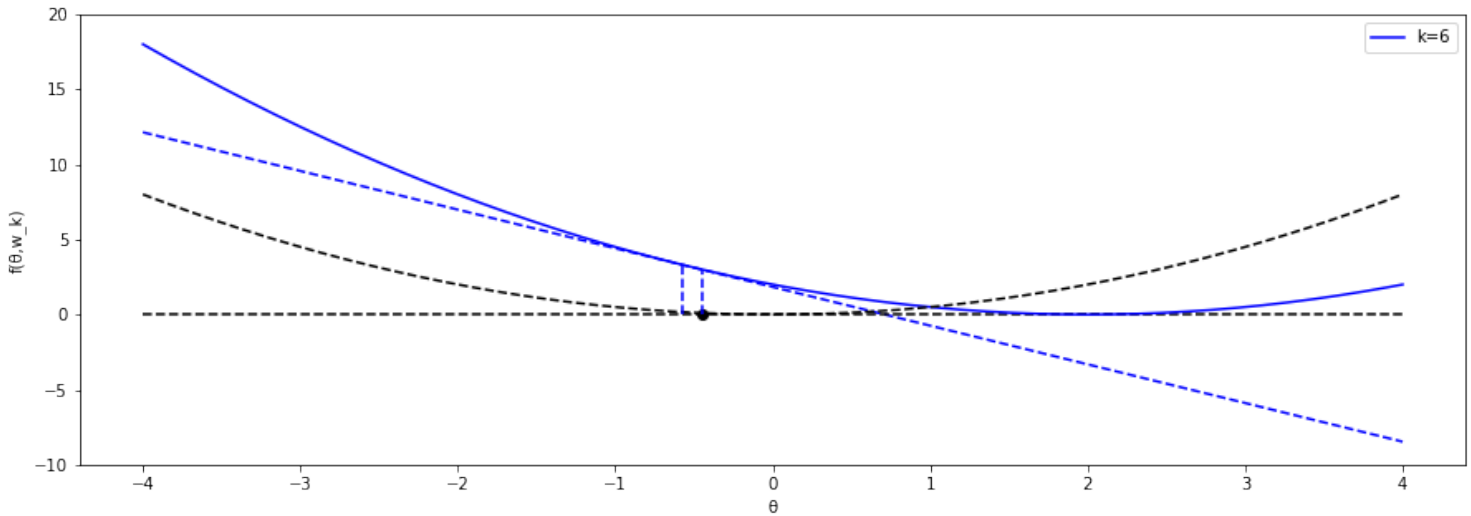
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



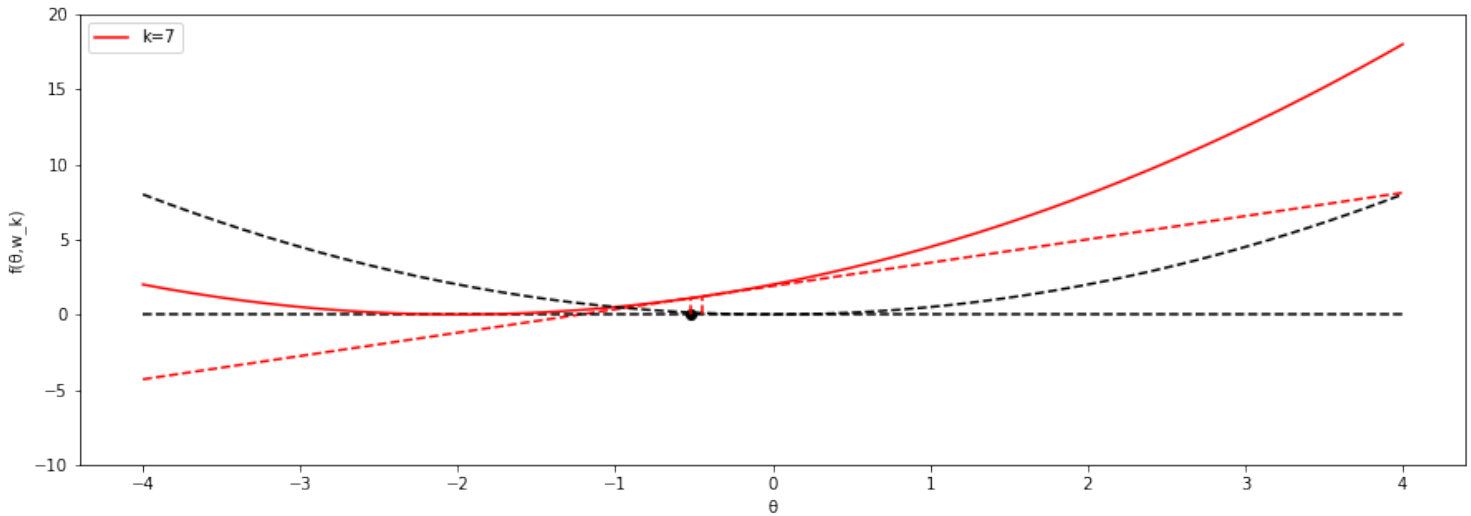
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



