2021-2022

Aprendizaje Automático

3. Técnicas de optimización

Francisco Casacuberta Nolla
(fcn@dsic.upv.es)

Enrique Vidal Ruiz
(evidal@dsic.upv.es)

Departament de Sistemas Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Septiembre, 2021

Index

Index

Clasificación, regresión y optimización

- Los modelos están parametrizados por un vector de parámetros $\mathbf{\Theta}$; es decir, $\mathcal{F} = \{f_{\mathbf{\Theta}} : \mathcal{X} \rightarrow \mathcal{Y}, \mathbf{\Theta} \in \mathbb{R}^D\}$
- Clasificación: $f_{\mathbf{\Theta}} : \mathcal{X} \rightarrow \{1, \ldots, C\}$. En muchos problemas $\mathcal{X} \equiv \mathbb{R}^d$
- Regresión: $f_{\mathbf{\Theta}} : \mathcal{X} \rightarrow \mathcal{Y}$. Típicamente $\mathcal{X} \equiv \mathbb{R}^d$ y $\mathcal{Y} \equiv \mathbb{R}$.
- Dos etapas:
  - Aprendizaje: Dado $S \subset \mathcal{X} \times \mathcal{Y}$, estimar $\hat{\mathbf{\Theta}} \in \mathbb{R}^D$

    Técnicas de optimización para aprendizaje:
    * Optimización analítica.
    * Optimización con restriciones: Multiplicadores de Lagrange.
    * Descenso (ascenso) por gradiente.
    * Optimización probabilística: Algoritmo EM.
  - Búsqueda o inferencia: Dados $\mathbf{\Theta}$ y $x \in \mathcal{X}$, estimar $\hat{y} = f_{\mathbf{\Theta}}(x)$.

    Técnicas de optimización para búsqueda:
    * Exhaustiva: por ejemplo, clasificación, si $C <<$.
    * Programación dinámica: algoritmo de Viterbi con modelos ocultos de Markov.
    * Inteligente: ramificación y poda, A*, etc.

Optimización y aprendizaje automático

- Dados:
  - N muestras de aprendizaje:
    $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$, $(x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$, $1 \leq n \leq N$,
  - un clasificador o regresor, $f_{\mathbf{\Theta}} : \mathcal{X} \rightarrow \mathcal{Y}$, parametrizado por $\mathbf{\Theta} \in \mathbb{R}^D$,
  - un criterio aprendizaje definido por una función objetivo, $q_S : \mathbb{R}^D \rightarrow \mathbb{R}$
- estimar $\hat{\mathbf{\Theta}}$ mediante optimización de q_S; es decir:

$$\hat{\mathbf{\Theta}} \equiv \mathbf{\Theta}^\star = \arg\min_{\mathbf{\Theta}} q_S(\mathbf{\Theta})$$

o bien:

$$\hat{\mathbf{\Theta}} \equiv \mathbf{\Theta}^\star = \arg\max_{\mathbf{\Theta}} q_S(\mathbf{\Theta})$$

Técnicas generales de AA basadas en optimización

- $f_{\mathbf{\Theta}}$ es una función cualquiera que depende de un vector de parámetros $\mathbf{\Theta}$: $q_S(\mathbf{\Theta})$ se basa en *funciones de error* y típicamente su optimización utiliza técnicas de *descenso/ascenso por gradiente* (caso particular de "hill-climbing").

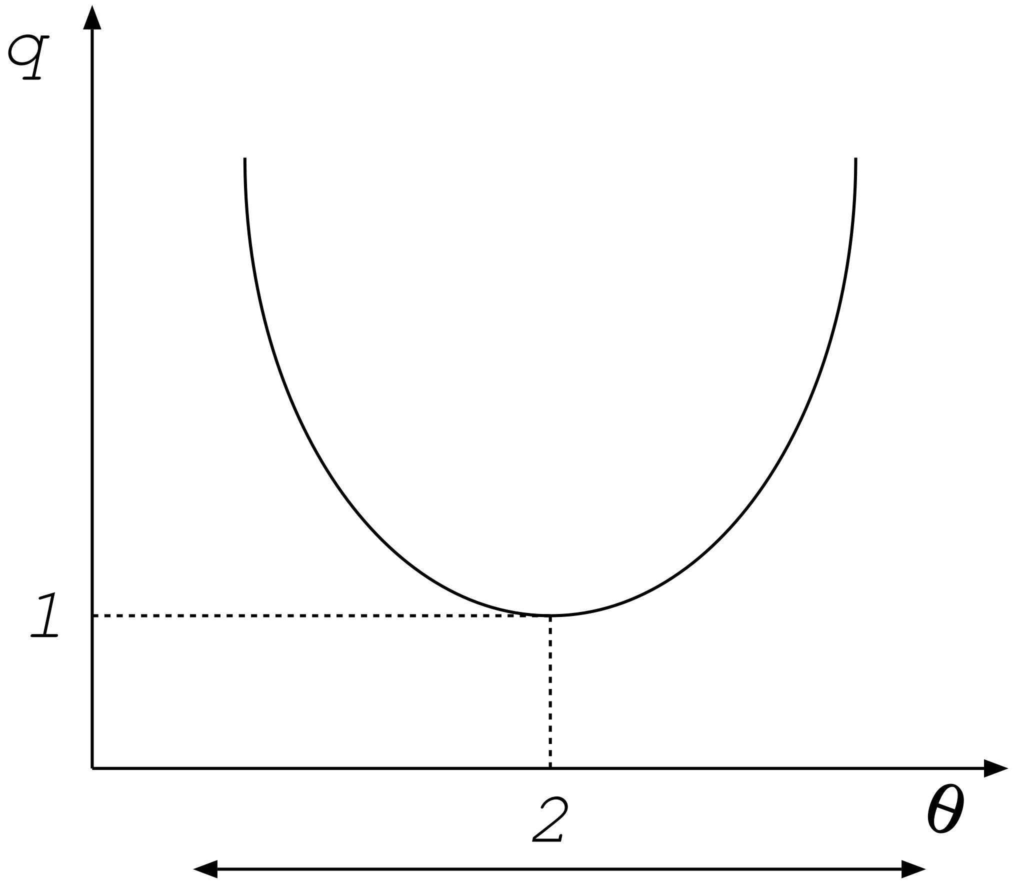
- $f_{\mathbf{\Theta}}$ es una función cualquiera dependiente de $\mathbf{\Theta}$, pero hay ciertas restricciones en los valores posibles de los parámetros $\mathbf{\Theta}$: *optimización con restricciones* de $q_S(\mathbf{\Theta})$ mediante la técnica de los *multiplicadores de Lagrange*.

- $f_{\mathbf{\Theta}}$ se basa en distribuciones (o densidades) de probabilidad: estimación de *máxima verosimilitud*. Frecuentemente hay restricciones en los parámetros a estimar y se requiere el uso de *multiplicadores de Lagrange*.

- $f_{\mathbf{\Theta}}$ se basa en distribuciones (o densidades) de probabilidad, pero hay variables aleatorias *"latentes"* u *"ocultas"*: La estimación de *máxima verosimilitud* generalmente requiere una técnica de optimización llamada *esperanza-maximización* (EM).

Index

Optimización analítica: ejemplo



- Dado $q(\theta) = 1 + (\theta - 2)^2$
- Calcular $\theta^\star = \arg\min_{\theta \in \mathbb{R}} q(\theta)$
- Procedimiento: $\frac{d\,q(\theta)}{d\,\theta} = 2\;(\theta - 2) = 0$
- Solución: $\theta^\star = 2$

Optimización analítica: gradiente

- Dada una función *convexa* $q : \mathbb{R}^D \to \mathbb{R}$, calcular $\displaystyle\operatorname*{arg\,min}_{\boldsymbol{\Theta}\in\mathbb{R}^D} q(\boldsymbol{\Theta})$
- Procedimiento:
  1. Calcular el gradiente de q: $\boldsymbol{\nabla} q(\boldsymbol{\Theta}) \stackrel{\text{def}}{=} \left(\frac{\partial q(\boldsymbol{\Theta})}{\partial \Theta_1}, \ldots, \frac{\partial q(\boldsymbol{\Theta})}{\partial \Theta_D}\right)^t$
  2. Resolver $\boldsymbol{\nabla} q(\boldsymbol{\Theta}) = \mathbf{0}$;

     es decir, resolver el sistema de ecuaciones $\displaystyle\frac{\partial q(\boldsymbol{\Theta})}{\partial \Theta_i} = 0, \; 1 \leq i \leq D$.

     Sean $\Theta_1^\star, \ldots, \Theta_D^\star$ las soluciones obtenidas.
- Solución: $\boldsymbol{\Theta}^\star = (\Theta_1^\star, \ldots, \Theta_D^\star)^t$
- Si q es convexa, $\boldsymbol{\nabla} q(\boldsymbol{\Theta}^\star) = \mathbf{0}$ es una condición *necesaria y suficiente* para que $\boldsymbol{\Theta}^\star$ sea (la única) solución.

Ejercicios:

a) Encontrar el vector $\boldsymbol{\theta}^\star \in \mathbb{R}^2$ que minimiza la función $q(\boldsymbol{\theta}) = (\theta_1 - 1)^2 + (\theta_2 - 2)^2$

b) Encontrar el vector $\boldsymbol{\theta}^\star \in \mathbb{R}^2$ que minimiza la función $q(\boldsymbol{\theta}) = (\theta_1 - 1)^2 + (\theta_2 - 2)^2 + \theta_1\theta_2$

c) ¿Qué ocurre si q *no* es convexa? Ejemplos: $\frac{1}{10}x^3 + 20\,x^2$, $\; x^4 - 10\,x^2$

Optimización analítica: otro ejemplo simple

- Estimar los parámetros[†] $\boldsymbol{\Theta} \equiv (\mu, \sigma)$ de una gaussiana univariada (en $\mathbb{R}^1$):

$$p(x; \boldsymbol{\Theta}) \stackrel{\text{def}}{=} p(x;\ \mu, \sigma) \ = \ \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Logaritmo de la verosimilitud de una muestra $S = \{x_1, \ldots, x_N\}$:

$$\begin{aligned} q_S(\boldsymbol{\Theta}) \ \equiv \ L_S(\mu, \sigma) \ &= \ \log \prod_{n=1}^{N} p(x_n; \mu, \sigma) \ = \ \sum_{n=1}^{N} \log p(x_n; \mu, \sigma) \\ &= \ N \ \log \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \end{aligned}$$

- Para una estimación de máxima verosimilitud basta hacer $\boldsymbol{\nabla} L_S(\boldsymbol{\Theta}) = \mathbf{0}$. En nuestro caso unidimensional (ejercicio):

$$\frac{\partial L_S(\mu, \sigma)}{\partial \mu} = 0 \ \Rightarrow \ \hat{\mu} \ = \ \frac{1}{N} \sum_{n=1}^{N} x_n \, ; \qquad \frac{\partial L_S(\mu, \sigma)}{\partial \sigma} = 0 \ \Rightarrow \ \hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

[†] media y desviación típica

Optimización analítica: otro ejemplo

- Estimar los parámetros de una gaussiana multivariada (en $\mathbb{R}^D$), con Σ dada:

$$p(\boldsymbol{x}; \boldsymbol{\Theta}) = (2\pi)^{-D/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right)$$

  donde $\boldsymbol{\Theta} \equiv (\boldsymbol{\mu}, \Sigma)$. Si Σ está prefijada, entonces $\boldsymbol{\Theta} \equiv \boldsymbol{\mu} \in \mathbb{R}^D$

- Logaritmo de la verosimilitud de una muestra $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$:

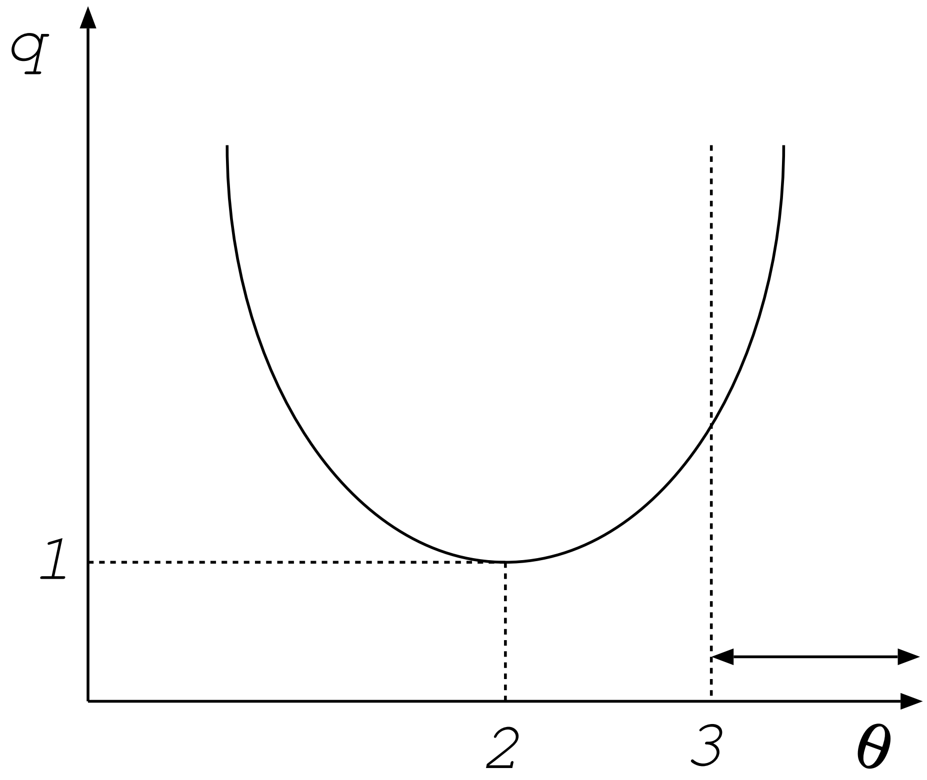
$$\begin{aligned} q_S(\boldsymbol{\Theta}) \equiv L_S(\boldsymbol{\Theta}) &= \log \prod_{n=1}^{N} p(\boldsymbol{x}_n; \boldsymbol{\Theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{x}_n; \boldsymbol{\Theta}) \\ &= N \log \left((2\pi)^{-D/2} |\Sigma|^{-1/2} \right) - \frac{1}{2} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu})^t \Sigma^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}) \end{aligned}$$

- Para una estimación de máxima verosimilitud basta hacer $\boldsymbol{\nabla} L_S(\boldsymbol{\Theta}) = \mathbf{0}$. Si Σ está prefijada, se obtiene (*ejercicio*):

