UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Aprendizaje Automático Probabilístico

Optimización

Alfons Juan

DSIC

Departamento de Sistemas Informáticos y Computación

Índice

8.0. Resumen

- ***Introducción:*** optimización continua, no discreta!
  - ***Dicotomía 1:*** local o global?
    - ↦ Global para problemas convexos. . . ahora son no convexos
    - ↦ Local es lo que se hace. . . condiciones de optimalidad
  - ***Dicotomía 2:*** con o sin restricciones?
    - ↦ Mejor sin restricciones (p.e. con ayuda de la softmax)
    - ↦ Las de igualdad requieren multiplicadores de Lagrange; las de desigualdad, si son pocas, quizás las podemos ignorar
  - ***Dicotomía 3:*** convexa o no?
    - ↦ Si un problema es convexo, un óptimo local es global!
  - ***Dicotomía 4:*** suave o no?
    - ↦ Suave si objetivo y restricciones son continuamente diferenciables. . . constante de Lipschitz
    - ↦ No suave “por poco”. . . subgradiente

- ***Métodos de primer orden:*** basados en derivadas de primer orden del objetivo... $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \, \boldsymbol{d}_t$
  - ***Dirección de descenso:*** $\boldsymbol{d}_t$ tal que $\mathcal{L}(\boldsymbol{\theta} + \eta \boldsymbol{d}_t) < \mathcal{L}(\boldsymbol{\theta})$
    - ***Descenso por gradiente:*** negativo del gradiente, $\boldsymbol{d}_t = -\boldsymbol{g}_t$
  - ***Tamaño de paso o factor de aprendizaje:*** $\{\eta_t\}$?
    - ***Constante:*** $\eta_t = \eta$ ... difícil de ajustar en la práctica
    - ***Búsqueda lineal:*** exacta o aproximada ... ***Armijo-Goldstein***
  - ***Ratios de convergencia:*** $\mu : |\mathcal{L}(\boldsymbol{\theta}_{t+1} - \mathcal{L}(\boldsymbol{\theta}_*))| \leq \mu |\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_*)|$
    - ***Objetivo cuadrático:*** $\mathbf{A} \succ 0$, $\mu = \left(\frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1} \right)^2$, $\kappa(\mathbf{A}) = \frac{\lambda_{\text{max}}(\mathbf{A})}{\lambda_{\text{min}}(\mathbf{A})}$
    - ***No cuadrático:*** $\approx$ cuadrático cerca de local $\rightarrow \kappa(\text{Hessiana})$
  - ***Momentum:*** heurístico para acelerar la convergencia
    - ***Estándar:*** $\boldsymbol{m}_t = \beta \, \boldsymbol{m}_{t-1} + \boldsymbol{g}_{t-1}$ y $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \, \boldsymbol{m}_t, \;\; \beta < 1$
      - ***EWMA:*** $\boldsymbol{m}_t = \sum_{\tau=0}^{t-1} \beta^\tau \boldsymbol{g}_{t-\tau-1} \overset{\boldsymbol{g}_{t-\tau-1} = \boldsymbol{g}}{=} \boldsymbol{g} \sum_{\tau=0}^{t-1} \beta^\tau \overset{t \to \infty}{=} \frac{\boldsymbol{g}}{1-\beta}$
    - ***Nesterov:*** extrapola $\boldsymbol{\theta}_{t+1}$ para amortiguar oscilaciones

- ***Métodos de segundo orden:*** añaden la Hessiana de $\mathcal{L}$ o aprox.
  - ***Método de Newton:*** $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \boldsymbol{g}_t$
    - ↦ Primero halla $\boldsymbol{d}_t$ tal que $\mathbf{H}_t \boldsymbol{d}_t = -\boldsymbol{g}_t$
    - ↦ Luego $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{d}_t$ con η_t hallado por búsqueda lineal
  - ***Métodos quasi-Newton:*** aproximan $\mathbf{H}_t$ con $\mathbf{B}_t$, obtenida iterativamente a partir de los gradientes hallados en cada paso
    - ↦ ***BFGS (Broyden–Fletcher–Goldfarb–Shanno):*** aplica actualizaciones sucesivas de rango dos . . . ***Wolfe;*** alt $\mathbf{C}_t \approx \mathbf{H}^{-1}$
    - ↦ ***Limited memory BFGS (L-BFGS):*** aproxima $\mathbf{H}_t^{-1} \boldsymbol{g}_t$ con las M actualizaciones más recientes
  - ***Métodos en regiones de confianza:*** no fijan $\boldsymbol{d}_t$ y luego η_t, sino al revés; aproximan $\mathcal{L}$ alrededor de $\boldsymbol{\theta}_t$ y buscan una dirección óptima. . . ***regularización de Tikhonov***

- ***Descenso por gradiente estocástico:*** descenso por gradiente aplicado a ***optimización estocástica,*** $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})]$

$$\boldsymbol{\theta}_{t+1} \overset{\boldsymbol{z}_t \sim q}{=} \boldsymbol{\theta}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{z}_t) \overset{q(\boldsymbol{z}) \text{ indep } \boldsymbol{\theta}}{=} \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_t$$

  - ***Aplicación a problemas de sumas finitas:*** en minimización del riesgo empírico con N muestras, observamos un ***minibatch*** de $B \ll N$ muestras en la iteración t
  - ***Ejemplo: SGD para ajustar regresión lineal:*** se conoce como ***mínimos cuadrados***, ***regla delta*** o ***Widrow-Hoff***
  - ***Elección del tamaño de paso: Robbins-Monro***
  - ***Promediado iterativo:*** EWMA para reducir la varianza
  - ***SGD precondicionado:*** $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{M}_t^{-1} \boldsymbol{g}_t$ con ***precondicionador*** diagonal $\mathbf{M}_t$ **ADAGRAD, RMSPROP, ADADELTA *o* ADAM**