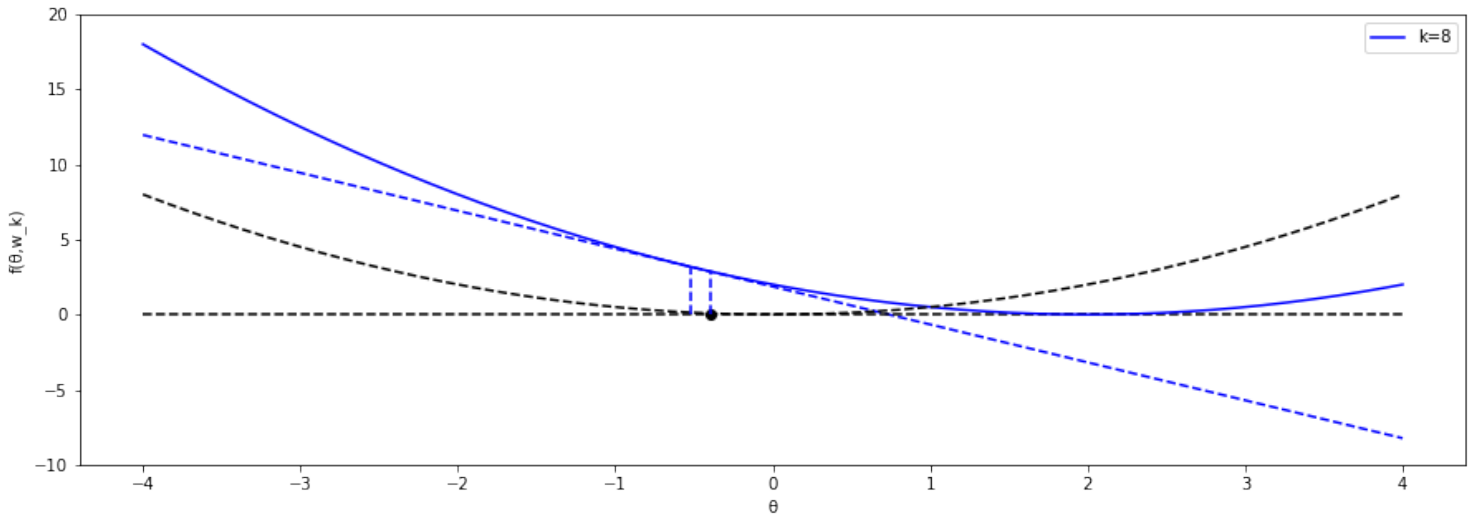
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



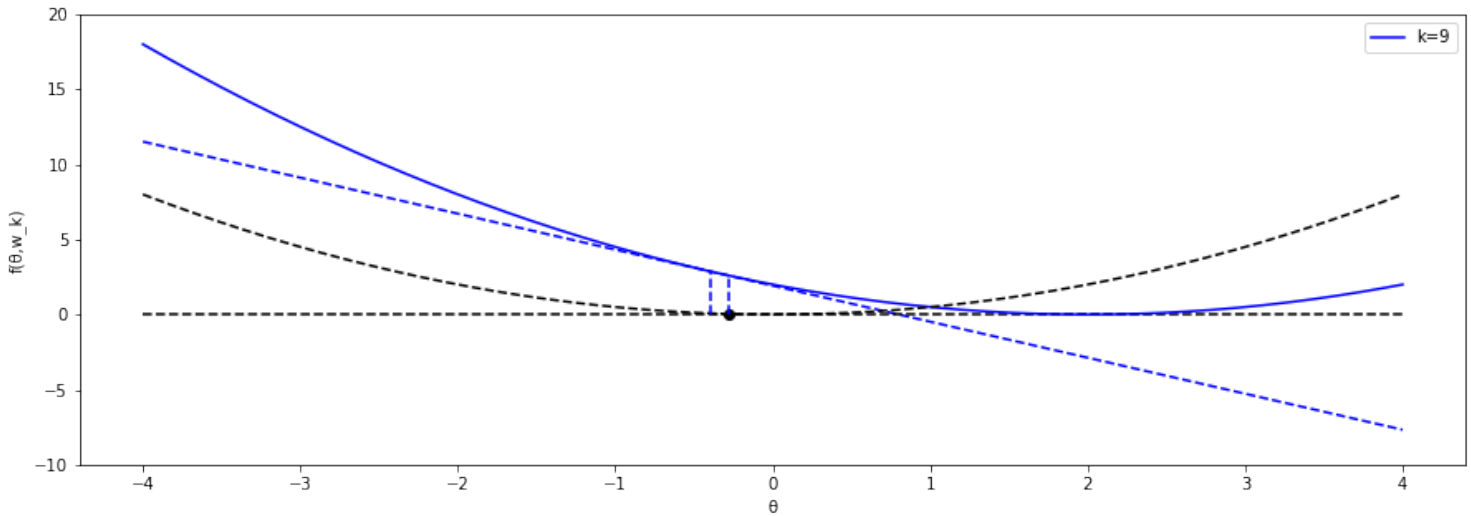
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