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n$$

Index

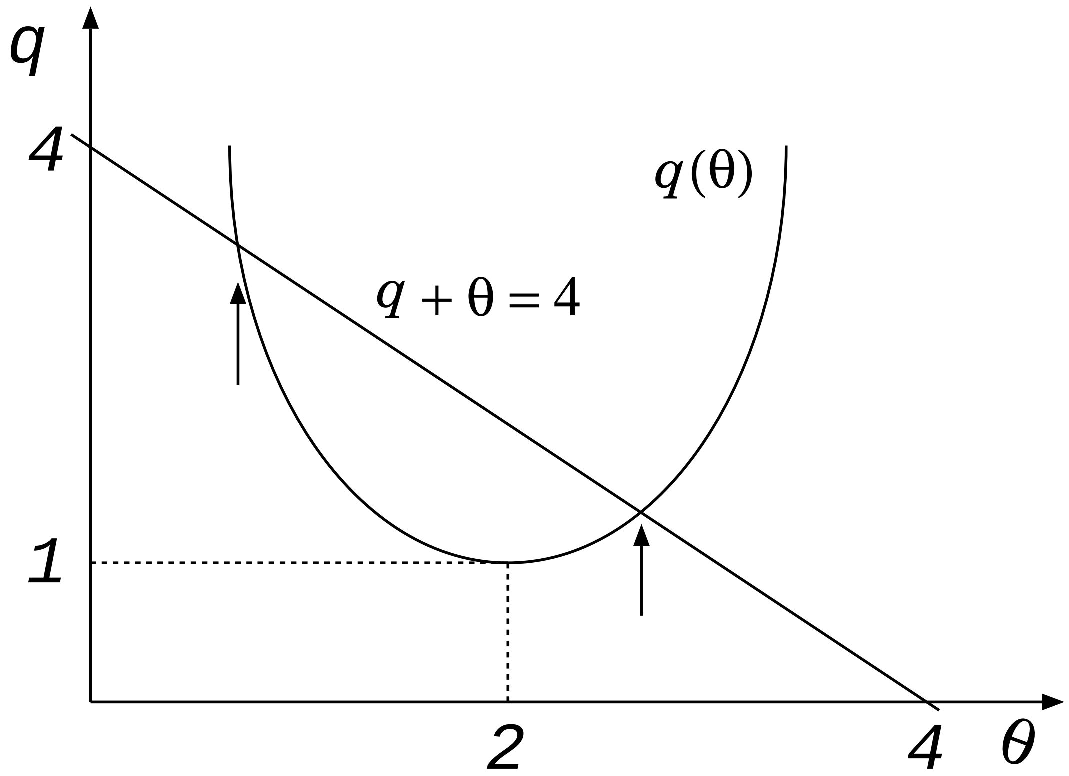
Optimización con restricciones: ejemplo simple 1



- Dado: $q(\theta) = 1 + (\theta - 2)^2$,
- Calcular: $\theta^\star = \underset{\theta:\,\theta \geq 3}{\arg\min}\, q(\theta)$

  (restricción de desigualdad: $\theta - 3 \geq 0$)
- Solución: ??

Optimización con restricciones: ejemplo simple 2



- Dado: $q(\theta) = 1 + (\theta - 2)^2$,

- Calcular: $\theta^\star = \underset{\theta:\, q+\theta=4}{\arg\min}\, q(\theta)$

  (restricción de igualdad: $q + \theta - 4 = 0$)

- Solución: ??

Optimización con restricciones: multiplicadores de Lagrange

Consideremos un problema de optimización definido por:

$$\begin{array}{lll} \text{minimizar} & q(\boldsymbol{\Theta}) & \boldsymbol{\Theta} \in \mathbb{R}^D \\ \text{sujeto a} & v_i(\boldsymbol{\Theta}) \geq 0 & 1 \leq i \leq k \\ & u_i(\boldsymbol{\Theta}) = 0 & 1 \leq i \leq m \end{array}$$

donde q es una función convexa y v_i, u_i son funciones que expresan restricciones. Equivalentemente, el problema consiste en calcular:

$$\boldsymbol{\Theta}^\star = \underset{\substack{\boldsymbol{\Theta} \in \mathbb{R}^D \\ v_i(\boldsymbol{\Theta}) \geq 0,\ 1 \leq i \leq k \\ u_i(\boldsymbol{\Theta}) = 0,\ 1 \leq i \leq m}}{\arg\min} q(\boldsymbol{\Theta})$$

Para resolver este problema, se define la *función Lagrangiana*:

$$\Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \stackrel{\text{def}}{=} q(\boldsymbol{\Theta}) - \sum_{i=1}^{k} \alpha_i v_i(\boldsymbol{\Theta}) + \sum_{i=1}^{m} \beta_i u_i(\boldsymbol{\Theta})$$

donde $\alpha_i \geq 0, 1 \leq i \leq k$ y $\beta_i, 1 \leq i \leq m,$ son los *multiplicadores de Lagrange*.

La técnica de los multiplicadores de Lagrange

1. *Definir multiplicadores de Lagrange y Lagrangiana:*

$$\Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \stackrel{\text{def}}{=} q(\boldsymbol{\Theta}) - \sum_{i=1}^{k} \alpha_i v_i(\boldsymbol{\Theta}) + \sum_{i=1}^{m} \beta_i u_i(\boldsymbol{\Theta})$$

2. *Obtener el minimizador $\boldsymbol{\Theta}^\star$ de la Lagrangiana $\Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$*, en función de $\boldsymbol{\alpha}, \boldsymbol{\beta}$ (resolviendo $\boldsymbol{\nabla}_{\boldsymbol{\Theta}} \Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$):

$$\boldsymbol{\Theta}^\star(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \arg\min_{\boldsymbol{\Theta}} \Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$$

3. *Obtener la función dual de Lagrange* (sustituir $\boldsymbol{\Theta}$ por $\boldsymbol{\Theta}^\star(\boldsymbol{\alpha}, \boldsymbol{\beta})$ en $\Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta})$):

$$\Lambda_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) \stackrel{\text{def}}{=} \Lambda(\boldsymbol{\Theta}^\star(\boldsymbol{\alpha}, \boldsymbol{\beta}), \boldsymbol{\alpha}, \boldsymbol{\beta})$$

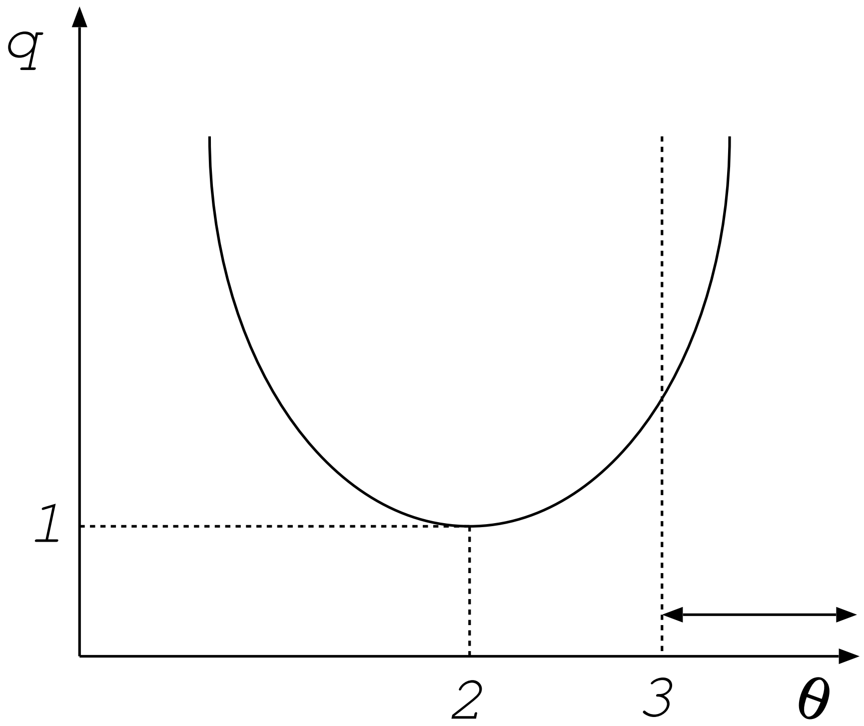
4. *Optimizar la función dual de Lagrange* (usualmente resolviendo $\boldsymbol{\nabla} \Lambda_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = 0$):

$$(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star) = \arg\max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha_i \geq 0} \Lambda_D(\boldsymbol{\alpha}, \boldsymbol{\beta})$$

5. *Solución final*:

$$\boldsymbol{\Theta}^\star = \boldsymbol{\Theta}^\star(\boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)$$

Multiplicadores de Lagrange: ejemplo



minimizar $q(\theta) = 1 + (\theta - 2)^2$ con $\theta \geq 3$

$$\Lambda(\theta, \alpha) = 1 + (\theta - 2)^2 - \alpha\ (\theta - 3)$$

$$\frac{\partial\ \Lambda(\theta, \alpha)}{\partial\ \theta} = 2\ (\theta - 2) - \alpha = 0 \Rightarrow \theta^\star(\alpha) = 2 + \frac{\alpha}{2}$$

$$\Lambda_D(\alpha) = \Lambda(\theta^\star(\alpha), \alpha) = 1 + \frac{\alpha^2}{4} - \alpha\ (\frac{\alpha}{2} - 1)$$

$$\frac{d\ \Lambda_D}{d\ \alpha} = \frac{\alpha}{2} - \frac{\alpha}{2} + 1 - \frac{\alpha}{2} = 1 - \frac{\alpha}{2} = 0 \Rightarrow$$

$$\alpha^\star = 2 \geq 0 \ \rightarrow \ \theta^\star = \theta^\star(\alpha^\star) = 3$$

Ejercicios:

a) minimizar $q(\theta) = 1 + (\theta - 2)^2$ con la condición de desigualdad $\theta \leq 3$?

b) minimizar $q(\theta) = 1 + (\theta - 2)^2$ con la condición de igualdad $q(\theta) + \theta = 4$

Multiplicadores de Lagrange: otro ejemplo

En una muestra S de una tarea de clasificación en *tres* clases se observan 4 datos de la clase $c = 1$, 2 datos de $c = 2$ y 1 dato de $c = 3$. Estimar por *máxima verosimilitud* las probabilidades a priori de las clases, $p_c, 1 \leq c \leq 3$.

Multiplicadores de Lagrange: otro ejemplo

En una muestra S de una tarea de clasificación en *tres* clases se observan 4 datos de la clase $c = 1$, 2 datos de $c = 2$ y 1 dato de $c = 3$. Estimar por *máxima verosimilitud* las probabilidades a priori de las clases, $p_c, 1 \leq c \leq 3$.

- Modelo: $P(c=1) = p_1,\ \ P(c=2) = p_2,\ \ P(c=3) = p_3,$
  $p_1 + p_2 + p_3 = 1, \quad \mathbf{\Theta} \equiv (p_1, p_2, p_3)^t$

Multiplicadores de Lagrange: otro ejemplo

En una muestra S de una tarea de clasificación en *tres* clases se observan 4 datos de la clase $c = 1$, 2 datos de $c = 2$ y 1 dato de $c = 3$. Estimar por *máxima verosimilitud* las probabilidades a priori de las clases, $p_c, 1 \leq c \leq 3$.

- Modelo: $P(c = 1) = p_1, \;\; P(c = 2) = p_2, \;\; P(c = 3) = p_3,$
  $p_1 + p_2 + p_3 = 1, \quad \boldsymbol{\Theta} \equiv (p_1, p_2, p_3)^t$

- Verosimilitud y logaritmo de la verosimilitud:

$$P(S \mid \boldsymbol{\Theta}) \;=\; \prod_{i=1}^{4} p_1 \prod_{j=1}^{2} p_2 \prod_{k=1}^{1} p_3 \;=\; p_1^4\, p_2^2\, p_3$$

$$q_S(\boldsymbol{\Theta}) \;=\; L_S(\boldsymbol{\Theta}) \;=\; \log P(S \mid \boldsymbol{\Theta}) \;=\; 4 \log p_1 + 2 \log p_2 + \log p_3$$

Multiplicadores de Lagrange: otro ejemplo

En una muestra S de una tarea de clasificación en *tres* clases se observan 4 datos de la clase $c=1$, 2 datos de $c=2$ y 1 dato de $c=3$. Estimar por *máxima verosimilitud* las probabilidades a priori de las clases, $p_c, 1 \leq c \leq 3$.

- Modelo: $P(c=1)=p_1,\ \ P(c=2)=p_2,\ \ P(c=3)=p_3,$
  $p_1+p_2+p_3=1, \quad \boldsymbol{\Theta} \equiv (p_1, p_2, p_3)^t$

- Verosimilitud y logaritmo de la verosimilitud:

$$P(S \mid \boldsymbol{\Theta}) \;=\; \prod_{i=1}^{4} p_1 \prod_{j=1}^{2} p_2 \prod_{k=1}^{1} p_3 \;=\; p_1^4\, p_2^2\, p_3$$

$$q_S(\boldsymbol{\Theta}) \;=\; L_S(\boldsymbol{\Theta}) \;=\; \log P(S \mid \boldsymbol{\Theta}) \;=\; 4\log p_1 + 2\log p_2 + \log p_3$$

- Estimación de máxima verosimilitud:

$$\boldsymbol{\Theta}^\star \;=\; \arg\max_{\boldsymbol{\Theta}} L_S(\boldsymbol{\Theta}) \;=\; \arg\max_{p_1,p_2,p_3} \;(4\log p_1 + 2\log p_2 + \log p_3)$$

Multiplicadores de Lagrange: otro ejemplo

En una muestra S de una tarea de clasificación en *tres* clases se observan 4 datos de la clase $c = 1$, 2 datos de $c = 2$ y 1 dato de $c = 3$. Estimar por *máxima verosimilitud* las probabilidades a priori de las clases, $p_c, 1 \leq c \leq 3$.

- Modelo: $P(c=1) = p_1,\ P(c=2) = p_2,\ P(c=3) = p_3,$
  $p_1 + p_2 + p_3 = 1, \quad \boldsymbol{\Theta} \equiv (p_1, p_2, p_3)^t$

- Verosimilitud y logaritmo de la verosimilitud:

$$P(S \mid \boldsymbol{\Theta}) \;=\; \prod_{i=1}^{4} p_1 \prod_{j=1}^{2} p_2 \prod_{k=1}^{1} p_3 \;=\; p_1^4\, p_2^2\, p_3$$

$$q_S(\boldsymbol{\Theta}) \;=\; L_S(\boldsymbol{\Theta}) \;=\; \log P(S \mid \boldsymbol{\Theta}) \;=\; 4 \log p_1 + 2 \log p_2 + \log p_3$$

- Estimación de máxima verosimilitud:

$$\boldsymbol{\Theta}^\star \;=\; \arg\max_{\boldsymbol{\Theta}} L_S(\boldsymbol{\Theta}) \;=\; \arg\max_{p_1, p_2, p_3} \; (4 \log p_1 + 2 \log p_2 + \log p_3)$$

Solución: Resolver $\boldsymbol{\nabla} L_S(\boldsymbol{\Theta}) = \mathbf{0}$

¿Es suficiente?