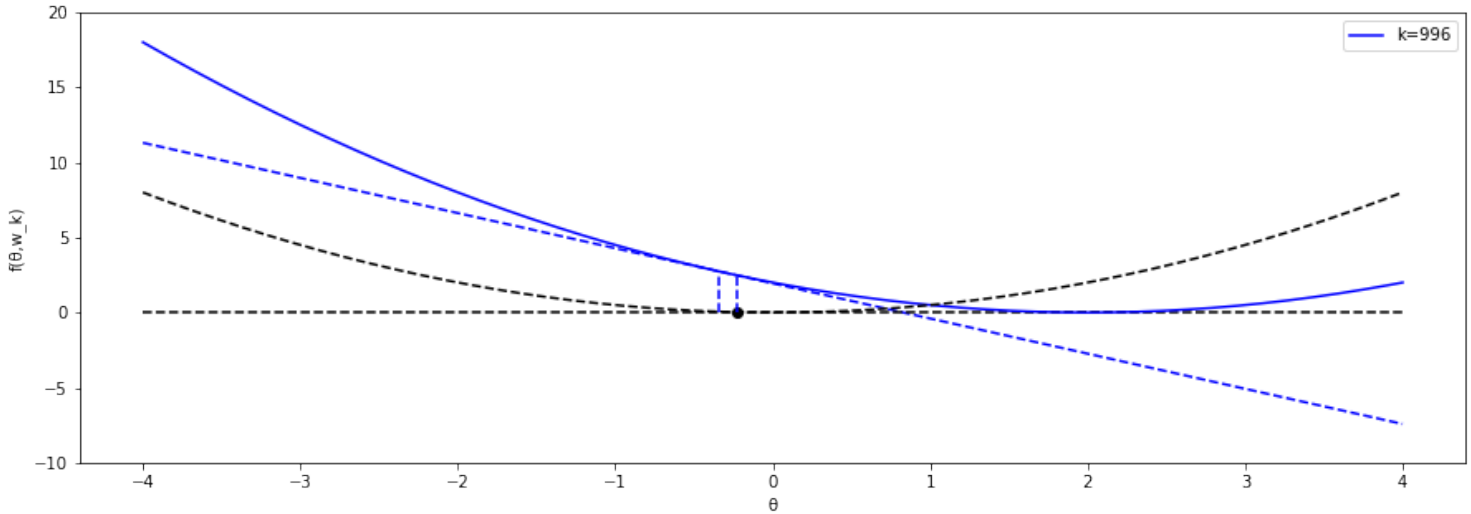
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



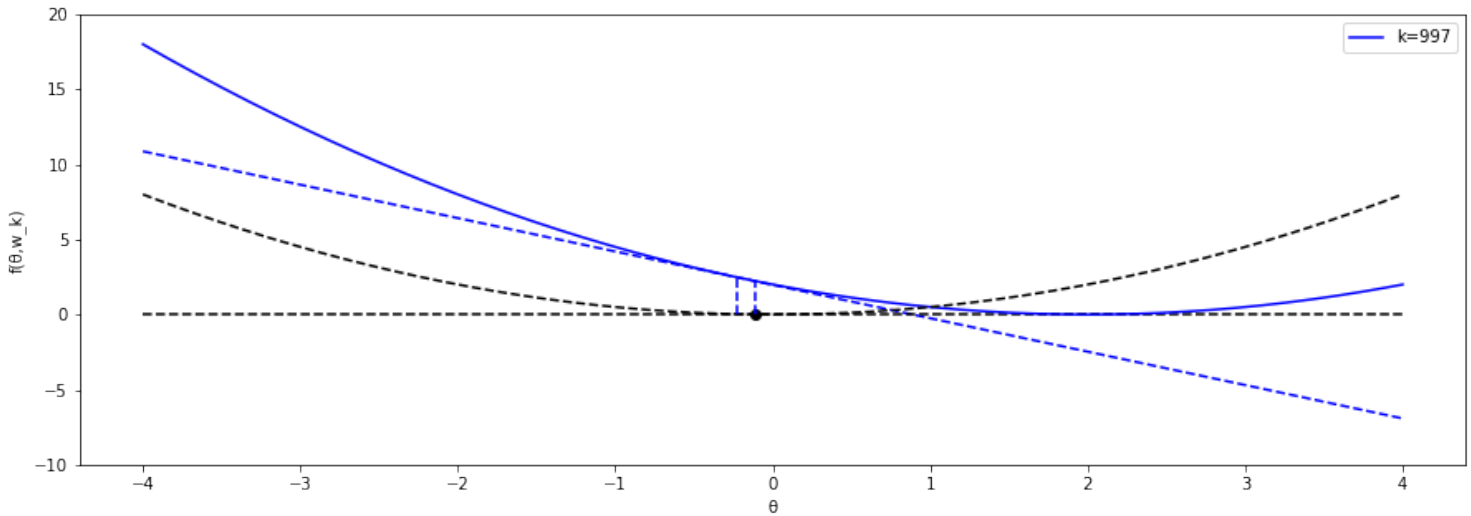
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



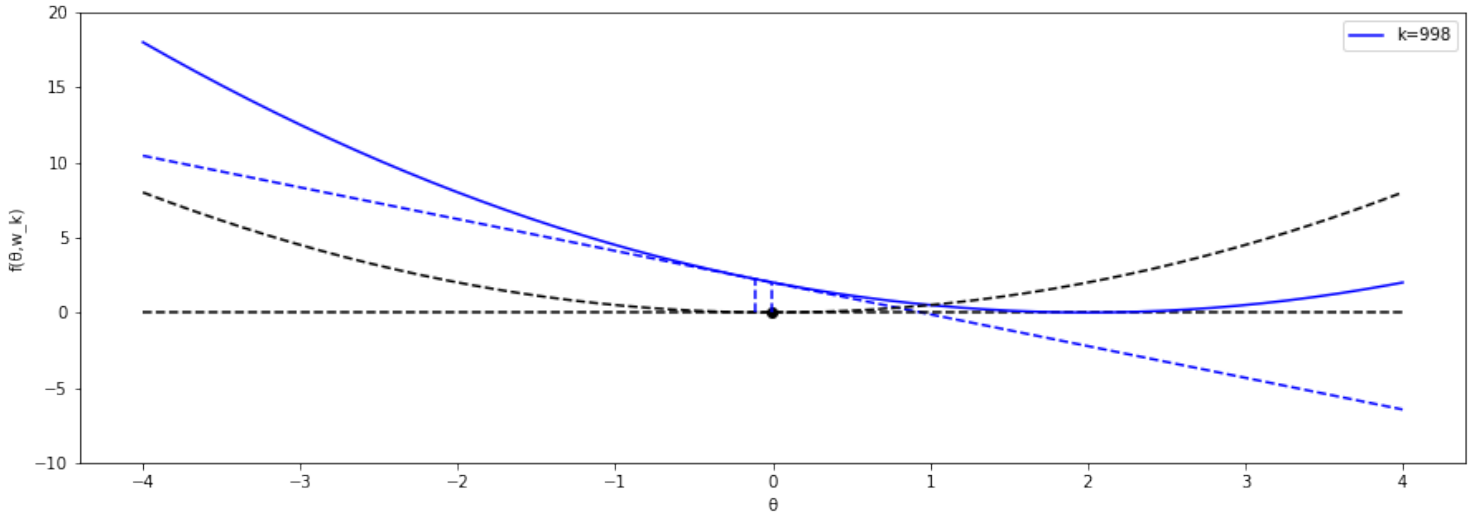
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