Multiplicadores de Lagrange: otro ejemplo

En una muestra S de una tarea de clasificación en *tres* clases se observan 4 datos de la clase $c=1$, 2 datos de $c=2$ y 1 dato de $c=3$. Estimar por *máxima verosimilitud* las probabilidades a priori de las clases, $p_c, 1 \leq c \leq 3$.

- Modelo: $P(c=1)=p_1,\ \ P(c=2)=p_2,\ \ P(c=3)=p_3,$
  $p_1+p_2+p_3=1, \quad \mathbf{\Theta} \equiv (p_1,p_2,p_3)^t$

- Verosimilitud y logaritmo de la verosimilitud:

$$P(S \mid \mathbf{\Theta}) \;=\; \prod_{i=1}^{4} p_1 \prod_{j=1}^{2} p_2 \prod_{k=1}^{1} p_3 \;=\; p_1^4\, p_2^2\, p_3$$

$$q_S(\mathbf{\Theta}) \;=\; L_S(\mathbf{\Theta}) \;=\; \log P(S \mid \mathbf{\Theta}) \;=\; 4\log p_1 + 2\log p_2 + \log p_3$$

- Estimación de máxima verosimilitud:

$$\mathbf{\Theta}^\star \;=\; \arg\max_{\mathbf{\Theta}} L_S(\mathbf{\Theta}) \;=\; \arg\max_{\substack{p_1,p_2,p_3 \\ p_1+p_2+p_3=1}} \;(4\log p_1 + 2\log p_2 + \log p_3)$$

- Problema de optimización con restricciones al que aplicaremos la técnica de los multiplicadores de Lagrange

Ejemplo: aplicación de la técnica de multiplicadores de Lagrange

- Lagrangiana: $\Lambda(p_1, p_2, p_3, \beta) = 4\log p_1 + 2\log p_2 + \log p_3 + \beta\,(1 - p_1 - p_2 - p_3)$
- Soluciones óptimas en función del multiplicador de Lagrange:

$$\left.\begin{array}{l} \dfrac{\partial\Lambda}{\partial p_1} = \dfrac{4}{p_1} - \beta = 0 \\ \dfrac{\partial\Lambda}{\partial p_2} = \dfrac{2}{p_2} - \beta = 0 \\ \dfrac{\partial\Lambda}{\partial p_3} = \dfrac{1}{p_3} - \beta = 0 \end{array}\right\} \qquad \begin{array}{l} p_1^\star(\beta) = \dfrac{4}{\beta} \\ p_2^\star(\beta) = \dfrac{2}{\beta} \\ p_3^\star(\beta) = \dfrac{1}{\beta} \end{array}$$

- Función dual de Lagrange:

$$\Lambda_D(\beta) \;=\; 4\log\frac{4}{\beta} + 2\log\frac{2}{\beta} + \log\frac{1}{\beta} + \beta(1 - \frac{4}{\beta} - \frac{2}{\beta} - \frac{1}{\beta}) \;=\; \beta - 7\log\beta - 7 + 10\log 2$$

- Valor óptimo del multiplicador de Lagrange: $\dfrac{d\Lambda_D}{d\beta} \;=\; 1 - \dfrac{7}{\beta} \;=\; 0 \;\Rightarrow\; \beta^\star \;=\; 7$
- Solución final: $p_1^\star = p_1^\star(\beta^\star) = \dfrac{4}{7} \qquad p_2^\star = p_2^\star(\beta^\star) = \dfrac{2}{7} \qquad p_3^\star = p_3^\star(\beta^\star) = \dfrac{1}{7}$

EJERCICIO: Demostrar que en cualquier problema de clasificación en C clases, la estimación de máxima verosimilitud de la probabilidad a priori de cada clase c, $1 \leq c \leq C$, es $\hat{p}_c = n_c/N$, donde $N = \sum_c n_c$ es el número total de datos observados y n_c es el número de datos de la clase c.

Teorema de Kuhn-Tucker

Consideremos un problema de optimización, $\mathcal{O}$, definido por:

$$\begin{array}{ll} \text{minimizar} & q(\boldsymbol{\Theta}), \quad \boldsymbol{\Theta} \in \mathbb{R}^D \\ \text{sujecto a} & v_i(\boldsymbol{\Theta}) \geq 0, \;\; 1 \leq i \leq k \\ & u_i(\boldsymbol{\Theta}) = 0, \;\; 1 \leq i \leq m \end{array}$$

y la correspondiente función Lagrangiana:

$$\Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = q(\boldsymbol{\Theta}) - \sum_{i=1}^{k} \alpha_i v_i(\boldsymbol{\Theta}) + \sum_{i=1}^{m} \beta_i u_i(\boldsymbol{\Theta})$$

Teorema de Kuhn-Tucker: si $\exists\ \boldsymbol{\Theta}^\star, \boldsymbol{\alpha}^\star$, $\boldsymbol{\beta}^\star$ tales que:

$$\boldsymbol{\nabla}_{\boldsymbol{\Theta}}\, \Lambda(\boldsymbol{\Theta}, \boldsymbol{\alpha}^\star, \boldsymbol{\beta}^\star)\big|_{\boldsymbol{\Theta}^\star} \;=\; \mathbf{0};$$

$$\alpha_i^\star \geq 0, \;\; v_i(\boldsymbol{\Theta}^\star) \geq 0, \;\; \alpha_i^\star\, v_i(\boldsymbol{\Theta}^\star) = 0, \quad 1 \leq i \leq k,$$

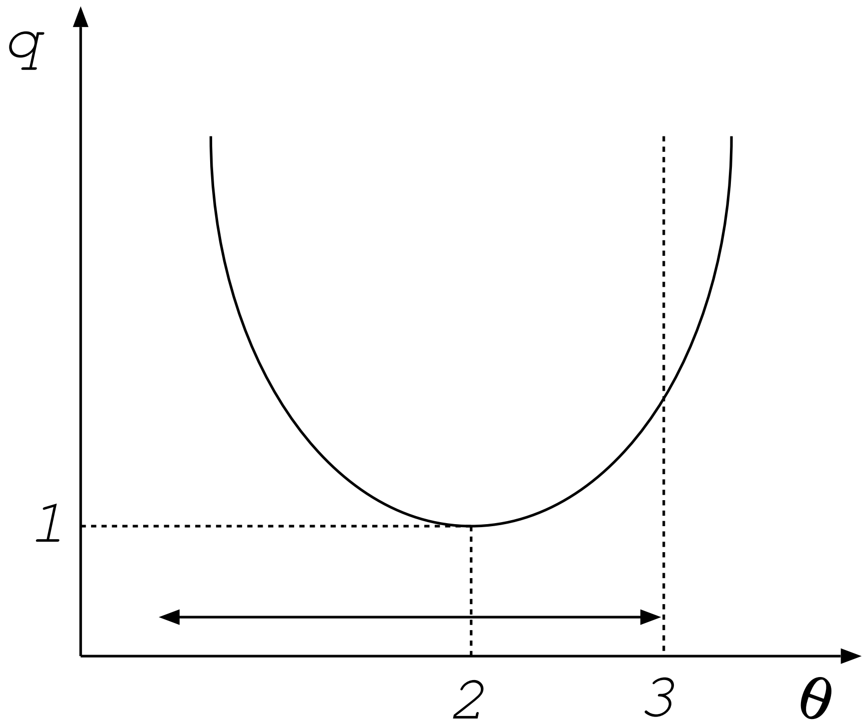
$$u_i(\boldsymbol{\Theta}^\star) = 0, \quad 1 \leq i \leq m$$

entonces $q(\boldsymbol{\Theta}^\star)$ es solución al problema $\mathcal{O}$.

$\alpha_i^\star v_i(\boldsymbol{\Theta}^\star) = 0,\ 1 \leq i \leq k$: *condiciones complementarias de Karush-Kuhn-Tucker (KKT)*.

Multiplicadores de Lagrange y KKT: ejemplo

En el ejemplo anterior, ¿qué ocurre si la condición de desigualdad es $\theta \leq 3$?



minimizar $q(\theta) = 1 + (\theta - 2)^2$ con $3 - \theta \geq 0$

$$\Lambda(\theta, \alpha) = 1 + (\theta - 2)^2 - \alpha\,(3 - \theta)$$

$$\frac{\partial\, \Lambda(\theta, \alpha)}{\partial\, \theta} = 2\,(\theta - 2) + \alpha = 0 \Rightarrow \theta^\star(\alpha) = 2 - \frac{\alpha}{2}$$

KKT: $\alpha^\star v(\theta^\star(\alpha^\star)) = 0 \;\Rightarrow\; \alpha^\star(3 - 2 + \frac{1}{2}\alpha^\star) = 0$

$$\Rightarrow \begin{cases} \alpha^\star = 0 \\ \alpha^\star = -2 < 0 \rightarrow \text{¡VIOLA } \alpha \geq 0\,\text{!} \end{cases}$$

KKT $\Rightarrow\; \alpha^\star = 0 \;\Rightarrow\; \theta^\star = \theta^\star(\alpha^\star) = 2$

Ejercicio: mediante el método de KKT, minimizar $q(\theta) = 1 + (\theta - 2)^2$ con $3 \leq \theta$.

Index

Descenso por gradiente

Problema: Minimización sin restricciones de una función objetivo $q : \mathbb{R}^D \rightarrow \mathbb{R}$, cuando una solución analítica no es viable:

$$\mathbf{\Theta}^\star = \arg\min_{\mathbf{\Theta}} \; q(\mathbf{\Theta})$$

- Una solución: construir una secuencia de puntos $\mathbf{\Theta}(1), \ldots, \mathbf{\Theta}(k), \ldots,$ que converja a $\mathbf{\Theta}^\star$.
- Cada valor $\mathbf{\Theta}(k)$ se contruye a partir del anterior $\mathbf{\Theta}(k-1)$ en la secuencia dependiendo de las derivadas de la función en el punto $\mathbf{\Theta}(k)$.
- Recordatorio: Gradiente de $q(\mathbf{\Theta})$ en el punto $\mathbf{\Theta} = \mathbf{\Theta}(k)$: vector formado por las derivadas parciales de la función calculadas en $\mathbf{\Theta}(k)$:

$$\nabla q \,\Big|_{\mathbf{\Theta}=\mathbf{\Theta}(k)} \equiv \left(\frac{\partial q}{\partial \theta_1}\Big|_{\mathbf{\Theta}=\mathbf{\Theta}(k)}, \ldots, \frac{\partial q}{\partial \theta_D}\Big|_{\mathbf{\Theta}=\mathbf{\Theta}(k)} \right)^t$$

Descenso por gradiente: algoritmo general

$$\begin{aligned}\boldsymbol{\Theta}(1) &= \text{arbitrario} \\ \boldsymbol{\Theta}(k+1) &= \boldsymbol{\Theta}(k) - \rho_k \boldsymbol{\nabla} q(\boldsymbol{\Theta})\,\big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}(k)}\end{aligned}$$

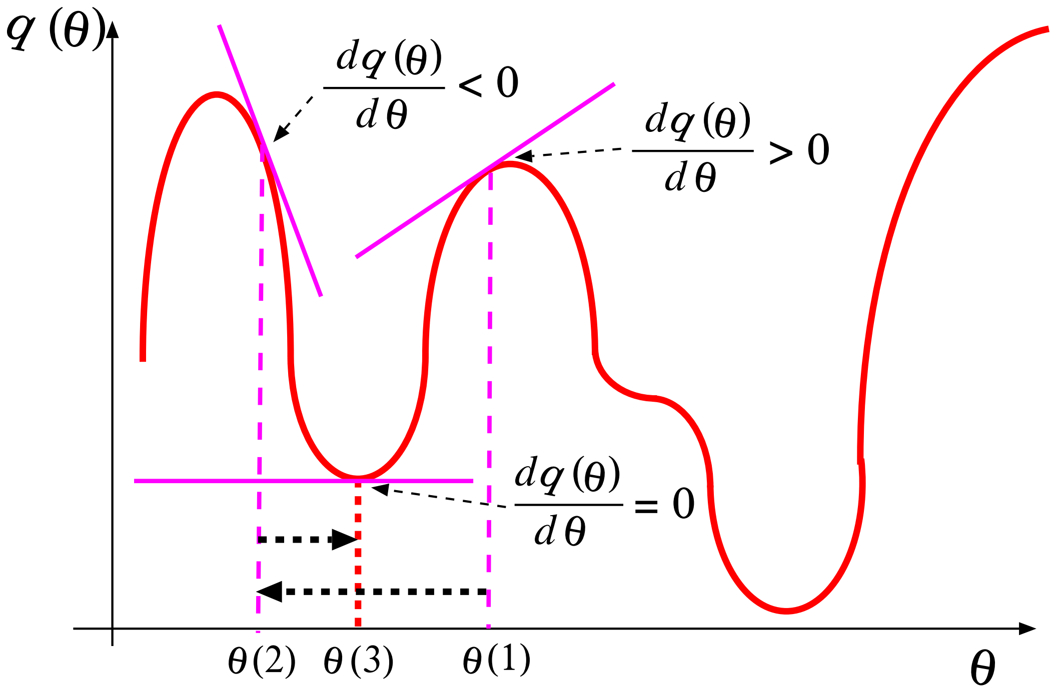
Donde $\rho_k \in \mathbb{R}^{>0}$ se denomina *tamaño del paso de descenso*

Ejemplo en $\mathbb{R}^1$ *con* $\boldsymbol{\Theta} \stackrel{def}{=} \theta$*:*