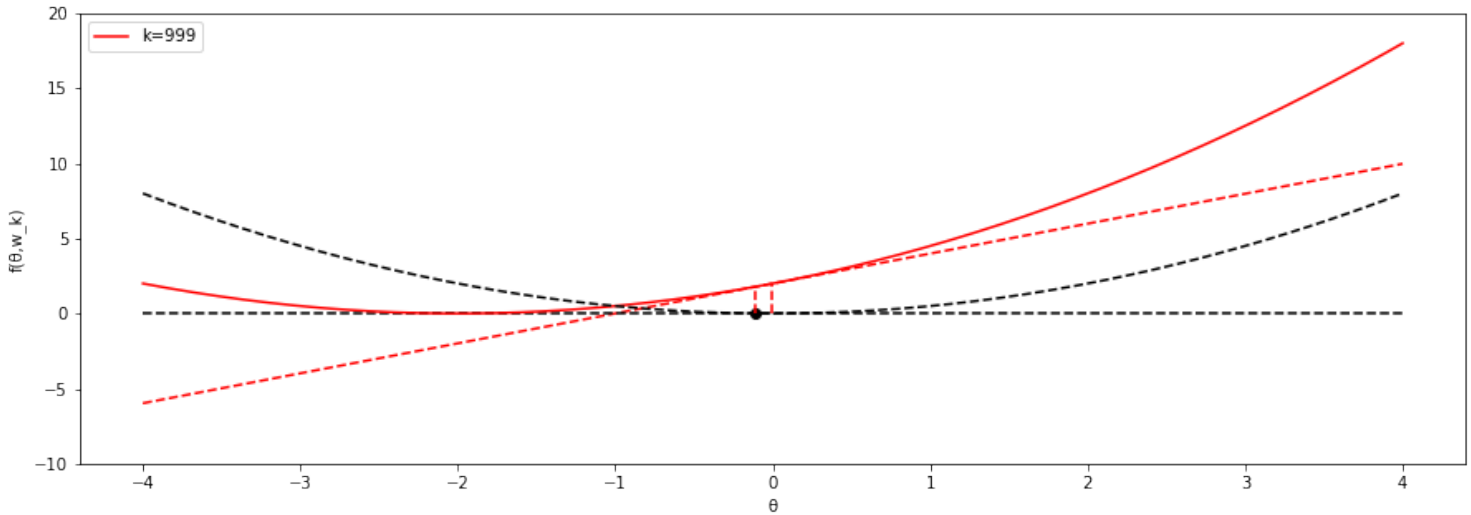
Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



Efecto del paso

- Volviendo a $\min_\theta \frac{1}{2} E_\xi[(\theta - w)^2]$ donde $\xi \in \{-1, 1\}$ con prob. $1/2$
- Consideremos un paso constante, e.g., $\alpha = 0{,}05$



- Con paso constante $\Rightarrow \theta_k$ orbita dentro de una bola
- Para convergencia se necesita $\alpha_k \to 0$ tal que $\sum_k \alpha_k = \infty$ y $\sum_k \alpha_k^2 < \infty$

Resultado formal (Robbins-Monro)

Dado $F(\theta, \xi) : \mathbb{R}^d \times \Xi \to \mathbb{R}$ y $f(\theta) = \mathbb{E}_\xi[F(\theta, \xi)]$ con $\theta^\star$ tal que $f(\theta^\star) = 0$. Si existen las constantes $c_1, c_2 > 0$, y el paso α_k tales que,

$$
\begin{aligned}
&(A1)\ \mathbb{E}_\xi[F(\theta, \xi)^2] \leq c_1 \\
&(A2)\ \textstyle\sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty \\
&(A3)\ c_2(\theta - \theta^\star)^2 \leq f(\theta)(\theta - \theta^\star)
\end{aligned}
$$

entonces la sucesión $\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \xi_k)$ con $\mathbb{E}[\theta_0^2] < \infty$ y $\xi_k \sim P_\xi$ i.i.d., converge a $\theta^\star$ en m.s.e.,

$$
\lim_{k \to \infty} \mathbb{E}\left[(\theta_k - \theta^\star)^2\right] = 0
$$

Prueba del Teorema de Robbins-Monro

- Definamos el error cuadrático medio (MSE)

$$e_k \triangleq \mathbb{E}_\xi\left[(\theta_k - \theta^\star)^2\right]$$

- Substituimos $\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \xi_k)$ en el MSE y expandimos

$$(\theta_k - \theta^\star - \alpha_k F(\theta_k, \xi_k))^2 = (\theta_k - \theta^\star)^2 - 2\alpha_k F_k(\theta_k)(\theta_k - \theta^\star) + \alpha_k^2 F_k^2(\theta_k)$$

$$(\theta_{k+1} - \theta^\star)^2 = (\theta_k - \theta^\star)^2 - 2\alpha_k F_k(\theta_k)(\theta_k - \theta^\star) + \alpha_k^2 F_k^2(\theta_k)$$

- Usamos $\mathbb{E}_\xi[\cdot]$ para obtener

$$e_{k+1} = e_k - 2\alpha_k \mathbb{E}_\xi\left[F_k(\theta_k)(\theta_k - \theta^\star)\right] + \alpha_k^2 \mathbb{E}_\xi\left[F_k^2(\theta_k)\right]$$

- Usando (A1) y (A3)

$$e_{k+1} \leq e_k - 2c_2\alpha_k e_k + c_1\alpha_k^2 = (1 - 2c_2\alpha_k)e_k + c_1\alpha_k^2$$

- Es un sistema contractivo $e_k \to 0$ bajo (A2)

Orbitas para paso constante

- Con paso constante $\alpha_k = \alpha$ tal que $|1 - 2c_2\alpha| < 1$

$$e_{k+1} \leq (1 - 2c_2\alpha)e_k + c_1\alpha^2$$

- Desenrollamos la sucesión

$$\begin{aligned} e_{k+1} &\leq (1 - 2c_2\alpha)^2 e_{k-1} + (1 - 2c_2\alpha)^2 c_1\alpha^2 + c_1\alpha^2 \\ &\leq (1 - 2c_2\alpha)^{k+1} e_0 + c_1\alpha^2 \sum_{l=0}^{k} (1 - 2c_2\alpha)^l \end{aligned}$$

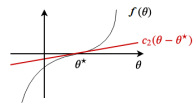
- Con $k \to \infty$ tenemos

$$\liminf_{k\to\infty} e_{k+1} \leq c_1\alpha^2 \frac{1}{1 - (1-2c_2\alpha)} = \frac{c_1}{2c_2}\alpha$$

- La bola de radio $\frac{c_1}{2c_2}\alpha$ es visitada infinitas veces

Comentarios de Robbins-Monro

- Hipótesis (A1)-(A2) son técnicas y habituales
- (A3) implica que el mínimo es único y asegura la contracción



- (A3) corresponde a un cruce franco por el cero de $f(\theta)$
- (A3) corresponde a convexidad fuerte cuando $f(\theta)$ es convexa
- El algoritmo $\theta_{k+1} = \theta_k - \alpha_k F(\theta_k, \xi_k)$ aprende de los datos $F_k = F(\theta_k, \xi_k)$
- Convergencia c.p.1, incluso debilitando (A3) y la hipótesis i.i.d.
- Este análisis requiere de martingalas

Convergencia con probabilidad 1

Dado $F(\theta, \xi) : \mathbb{R}^d \times \Xi \to \mathbb{R}$ y $f(\theta) = \mathbb{E}_\xi[F(\theta, \xi)] = \frac{dg}{d\theta}$ con $\theta^\star$ tal que $f(\theta^\star) = 0$, si existen $c_1, c_2 > 0$, y el paso α_k tales que,

$$(A1)\ \mathbb{E}_\xi[F(\theta, \xi)^2] \leq c_1$$
$$(A2)\ \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \alpha_k^2 < \infty$$
$$(A3)\ \ g(\theta) \text{ es convexa}$$

entonces la sucesión

$$\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k F(\hat{\theta}_k, \xi_k)$$

converge a $\theta^\star$ con prob. 1.

$$\lim_{k \to \infty} E\left[\left(\hat{\theta}_k - \theta^\star\right)^2\right] = 0$$

Esquema de la prueba

- Idea: usar la convexidad para probar que $S_k = \|\hat{\theta}_k - \theta^*\|^2$ es una super-martingala, i.e.,

$$E_w[\|\hat{\theta}_{k+1} - \theta^*\|^2 | \theta_k] \leq \|\hat{\theta}_k - \theta^*\|^2$$

- SMs son la contraparte aleatoria de sucesiones decrecientes $\Rightarrow$ convergen (Doob)
- Se completa la prueba mostrando que el límite es cero
- Una corrección técnica es necesaria $S_k := \|\hat{\theta}_k - \theta^*\|^2 + c_1 \sum_{i=k}^{\infty} \delta_i^2$