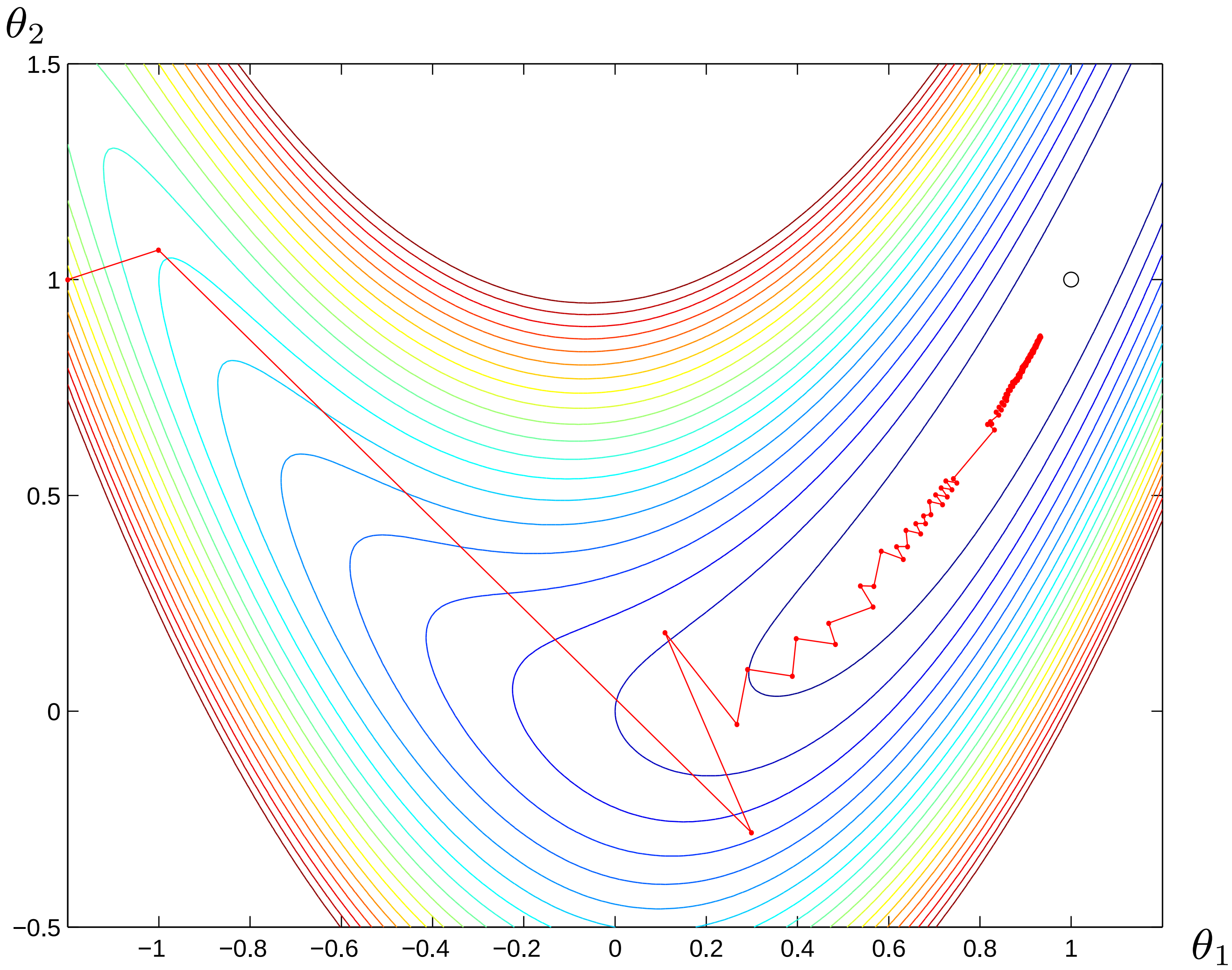
$$\theta(2) = \theta(1) - \rho_1 \frac{dq}{d\theta}\bigg|_{\theta(1)}$$

$$\theta(3) = \theta(2) - \rho_2 \frac{dq}{d\theta}\bigg|_{\theta(2)}$$

$$\theta(3) \equiv \theta^\star, \qquad \frac{dq}{d\theta}\bigg|_{\theta(3)} = 0$$



Descenso por gradiente: Ejemplo en $\mathbb{R}^2$



Curvas de nivel de la función de Rosenbrock[1] $q(\theta_1, \theta_2) = 10(\theta_2 - \theta_1^2)^2 + (1 - \theta_1)^2$ y trayectoria seguida por el vector de parámetros $\boldsymbol{\theta} = (\theta_1, \theta_2)^t$.

1. Figuras basadas en la presentación de R. Hauser http://people.maths.ox.ac.uk/hauser/hauser_lecture2.pdf.

Convergencia y tamaño del paso

- TEOREMA GENERAL DE CONVERGENCIA:

  Sea $H(q, \boldsymbol{\Theta})$ la matriz de segundas derivadas (*Hessiana*) de q evaluada en $\boldsymbol{\Theta}$:

  $$H_{ij}(q, \boldsymbol{\Theta}) \stackrel{\text{def}}{=} \frac{\partial^2 q(\boldsymbol{\Theta})}{\partial\theta_i \, \partial\theta_j}$$

  Sean $\lambda_l(k)$ los valores propios de $H(q, \boldsymbol{\Theta}(k))$ en el paso k-ésimo del algoritmo de *descenso por gradiente*.

  Si $|1-\lambda_l(k)\rho_k|<1 \; \forall l$, *entonces* $\boldsymbol{\Theta}(k)$ *tiende a un minimo local de* $q(\boldsymbol{\Theta})$ *cuando* $k \to \infty$

- INFLUENCIA DEL TAMAÑO DEL PASO:
  - $\rho < 2/\lambda_{\max}$ garantiza la convergencia
  - $\rho \gg \Rightarrow$ convergencia rápida y tendencia a oscilar
  - $\rho \ll \Rightarrow$ convergencia lenta

Ejemplo: clasificador lineal en dos clases

- Clasificador en *dos* clases basado *funciones discriminantes lineales* (FDL):

$$f(\boldsymbol{x}) = \underset{1\leq c\leq 2}{\arg\max}\ \phi_c(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, \ldots, x_d)^t \in \mathbb{R}^d, \quad \phi_c : \mathbb{R}^d \to \mathbb{R},\ 1 \leq c \leq 2$$

Cada FDL ϕ_c está definida por un vector de pesos $\boldsymbol{\theta}_c \in \mathbb{R}^D$ donde $D = d + 1$

En *notación homogénea* se añade una componente, $x_0 \equiv 1$, a $\boldsymbol{x}$, con lo que:

$$\phi_c(\boldsymbol{x}) \ = \ \sum_{j=1}^{d} \theta_{c_j} x_j + \theta_{c_0} \ = \ \sum_{j=0}^{d} \theta_{c_j} x_j \ = \ {\boldsymbol{\theta}_c}^t \boldsymbol{x}\,, \quad 1 \leq c \leq 2$$

La *frontera de decisión* se define como $F(\phi_1, \phi_2) \equiv F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \{\boldsymbol{x} : {\boldsymbol{\theta}_1}^t \boldsymbol{x} = {\boldsymbol{\theta}_2}^t \boldsymbol{x}\}$

Ejemplo: clasificador lineal en dos clases

- Clasificador en *dos* clases basado *funciones discriminantes lineales* (FDL):

$$f(\boldsymbol{x}) = \operatorname*{arg\,max}_{1\leq c\leq 2} \ \phi_c(\boldsymbol{x}), \quad \boldsymbol{x} = (x_1, \ldots, x_d)^t \in \mathbb{R}^d, \quad \phi_c : \mathbb{R}^d \to \mathbb{R}, \ 1 \leq c \leq 2$$

Cada FDL ϕ_c está definida por un vector de pesos $\boldsymbol{\theta}_c \in \mathbb{R}^D$ donde $D = d + 1$

En *notación homogénea* se añade una componente, $x_0 \equiv 1$, a $\boldsymbol{x}$, con lo que:

$$\phi_c(\boldsymbol{x}) \ = \ \sum_{j=1}^{d} \theta_{c_j} x_j + \theta_{c_0} \ = \ \sum_{j=0}^{d} \theta_{c_j} x_j \ = \ \boldsymbol{\theta}_c{}^t \boldsymbol{x}\,, \quad 1 \leq c \leq 2$$

La *frontera de decisión* se define como $F(\phi_1, \phi_2) \equiv F(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \{\boldsymbol{x} : \boldsymbol{\theta}_1{}^t \boldsymbol{x} = \boldsymbol{\theta}_2{}^t \boldsymbol{x}\}$

- Simplificación (si $C = 2$): etiquetar las clases $\{1, 2\}$ como $\{+1, -1\}$ y usar un único vector de pesos $\boldsymbol{\theta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$. $(F(\boldsymbol{\theta}) = \{\boldsymbol{x} : \boldsymbol{\theta}^t \boldsymbol{x} = 0\})$

Clasificador, $f_{\boldsymbol{\theta}} : \mathbb{R}^D \to \{-1, +1\}$:

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) \ = \ \begin{cases} +1 & \text{si } \ \boldsymbol{\theta}^t \boldsymbol{x} \geq 0 \qquad\qquad (\phi_1(\boldsymbol{x}) \geq \phi_2(\boldsymbol{x}) \Rightarrow \boldsymbol{\theta}_1{}^t \boldsymbol{x} \geq \boldsymbol{\theta}_2{}^t \boldsymbol{x}) \\ -1 & \text{si } \ \boldsymbol{\theta}^t \boldsymbol{x} < 0 \qquad\qquad (\phi_1(\boldsymbol{x}) < \phi_2(\boldsymbol{x}) \Rightarrow \boldsymbol{\theta}_1{}^t \boldsymbol{x} < \boldsymbol{\theta}_2{}^t \boldsymbol{x}) \end{cases}$$

Aprendizaje de funciones discriminantes lineales

- Sea $S = \{(\boldsymbol{x}_1, c_1), \dots, (\boldsymbol{x}_N, c_N)\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $c_n \in \{+1, -1\}$ una muestra de entrenamiento. S es *linealmente separable* (LS) si $\exists\, \boldsymbol{\theta} \in \mathbb{R}^D$ $(D = d + 1)$ tal que:

$$\forall n,\ 1 \leq n \leq N, \quad \boldsymbol{\theta}^t \boldsymbol{x}_n \begin{cases} \geq 0 & \text{if } c_n = +1 \\ < 0 & \text{if } c_n = -1 \end{cases}; \quad \text{es decir,} \quad c_n\, \boldsymbol{\theta}^t \boldsymbol{x}_n \geq 0$$

- Aprendizaje: Dada S, encontrar un vector de pesos $\hat{\boldsymbol{\theta}}$ que la separe; es decir, que satisfaga el sistema de N inecuaciones:

$$c_n\, \boldsymbol{\theta}^t \boldsymbol{x}_n \geq 0, \;\; 1 \leq n \leq N$$

Aprendizaje de funciones discriminantes lineales

- Sea $S = \{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $c_n \in \{+1, -1\}$ una muestra de entrenamiento. S es *linealmente separable* (LS) si $\exists\, \boldsymbol{\theta} \in \mathbb{R}^D$ $(D = d+1)$ tal que:

$$\forall n,\ 1 \leq n \leq N, \quad \boldsymbol{\theta}^t \boldsymbol{x}_n \begin{cases} \geq 0 & \text{if } c_n = +1 \\ < 0 & \text{if } c_n = -1 \end{cases}; \quad \text{es decir,} \quad c_n\, \boldsymbol{\theta}^t \boldsymbol{x}_n \geq 0$$

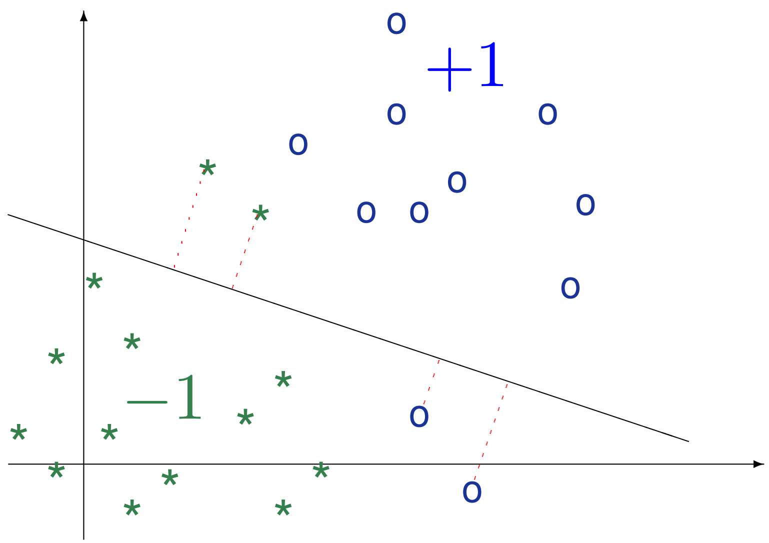
- Aprendizaje: Dada S, encontrar un vector de pesos $\hat{\boldsymbol{\theta}}$ que la separe; es decir, que satisfaga el sistema de N inecuaciones:

$$c_n\, \boldsymbol{\theta}^t \boldsymbol{x}_n \geq 0, \quad 1 \leq n \leq N$$

- Planteamiento equivalente:
  minimizar la función: $q_S : \mathbb{R}^D \to \mathbb{R}^{\geq 0}$:

$$q_S(\boldsymbol{\theta}) = \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\, \boldsymbol{\theta}^t \boldsymbol{x} < 0}} -c\, \boldsymbol{\theta}^t \boldsymbol{x}$$

proporcional a la suma de *segmentos punteados* $\longrightarrow$



Aprendizaje de funciones discriminantes lineales

- Sea $S = \{(\boldsymbol{x}_1, c_1), \dots, (\boldsymbol{x}_N, c_N)\}$, $\boldsymbol{x}_n \in \mathbb{R}^D$, $c_n \in \{+1, -1\}$ una muestra de entrenamiento. S es *linealmente separable* (LS) si $\exists\, \boldsymbol{\theta} \in \mathbb{R}^D$ $(D = d+1)$ tal que:

$$\forall n,\ 1 \leq n \leq N, \quad \boldsymbol{\theta}^t \boldsymbol{x}_n \begin{cases} \geq 0 & \text{if } c_n = +1 \\ < 0 & \text{if } c_n = -1 \end{cases}; \quad \text{es decir,} \quad c_n\, \boldsymbol{\theta}^t \boldsymbol{x}_n \geq 0$$

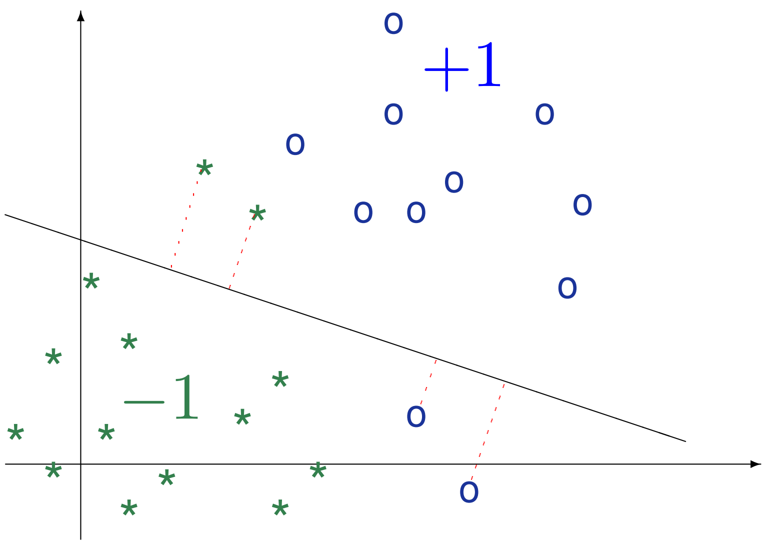
- Aprendizaje: Dada S, encontrar un vector de pesos $\hat{\boldsymbol{\theta}}$ que la separe; es decir, que satisfaga el sistema de N inecuaciones:

$$c_n\, \boldsymbol{\theta}^t \boldsymbol{x}_n \geq 0, \quad 1 \leq n \leq N$$

- Planteamiento equivalente:
  minimizar la función: $q_S : \mathbb{R}^D \rightarrow \mathbb{R}^{\geq 0}$:

$$q_S(\boldsymbol{\theta}) = \sum_{\substack{(\boldsymbol{x}, c) \in S \\ c\, \boldsymbol{\theta}^t \boldsymbol{x} < 0}} -c\, \boldsymbol{\theta}^t \boldsymbol{x}$$

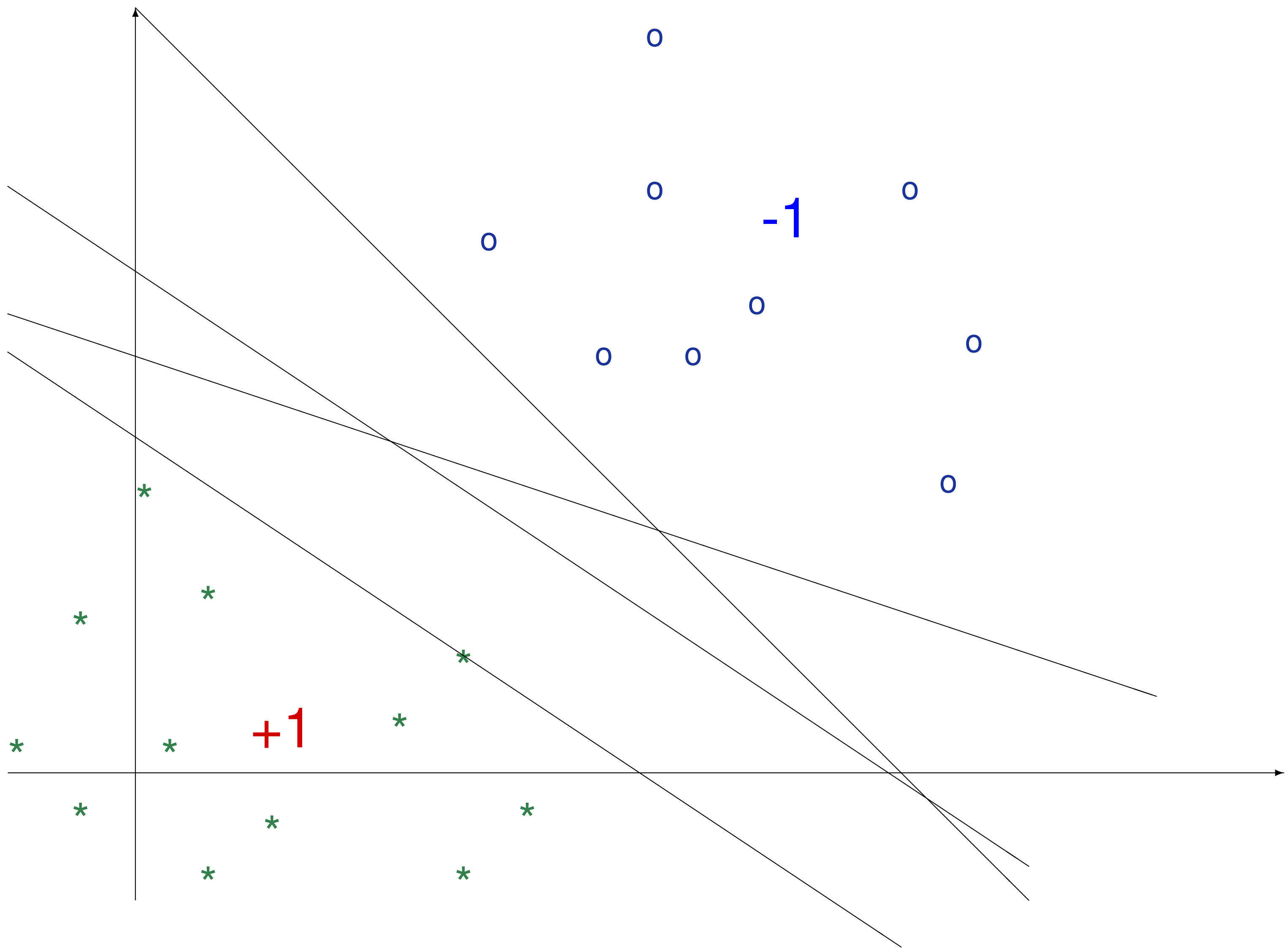
  proporcional a la suma de *segmentos punteados* $\longrightarrow$



- Sea $\boldsymbol{\theta}^\star = \arg\min_{\boldsymbol{\theta}} q_S(\boldsymbol{\theta})$. Si S es LS, entonces $q_S(\boldsymbol{\theta}^\star) = 0$, $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^\star$.

Aprendizaje de funciones discriminantes lineales

Ejemplo de muestras linealmente separables en $\mathbb{R}^2$ y posibles soluciones



Algoritmo perceptrón

$$\boldsymbol{\nabla} q_S(\boldsymbol{\theta}) = \boldsymbol{\nabla} \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\ \boldsymbol{\theta}^t\boldsymbol{x} < 0}} -c\ \boldsymbol{\theta}^t\boldsymbol{x} = \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\ \boldsymbol{\theta}^t\boldsymbol{x} < 0}} -c\ \boldsymbol{x}$$

$$\boxed{\begin{aligned} \boldsymbol{\theta}(1) &= \text{arbitrario} \\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\ \boldsymbol{\theta}(k)^t\boldsymbol{x} < 0}} c\ \boldsymbol{x} \end{aligned}}$$

El *tamaño del paso* aquí se denomina factor de aprendizaje

Algoritmo perceptrón

$$\nabla q_S(\boldsymbol{\theta}) = \nabla \sum_{\substack{(\boldsymbol{x},c)\in S\\ c\,\boldsymbol{\theta}^t\boldsymbol{x}<0}} -c\,\boldsymbol{\theta}^t\boldsymbol{x} = \sum_{\substack{(\boldsymbol{x},c)\in S\\ c\,\boldsymbol{\theta}^t\boldsymbol{x}<0}} -c\,\boldsymbol{x}$$

$$\boxed{\begin{aligned} \boldsymbol{\theta}(1) &= \text{arbitrario} \\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k \sum_{\substack{(\boldsymbol{x},c)\in S\\ c\,\boldsymbol{\theta}(k)^t\boldsymbol{x}<0}} c\,\boldsymbol{x} \end{aligned}}$$

El *tamaño del paso* aquí se denomina factor de aprendizaje

Algoritmo perceptrón muestra a muestra ("online"): (DEMO)

$$\begin{aligned} \boldsymbol{\theta}(1) &= \text{arbitrario} \\ \boldsymbol{\theta}(k+1) &= \begin{cases} \boldsymbol{\theta}(k) & c(k)\,\boldsymbol{\theta}^t\boldsymbol{x}(k) \geq 0 \\ \boldsymbol{\theta}(k) + \rho_k\, c(k)\,\boldsymbol{x}(k) & c(k)\,\boldsymbol{\theta}^t\boldsymbol{x}(k) < 0 \end{cases} \end{aligned}$$

Algoritmo perceptrón

$$\nabla q_S(\boldsymbol{\theta}) = \nabla \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\ \boldsymbol{\theta}^t\boldsymbol{x}<0}} - c\ \boldsymbol{\theta}^t\boldsymbol{x} = \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\ \boldsymbol{\theta}^t\boldsymbol{x}<0}} - c\ \boldsymbol{x}$$

$$\boxed{\begin{aligned} \boldsymbol{\theta}(1) &= \text{arbitrario} \\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k \sum_{\substack{(\boldsymbol{x},c)\in S \\ c\ \boldsymbol{\theta}(k)^t\boldsymbol{x}<0}} c\ \boldsymbol{x} \end{aligned}}$$

El *tamaño del paso* aquí se denomina factor de aprendizaje

Algoritmo perceptrón muestra a muestra ("online"): (DEMO)

$$\boldsymbol{\theta}(1) = \text{arbitrario}$$

$$\boldsymbol{\theta}(k+1) = \begin{cases} \boldsymbol{\theta}(k) & c(k)\ \boldsymbol{\theta}^t\boldsymbol{x}(k) \ge 0 \\ \boldsymbol{\theta}(k) + \rho_k\ c(k)\ \boldsymbol{x}(k) & c(k)\ \boldsymbol{\theta}^t\boldsymbol{x}(k) < 0 \end{cases}$$

TEOREMA DEL PERCEPTRÓN:

Si S es LS y ρ_k es positivo y decreciente o creciente sublinealmente con k, el algorithmo perceptrón converge a una solución en un número finito de iteraciones

Regresión lineal mediante descenso por gradiente

- Sea $f_{\boldsymbol{\theta}} : \mathbb{R}^D \to \mathbb{R}$ una función *lineal* ($D = d + 1$):

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) \stackrel{\text{def}}{=} \boldsymbol{\theta}^t \boldsymbol{x}, \quad \boldsymbol{x}, \boldsymbol{\theta} \in \mathbb{R}^D$$

  y S una muestra de entrenamiento:

$$S = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}, \quad \boldsymbol{x}_n \in \mathbb{R}^D, \; y_n \in \mathbb{R}$$

- Aprendizaje: Calcular $\hat{\boldsymbol{\theta}}$ tal que;

$$\hat{\boldsymbol{\theta}}^t \boldsymbol{x}_n \; \approx \; y_n, \quad 1 \leq n \leq N$$

- *Aproximación por mínimos cuadrados*:
  minimizar la función de Widrow-Hoff:

$$q_S(\boldsymbol{\theta}) = \frac{1}{2} \sum_{n=1}^{N} \left(\boldsymbol{\theta}^t \boldsymbol{x}_n - y_n\right)^2$$

- Solución: descenso por gradiente.

Algoritmo de Widrow-Hoff (Adaline)

$$\boldsymbol{\nabla} q_S(\boldsymbol{\theta}) \;=\; \boldsymbol{\nabla} \frac{1}{2} \sum_{n=1}^{N} \left(\boldsymbol{\theta}^t \boldsymbol{x}_n - y_n\right)^2 \;=\; \sum_{n=1}^{N} \left(\boldsymbol{\theta}^t \boldsymbol{x}_n - y_n\right) \boldsymbol{x}_n$$

$$\boxed{\begin{aligned} \boldsymbol{\theta}(1) \;&=\; \text{arbitrario} \\ \boldsymbol{\theta}(k+1) \;&=\; \boldsymbol{\theta}(k) + \rho_k \sum_{n=1}^{N} (y_n - \boldsymbol{\theta}(k)^t \boldsymbol{x}_n)\, \boldsymbol{x}_n \end{aligned}}$$

Algoritmo de Widrow-Hoff (Adaline)

$$\nabla q_S(\boldsymbol{\theta}) \;=\; \nabla \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{\theta}^t \boldsymbol{x}_n - y_n\right)^2 \;=\; \sum_{n=1}^{N}\left(\boldsymbol{\theta}^t \boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n$$

$$\boxed{\begin{aligned} \boldsymbol{\theta}(1) &= \text{arbitrario} \\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k \sum_{n=1}^{N}(y_n - \boldsymbol{\theta}(k)^t \boldsymbol{x}_n)\,\boldsymbol{x}_n \end{aligned}}$$

Algoritmo muestra a muestra: (DEMO)

$$\begin{aligned} \boldsymbol{\theta}(1) &= \text{arbitrario} \\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k \left(y(k) - \boldsymbol{\theta}(k)^t \boldsymbol{x}(k)\right)\boldsymbol{x}(k) \end{aligned}$$

Algoritmo de Widrow-Hoff (Adaline)

$$\nabla q_S(\boldsymbol{\theta}) = \nabla \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{\theta}^t\boldsymbol{x}_n - y_n\right)^2 = \sum_{n=1}^{N}\left(\boldsymbol{\theta}^t\boldsymbol{x}_n - y_n\right)\boldsymbol{x}_n$$

$$\boxed{\begin{aligned}\boldsymbol{\theta}(1) &= \text{arbitrario}\\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k \sum_{n=1}^{N}(y_n - \boldsymbol{\theta}(k)^t\boldsymbol{x}_n)\,\boldsymbol{x}_n\end{aligned}}$$

Algoritmo muestra a muestra: (DEMO)

$$\begin{aligned}\boldsymbol{\theta}(1) &= \text{arbitrario}\\ \boldsymbol{\theta}(k+1) &= \boldsymbol{\theta}(k) + \rho_k\,\left(y(k) - \boldsymbol{\theta}(k)^t\boldsymbol{x}(k)\right)\,\boldsymbol{x}(k)\end{aligned}$$

TEOREMA:

Si $\rho_k = \rho_1/k,\ \rho_1 > 0,\ \hat{\boldsymbol{\theta}} = \lim_{k\to\infty}\boldsymbol{\theta}(k)$ *satisface* $\nabla q_S(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$

Ejercicios propuestos

Ejercicio:
Mostrar la traza de tres iteraciones de descenso por gradiente para minimizar $q(\boldsymbol{\theta}) = (\theta_1 - 1)^2 + (\theta_2 - 2)^2 + \theta_1\theta_2$, con $\rho_k = 1/(2k)$ y $\boldsymbol{\Theta}(1) = (-1, 1)^t$

Ejercicio:
Existe una variante de la función de Widrow-Hoff que incluye un término de regularización con el objetivo de que los pesos no se hagan demasiado grandes:

$$q_S(\boldsymbol{\theta}) = \frac{1}{2}\sum_{n=1}^{N}\left(\boldsymbol{\theta}^t\boldsymbol{x}_n - y_n\right)^2 + \frac{\boldsymbol{\theta}^t\boldsymbol{\theta}}{2}$$

Aplicando la técnica de descenso por gradiente, obtener la correspondiente variante del algoritmo de Widrow-Hoff y la correspondiente versión muestra a muestra. e

Index

Aprendizaje de modelos probabilísticos con variables latentes

- Se suele usar el criterio de *máxima verosimilitud*; es decir, $q_S(\mathbf{\Theta}) \equiv L_S(\mathbf{\Theta})$.
- En ocasiones los datos observados *no* contienen suficiente información sobre cómo han sido generados por los modelos probabilísticos asumidos.
- Por ejemplo, en los modelos ocultos de Markov los datos de entrenamiento son cadenas de símbolos, sin información sobre qué *secuencia de estados* ha producido cada cadena.
- Otro ejemplo típico son los modelos definidos como combinación lineal ("mezcla" o "mixtura") de distribuciones de probabilidad. Los coeficientes de combinación son parámetros a aprender, pero los datos de entrenamiento no contienen información sobre la distribución con que se ha generado cada dato.
- La información ausente en los datos de entrenamiento generalmente se denomina datos *perdidos*, o variables *latentes* u *ocultas*.
- Las técnicas simples de optimización resultan insuficientes para la estimación de los parámetros de estos modelos.

Ejemplo: mezcla de 2 gaussianas en $\mathbb{R}^2$ y modelo generador

- En el caso de una única gaussiana las muestras se generan en un paso:

  1. Escoger $\boldsymbol{x} \in \mathbb{R}^2$, de acuerdo con la distribución $p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)$

- Si el modelo es una *mezcla* de 2 gaussianas, el proceso de generación se compone de dos etapas:

  1. Con probabilidad $P(1) = \alpha_1 \equiv \alpha$ escoger la primera componente de la mezcla o con $P(2) = \alpha_2 = 1 - \alpha$ escoger la segunda componente con la que se va a generar $\boldsymbol{x}$. Sea $k \in \{1, 2\}$ la distribución escogida.
  2. Escoger $\boldsymbol{x}$, según la distribución definida por la k-ésima gaussiana, $p(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k)$

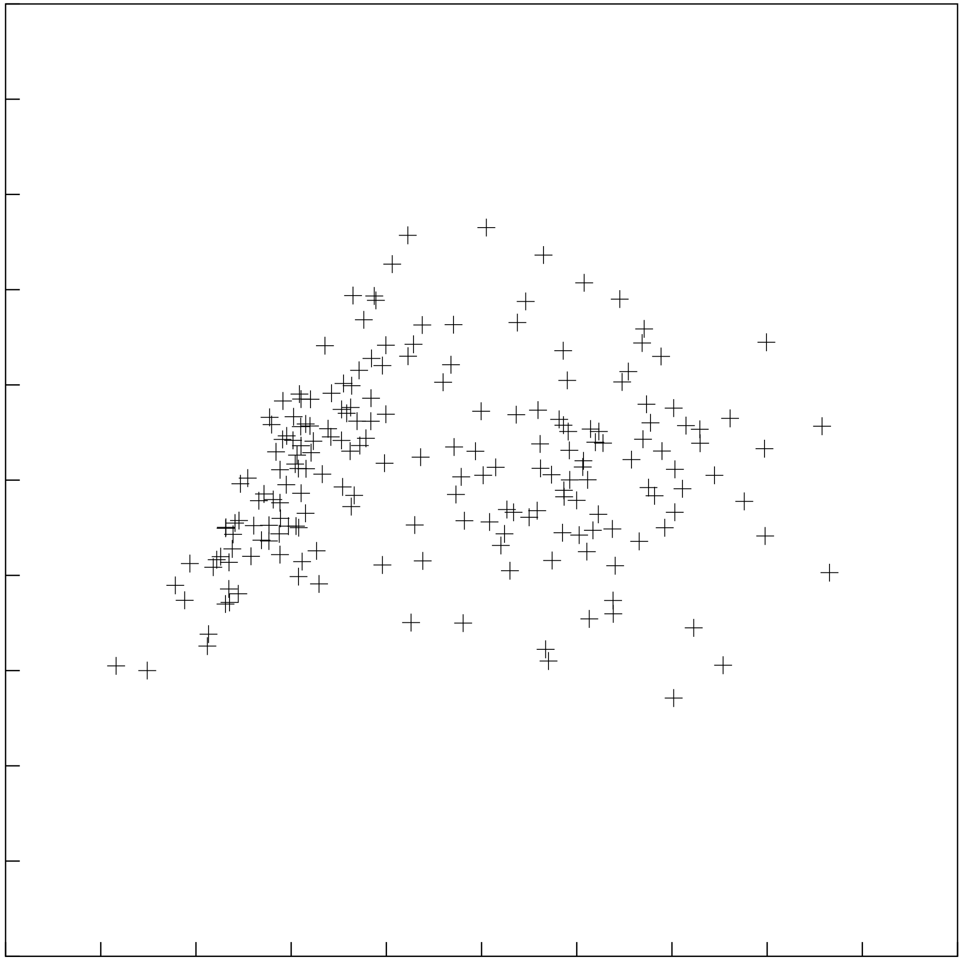
  – $\boldsymbol{x}$ es el dato observable y k es el valor de una variable oculta z. Los datos observables junto con los ocultos se denominan datos completos.

  – Probabilidad con la que se genera $\boldsymbol{x}$ según este proceso:

$$p(\boldsymbol{x}) = \sum_{k=1}^{2} p(z = k, \boldsymbol{x}) = \sum_{k=1}^{2} P(z = k)\; p(\boldsymbol{x} \mid k) \equiv \alpha\; p(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1) + (1 - \alpha) p(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)$$

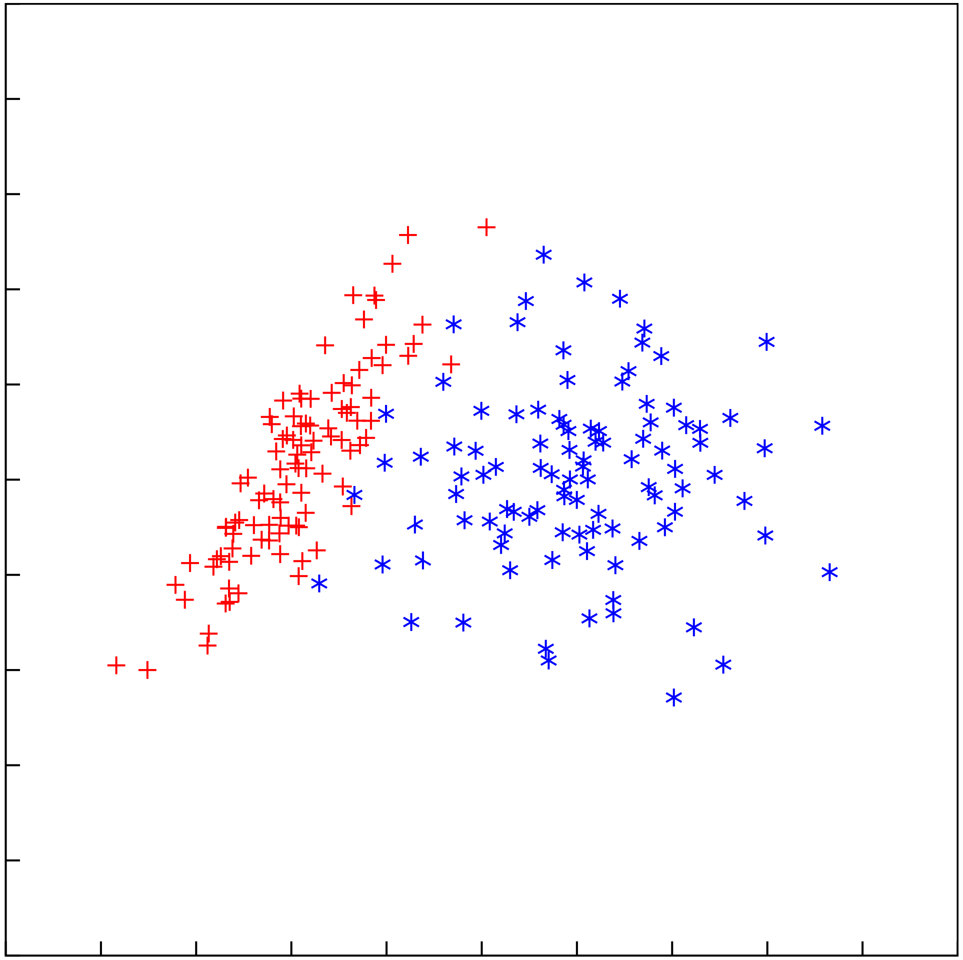
$$\equiv \quad p(\boldsymbol{x}; \boldsymbol{\Theta}); \quad \boldsymbol{\Theta} \stackrel{\text{def}}{=} [\alpha, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \Sigma_1, \Sigma_2]$$

Mezcla de gaussianas: ilustración en $\mathbb{R}^2$



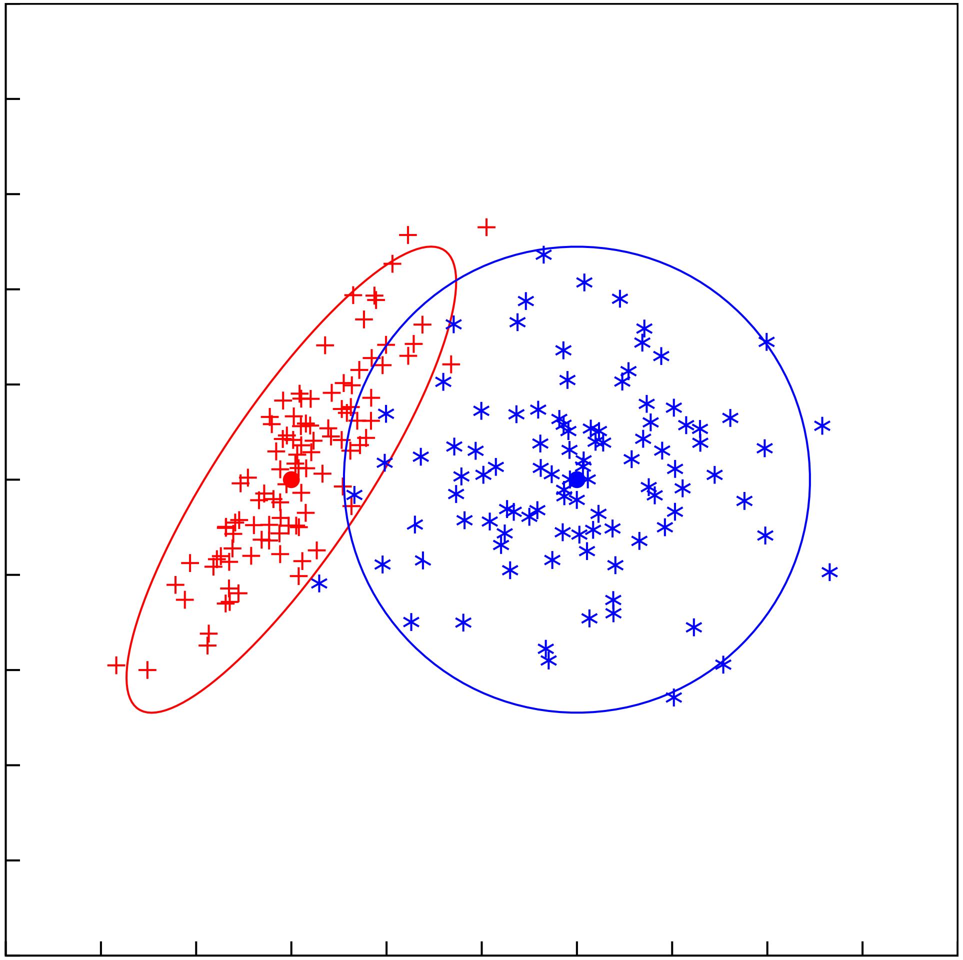
Datos *observables* generados por una mezcla de dos gaussianas.

Mezcla de gaussianas: ilustración en $\mathbb{R}^2$



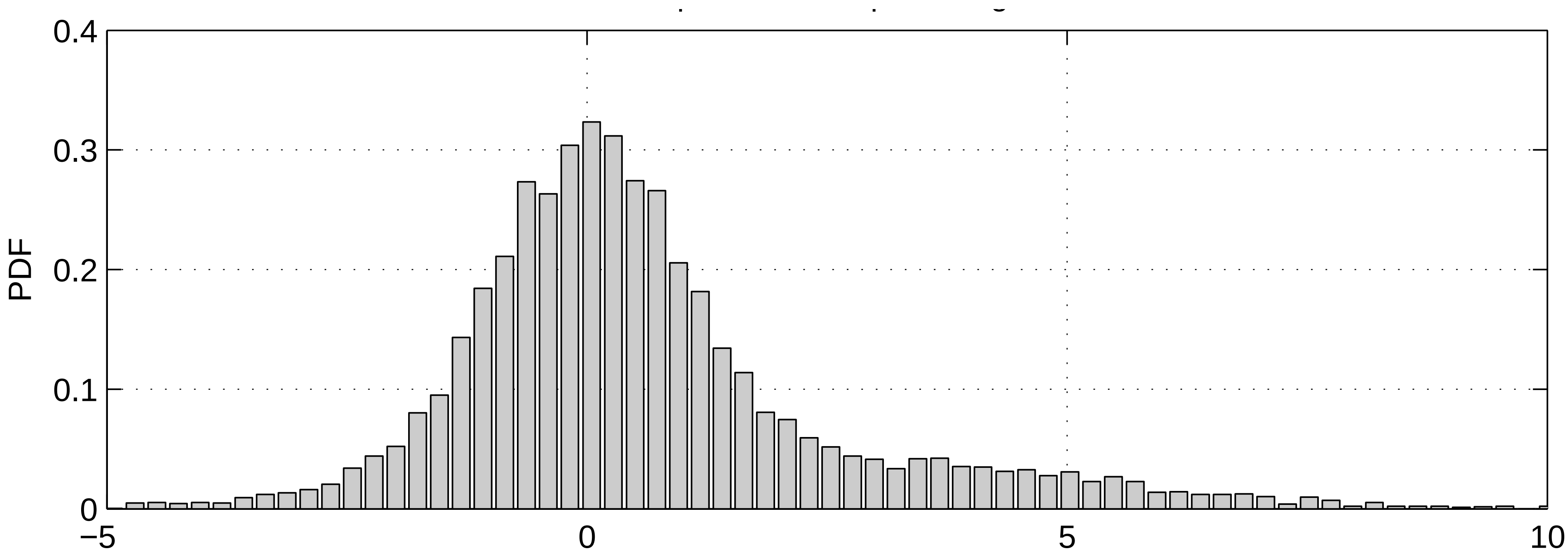
Datos de una mezcla de dos gaussianas con la variable oculta expuesta.

Mezcla de gaussianas: ilustración en $\mathbb{R}^2$



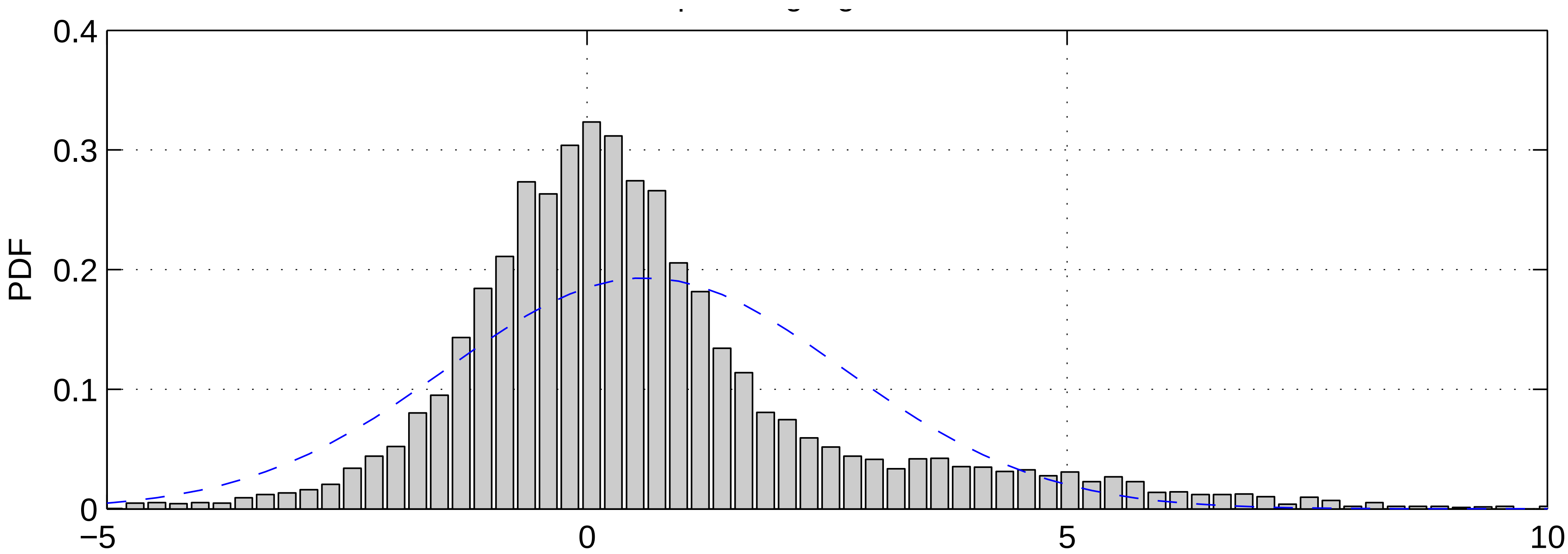
Datos de una mezcla de dos gaussianas con la variable oculta expuesta.
Las elipses muestran los parámetros de las gaussianas del modelo generador.

Mezcla de gaussianas: otro ejemplo en $\mathbb{R}$



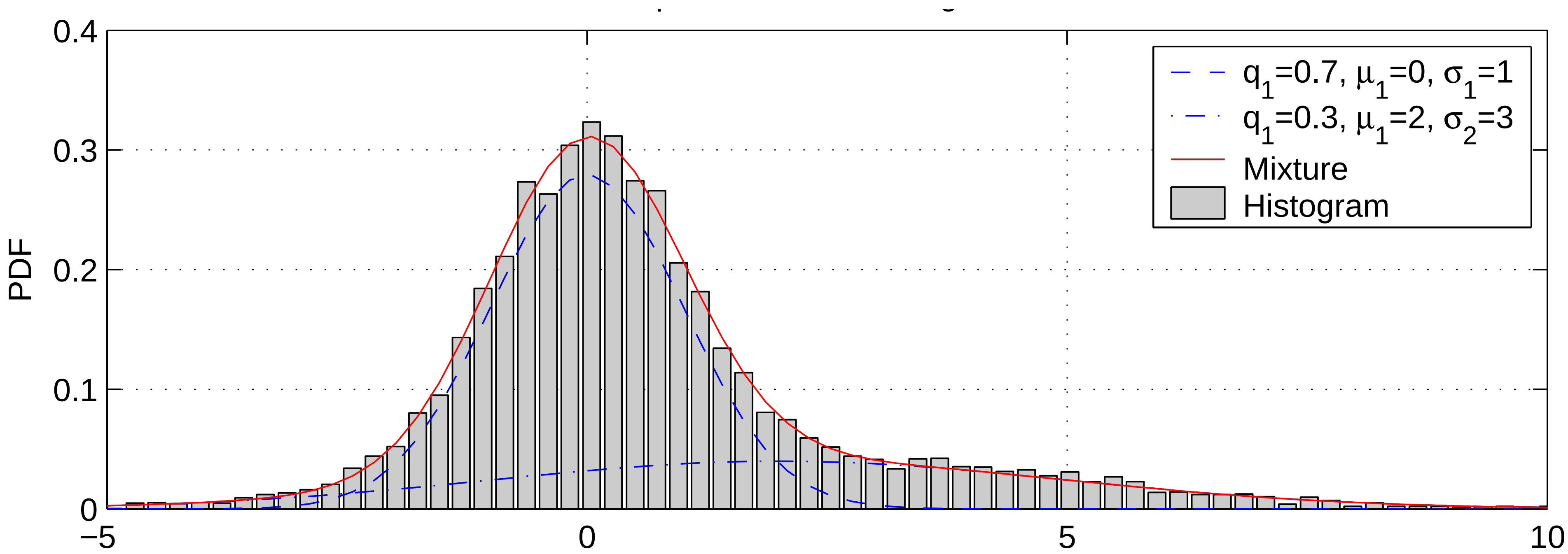
Histograma de una muestra de una variable unidimensional

Mezcla de gaussianas: otro ejemplo en $\mathbb{R}$



Una única gaussiana estimada a partir de la muestra

Mezcla de gaussianas: otro ejemplo en $\mathbb{R}$



Mezcla de dos gaussianas con la que se ha generado la muestra

Aprendizaje de modelos probabilísticos con variables latentes

- El logaritmo de la verosimilitud de la muestra $S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ con $\boldsymbol{x}_i \in \mathbb{R}^d$ para $1 \leq i \leq N$ y variables latentes $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$, es:

$$L_S(\boldsymbol{\Theta}) \;=\; \log \prod_{n=1}^{N} P(\boldsymbol{x}_n; \boldsymbol{\Theta}) \;=\; \sum_{n=1}^{N} \log \left(\sum_{\boldsymbol{z}_n} P(\boldsymbol{x}_n, \boldsymbol{z}_n; \boldsymbol{\Theta}) \right)$$

- Problema: Estimar de $\boldsymbol{\Theta}$ por máxima verosimilitud: $\boldsymbol{\Theta}^* \;=\; \arg\max_{\boldsymbol{\Theta}} L_S(\boldsymbol{\Theta})$

- Se define una ***función auxiliar*** $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}')$

$$Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') \;=\; \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} P(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \boldsymbol{\Theta}') \log P(\boldsymbol{x}_n, \boldsymbol{z}_n; \boldsymbol{\Theta})$$

- Teorema: $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}') \leq L_S(\boldsymbol{\Theta})$ para cualquier $\boldsymbol{\Theta}'$.

- Estimación de $\boldsymbol{\Theta}$: El algoritmo EM utilizando $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}')$.

Algoritmo esperanza-maximización (EM)

- Inicialización: $t = 1$, $\mathbf{\Theta}(1)$ arbitrario.

- Iteración hasta la convergencia

  – Paso E (esperanza de las variables ocultas). A partir de un conjunto de parámetros dado $\mathbf{\Theta}(t)$ y para todo $1 \leq n \leq N$,

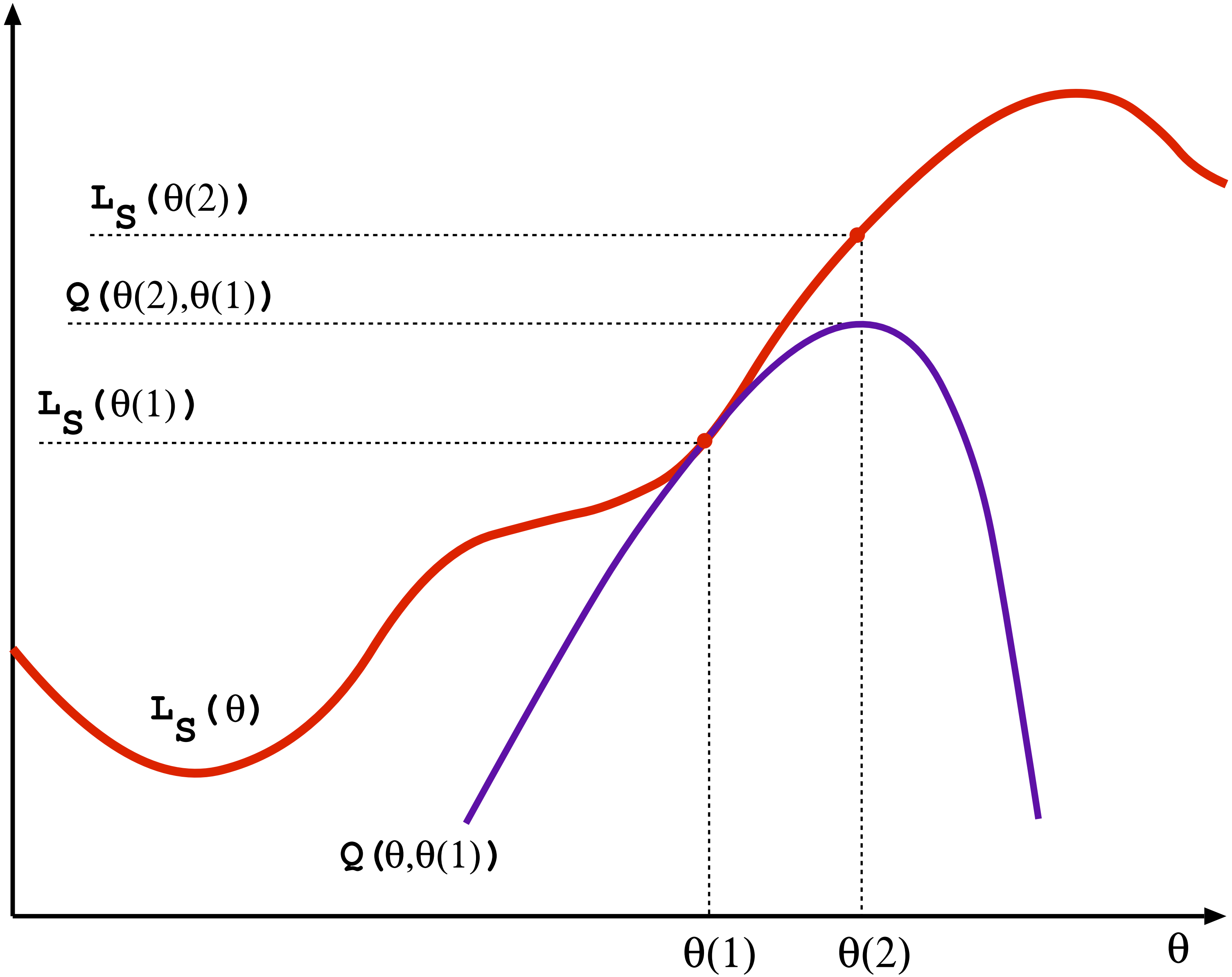
$$P(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \mathbf{\Theta}(t)) \;=\; \frac{P(\boldsymbol{x}_n \mid \boldsymbol{z}_n; \mathbf{\Theta}(t)) P(\boldsymbol{z}_n; \mathbf{\Theta}(t))}{P(\boldsymbol{x}_n)}$$

  – Paso M (maximización de la función auxiliar $Q(\mathbf{\Theta}, \mathbf{\Theta}(t))$ con respecto a $\mathbf{\Theta}$)

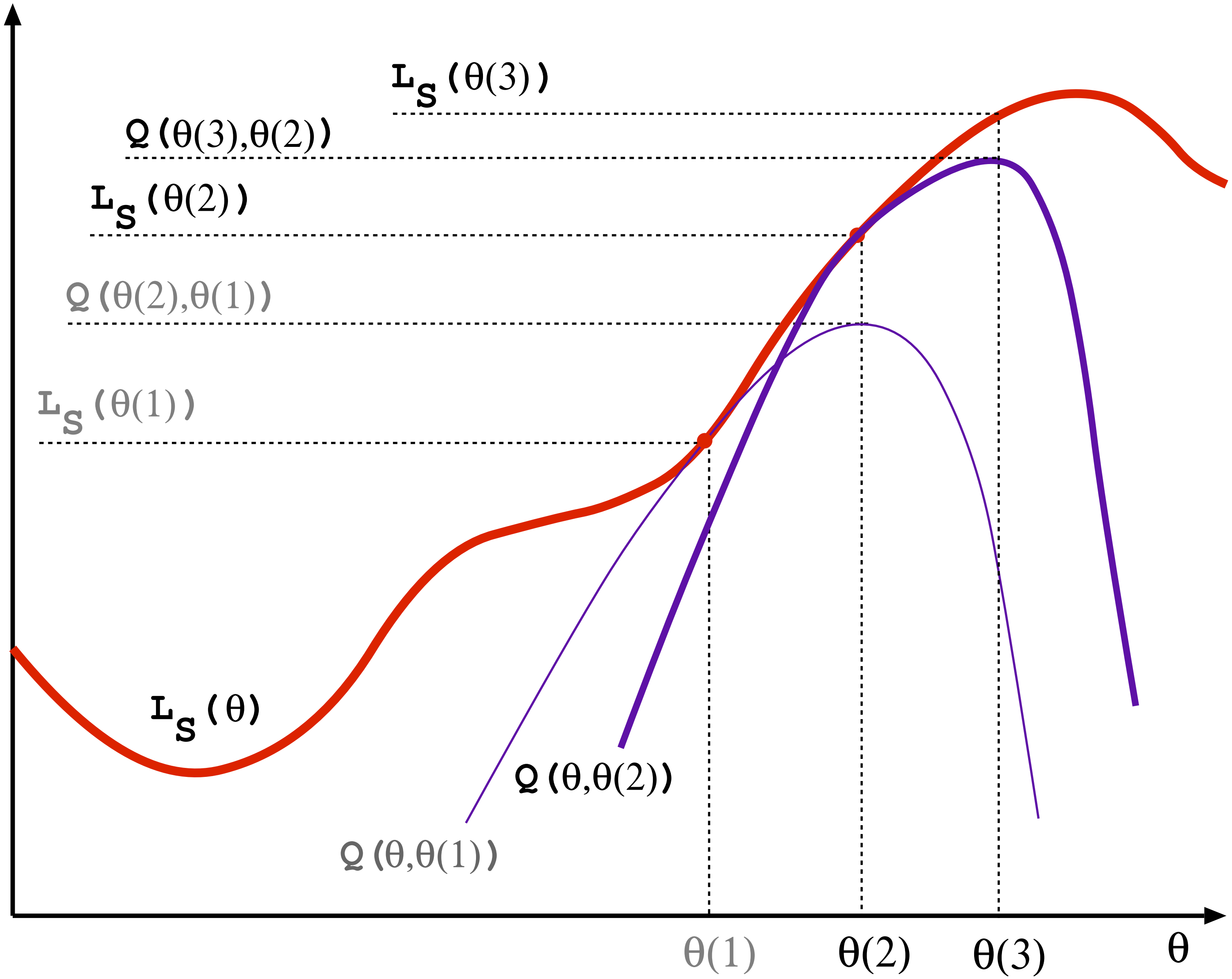
$$\begin{aligned}\mathbf{\Theta}(t+1) &= \arg\max_{\mathbf{\Theta}} Q(\mathbf{\Theta}, \mathbf{\Theta}(t)) \\ &= \arg\max_{\mathbf{\Theta}} \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} P(\boldsymbol{z}_n \mid \boldsymbol{x}_n; \mathbf{\Theta}(t)) \log P(\boldsymbol{x}_n, \boldsymbol{z}_n; \mathbf{\Theta})\end{aligned}$$

  – $t = t + 1$

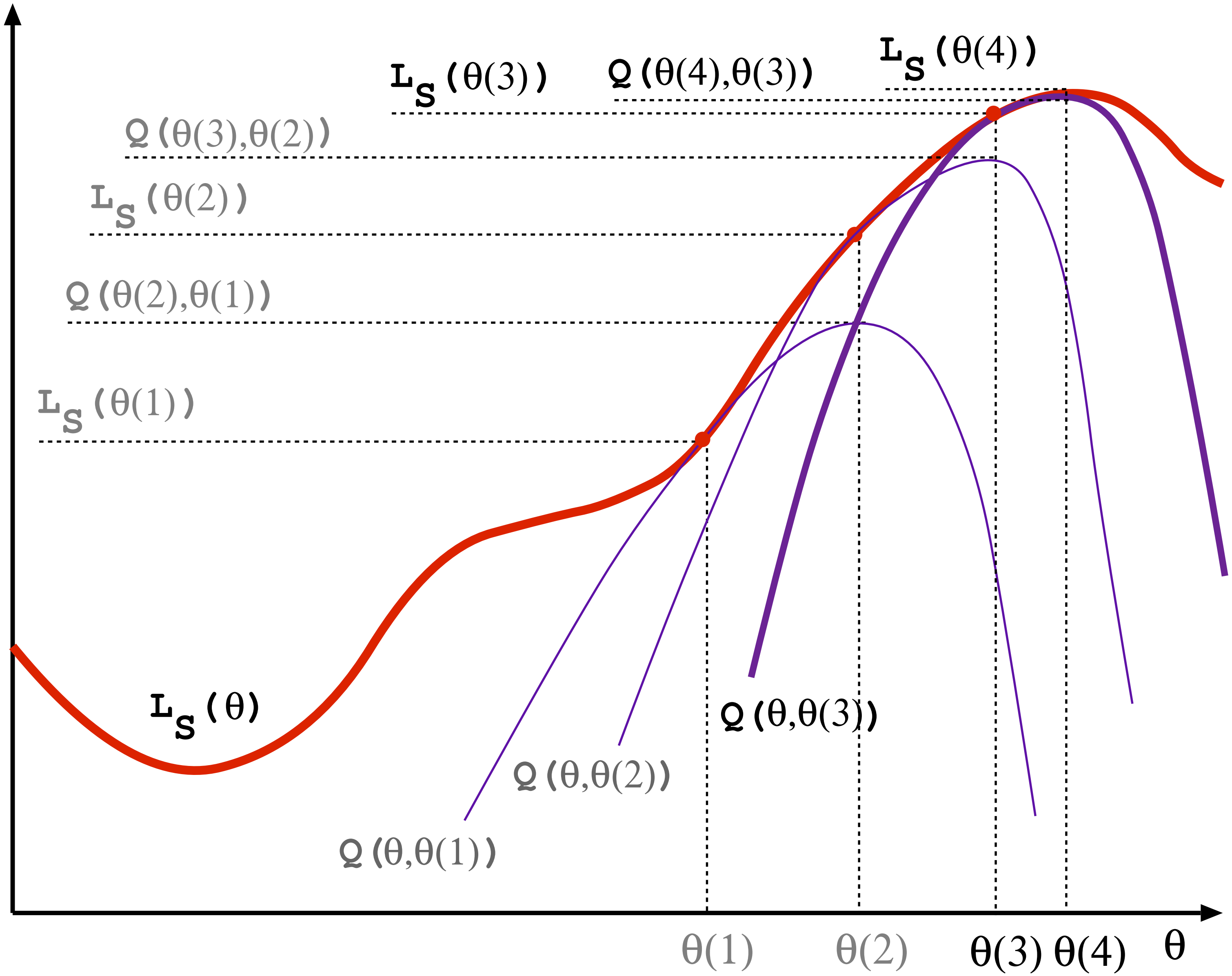
Propiedades y convergencia del EM



Propiedades y convergencia del EM



Propiedades y convergencia del EM



Ejemplo: mezcla de 2 gaussianas en $\mathbb{R}$

- Dada una muestra $S = \{x_1, \ldots, x_N\}$ con $x_i \in \mathbb{R}$ para $1 \leq i \leq N$, el problema es estimar $\boldsymbol{\Theta} \stackrel{\text{def}}{=} (\alpha, \mu_1, \mu_2)^t$ (suponemos que σ_1 y σ_2 están dados)

$$\begin{aligned}
(\alpha^*, \mu_1^*, \mu_2^*) &= \arg\max_{\mu_1,\mu_2,\alpha} L_S(\alpha, \mu_1, \mu_2) \\
&= \arg\max_{\mu_1,\mu_2,\alpha} \sum_{n=1}^{N} \log p(x_n; \mu_1, \mu_2, \alpha) \\
&= \arg\max_{\mu_1,\mu_2,\alpha} \sum_{n=1}^{N} \log \sum_{i=1}^{2} p(x_n, z_n = i; \mu_1, \mu_2, \alpha) \\
&= \arg\max_{\mu_1,\mu_2,\alpha} \sum_{n=1}^{N} \log \sum_{i=1}^{2} P(z_n = i; \alpha)\, p(x_n \mid z_n = i; \mu_i)
\end{aligned}$$

- Recordemos que,

$$P(z_n = 1; \alpha) = \alpha \quad ; \quad P(z_n = 2; \alpha) = (1 - \alpha)$$

$$p(x_n \mid z_n = i; \mu_i) = p(x_n; \mu_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right) \text{ para } i = 1, 2$$

$$\log p(x_n \mid z_n = i; \mu_i) = -\log \sqrt{2\pi}\sigma_i - \frac{(x - \mu_i)^2}{2\sigma_i^2}$$

Ejemplo: mezcla de 2 gaussianas en $\mathbb{R}$

- Apliquemos el paso E para $i=1,2$ en la iteración t para $1 \leq n \leq N$ e $i=1,2$:

$$P(z_n = i \mid x_n; \mathbf{\Theta}(t)) = \frac{P(z_n; \mathbf{\Theta}(t))\ p(x_n \mid z_n; \mathbf{\Theta}(t))}{\sum_{i'=1}^{2} P(z_n = i'; \mathbf{\Theta}(t))\ p(x_n \mid z_n = i'; \mathbf{\Theta}(t))}$$

- Sustituyendo:

$$P(z_n = 1 \mid x_n; \mathbf{\Theta}(t)) = \frac{\alpha(t)\ p(x_n; \mu_1(t))}{\alpha(t)\ p(x_n; \mu_1(t)) + (1-\alpha(t))\ p(x_n; \mu_2(t))}$$

$$P(z_n = 2 \mid x_n; \mathbf{\Theta}(t)) = \frac{(1-\alpha(t))\ p(x_n; \mu_2(t))}{\alpha(t)\ p(x_n; \mu_1(t)) + (1-\alpha(t))\ p(x_n; \mu_2(t))}$$

Ejemplo: mezcla de 2 gaussianas en $\mathbb{R}$

- Cálculo de $Q(\mathbf{\Theta}, \mathbf{\Theta}(t))$ en t que toma la forma de (ejercicio):

$$Q(\mathbf{\Theta}, \mathbf{\Theta}(t)) = \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} P(z_n \mid x_n; \mathbf{\Theta}(t)) \log p(x_n, z_n; \mathbf{\Theta})$$

$$= \sum_{n=1}^{N} \Bigg[P(z_n = 1 \mid x_n; \mathbf{\Theta}(t)) \left(\log \alpha - \log \sqrt{2\pi}\sigma_1 - \frac{(x_n - \mu_1)^2}{2\sigma_1^2} \right) + P(z_n = 2 \mid x_n; \mathbf{\Theta}(t)) \left(\log(1-\alpha) - \log \sqrt{2\pi}\sigma_2 - \frac{(x_n - \mu_2)^2}{2\sigma_2^2} \right) \Bigg]$$

- Apliquemos el paso M, maximizando $Q(\mathbf{\Theta}, \mathbf{\Theta}(t))$ con respecto a $\mathbf{\Theta} = (\mu_1, \mu_2, \alpha)^t$ (ejercicio)

Resolver $\quad \dfrac{\partial Q(\mathbf{\Theta}, \mathbf{\Theta}(t))}{\partial \alpha} = 0, \quad \dfrac{\partial Q(\mathbf{\Theta}, \mathbf{\Theta}(t))}{\partial \mu_1} = 0, \quad \dfrac{\partial Q(\mathbf{\Theta}, \mathbf{\Theta}(t))}{\partial \mu_2} = 0$

Algoritmo EM para la mezcla de 2 gausianas en $\mathbb{R}$

- Inicialización: $t = 1$, $\boldsymbol{\Theta}(1)$ arbitrarios ($\boldsymbol{\Theta}(1) = (\mu_1(1), \mu_2(1), \alpha(1))^t$).
- Iteración hasta la convergencia
  - Paso E (esperanza). A partir de $\boldsymbol{\Theta}(t) = (\mu_1(t), \mu_2(t), \alpha(t))^t$ y para $1 \leq n \leq N$:

$$\begin{aligned}
P(z_n = 1 \mid x_n; \boldsymbol{\Theta}(t)) &= \frac{\alpha(t)\, p(x_n; \mu_1(t))}{\alpha(t)\, p(x_n; \mu_1(t)) + (1 - \alpha(t))\, p(x_n; \mu_2(t))} \\
P(z_n = 2 \mid x_n; \boldsymbol{\Theta}(t)) &= \frac{(1 - \alpha(t))\, p(x_n; \mu_2(t))}{\alpha(t)\, p(x_n; \mu_1(t)) + (1 - \alpha(t))\, p(x_n; \mu_2(t))}
\end{aligned}$$

  - Paso M (maximización)

$$\begin{aligned}
\alpha(t+1) &= \frac{1}{N} \sum_{n=1}^{N} P(z_n = 1 \mid x_n; \boldsymbol{\Theta}(t)) \\
\mu_1(t+1) &= \frac{1}{N\, \alpha(t+1)} \sum_{n=1}^{N} P(z_n = 1 \mid x_n; \boldsymbol{\Theta}(t))\, x_n \\
\mu_2(t+1) &= \frac{1}{N\, (1 - \alpha(t+1))} \sum_{n=1}^{N} P(z_n = 2 \mid x_n; \boldsymbol{\Theta}(t))\, x_n
\end{aligned}$$

  - $t = t + 1$

Ejemplo: mezcla de gaussianas y modelo generador en $\mathbb{R}^D$

- En el caso de una única gaussiana las muestras se generan en un paso:
  1. Escoger $\boldsymbol{x}$, de acuerdo con la distribución $p(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma)$

- Si el modelo es una *mezcla* de K gaussianas, el proceso de generación se compone de dos etapas:
  1. De acuerdo con la distribución $P(k) = \alpha_k$, escoger la componente k-ésima de la mezcla con la que se va a generar $\boldsymbol{x}$
  2. Escoger $\boldsymbol{x}$, según la distribución definida por la k-ésima gaussiana, $p(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \Sigma_k)$

  - $\boldsymbol{x}$ es el dato observable y k es el valor de una variable oculta z. Los datos observables junto con los ocultos se denominan datos completos

  - Probabilidad con la que se genera $\boldsymbol{x}$ según este proceso:

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{k=1}^{K} p(z=k, \boldsymbol{x}) = \sum_{k=1}^{K} P(z=k)\, p(\boldsymbol{x} \mid k) \equiv \sum_{k=1}^{K} \alpha_k\, p(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k) \\
&\equiv p(\boldsymbol{x}; \boldsymbol{\Theta}); \quad \boldsymbol{\Theta} \stackrel{\text{def}}{=} (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \Sigma_1 \ldots, \Sigma_K, \alpha_1, \ldots, \alpha_K)^t
\end{aligned}
$$

Algoritmo EM para la mezcla de K gausianas

- $\mathbf{\Theta} \stackrel{\text{def}}{=} (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \alpha_1, \dots, \alpha_K)^t$, supondremos que $\Sigma_1 \dots, \Sigma_K$ están dados.
- Inicialización: $t = 1,\ \mathbf{\Theta}(1) = (\boldsymbol{\mu}_1(1), \dots, \boldsymbol{\mu}_K(1),\ \alpha_1(1), \dots, \alpha_K(1))^t$ arbitrarios.
- Iteración hasta la convergencia
  - Paso E (esperanza). A partir de $\mathbf{\Theta}(t)$ y para $1 \leq n \leq N$ y $1 \leq k \leq K$:

$$P(z_n = k \mid \boldsymbol{x}_n; \mathbf{\Theta}(t)) \;=\; \frac{\alpha_k(t)\ p(\boldsymbol{x}_n; \boldsymbol{\mu}_k(t))}{\sum_{k'} \alpha_{k'}(t)\ p(\boldsymbol{x}_n; \boldsymbol{\mu}_{k'}(t))}$$

  - Paso M (maximización). Para todo $1 \leq k \leq K$:

$$\alpha_k(t+1) \;=\; \frac{1}{N} \sum_{n=1}^{N} P(z_n = k \mid \boldsymbol{x}_n; \mathbf{\Theta}(t))$$

$$\boldsymbol{\mu}_k(t+1) \;=\; \frac{1}{N\ \alpha_k(t+1)} \sum_{n=1}^{N} P(z_n = k \mid \boldsymbol{x}_n; \mathbf{\Theta}(t))\ \boldsymbol{x}_n$$

  - $t = t + 1$

Index

Notación

- $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_D)^t$: vector de parámetros. Como los vectores son matrices columna, para representar las componentes en fila se usa la t (*"transpuesta"*).

- $q_S(\boldsymbol{\Theta})$: función objetivo a optimizar definida sobre un conjunto de entrenamiento S, cuyos de parámetros son $\boldsymbol{\Theta}$

- $\boldsymbol{\nabla} q(\boldsymbol{\Theta}) \stackrel{\text{def}}{=} \left(\frac{\partial q(\boldsymbol{\Theta})}{\partial \Theta_1}, \ldots, \frac{\partial q(\boldsymbol{\Theta})}{\partial \Theta_D}\right)^t$: gradiente de la función q_s

  (vector de derivadas parciales con respecto a cada componente de $\boldsymbol{\Theta}$)

- $\boldsymbol{\nabla} q(\boldsymbol{\Theta}) \mid_{\boldsymbol{\Theta}=\boldsymbol{\Theta}(k)} \equiv \left(\frac{\partial q}{\partial \Theta_1}\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}(k)}, \ldots, \frac{\partial q}{\partial \Theta_D}\Big|_{\boldsymbol{\Theta}=\boldsymbol{\Theta}(k)}\right)^t$: gradiente de la función q

  calculado en $\boldsymbol{\Theta}(k)$

- Σ : matriz de covarianzas de una distribución gaussiana