- ***Optimización con restricciones:*** de igualdad y desigualdad
  - ***Multiplicadores de Lagrange:*** solo restricciones de igualdad
  - ***Las condiciones KKT:*** para el caso general . . . ***estacionaridad, factibilidad primal y dual, y holgura complementaria***
  - ***Programación lineal:*** objetivo lineal con restricciones lineales
  - ***Programación cuadrática:*** objetivo cuadrático con restricciones lineales

- ***Optimización acotada:*** basada en una cota inferior del objetivo

  - ***Algoritmo majorize-minorize (MM):*** basado en una ***función sustituta,*** cota inferior $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) \leq \mathrm{LL}(\boldsymbol{\theta})$ que toca el objetivo en $\boldsymbol{\theta}^t$, $Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \mathrm{LL}(\boldsymbol{\theta}^t)$: $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$

  - ***Algoritmo expectation maximization (EM):*** algoritmo MM para calcular el MLE o MAP de modelos con datos perdidos

    $$\mathrm{LL}(\boldsymbol{\theta}) = \sum_{n=1}^N \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta}) = \sum_{n=1}^N \log \left[\sum_{\boldsymbol{z}_n} p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})\right]$$

    - ↦ ***Evidence lower bound (ELBO):*** $Ł(\boldsymbol{\theta}, q_{1:N}) \leq \mathrm{LL}(\boldsymbol{\theta})$ (Jensen)
    - ↦ ***Paso E:*** cálculo de $q_n^* = p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta}) \rightarrow Ł(\boldsymbol{\theta}, q_n^*) = \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta})$
    - ↦ ***Paso M:*** maximización de la log-verosimilitud completa esperada, $\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \sum_n \mathbb{E}_{q_n^t(\boldsymbol{z}_n)}[\log p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})]$
    - ↦ ***Aplicación a mixturas de Gaussianas***
    - ↦ ***Estimación MAP:*** para mixturas de Gaussianas (robusta)
    - ↦ ***Noconvexidad de la NLL:*** label switching problem

- ***Optimización sin derivadas y caja-negra:*** optimización mediante búsqueda en rejilla para selección de modelos
  - Para exploración de hiperparámetros y cuesta “horrores”

8.2. Métodos de primer orden

- Los ***métodos de primer orden*** son métodos iterativos basados en derivadas de primer orden del objetivo

- Dado un punto de inicio $\boldsymbol{\theta}_0$, la iteración t consiste en hacer:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \, \boldsymbol{d}_t \tag{22}$$

- η_t es el ***tamaño del paso (step size)*** o ***factor de aprendizaje (learning rate)***

- $\boldsymbol{d}_t$ es la ***dirección de descenso***, como el negativo del ***gradiente,*** $\boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_t}$

- Se termina al alcanzar un punto estacionario, de gradiente nulo

8.2.1. Dirección de descenso

- $\boldsymbol{d}$ es una ***dirección de descenso*** si existe un $\eta_{\text{max}} > 0$ tal que

$$\mathcal{L}(\boldsymbol{\theta} + \eta \boldsymbol{d}) < \mathcal{L}(\boldsymbol{\theta}) \qquad \text{para todo } 0 < \eta < \eta_{\text{max}} \tag{23}$$

- La dirección de máximo ascenso en f es la del gradiente actual:

$$\boldsymbol{g}_t \triangleq \nabla \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_t} = \mathcal{L}(\boldsymbol{\theta}_t) = \boldsymbol{g}(\boldsymbol{\theta}_t) \tag{24}$$

- $\boldsymbol{d}$ es dirección de descenso si el ángulo θ entre $\boldsymbol{d}$ y $-\boldsymbol{g}_t$ es menor de 90 grados y satisface:

$$\boldsymbol{d}^t \boldsymbol{g}_t = \|\boldsymbol{d}\| \|\boldsymbol{g}_t\| \cos(\theta) < 0 \tag{25}$$

- ***Descenso por gradiente (gradient descent)*** o ***más pronunciado (steepest descent):*** escogemos el negativo del gradiente

$$\boldsymbol{d}_t = -\boldsymbol{g}_t \tag{26}$$

8.2.2. Tamaño de paso o factor de aprendizaje

- ***Learning rate schedule:*** secuencia de tamaños de paso $\{\eta_t\}$

8.2.2.1. Tamaño de paso constante

- La opción más simple consiste en usar un learning rate constante

$$\eta_t = \eta \tag{27}$$

- Si η es demasiado grande, el método puede no converger
- Si η es demasiado pequeño, el método covergerá muy lentamente
- *Ejemplo:* $\mathcal{L}(\boldsymbol{\theta}) = 0.5(\theta_1^2 - \theta_2)^2 + 0.5(\theta_1 - 1)^2$
  - Con $\eta = 0.1$ converge lentamente
  - Con $\eta = 0.6$ oscila y no converge

8.2.2.2. Búsqueda lineal

- ***Búsqueda lineal*** consiste en hallar el tamaño de paso óptimo en la dirección escogida mediante optimización:

$$\eta_t = \arg\min_{\eta>0} \phi_t(\eta) \tag{31}$$

$$= \arg\min_{\eta>0} \mathcal{L}(\boldsymbol{\theta}_t + \eta\, \boldsymbol{d}_t) \tag{32}$$

- ***Búsqueda lineal exacta*** consiste en resolver analíticamente la optimización anterior, si se puede
  - En particular, si $\mathcal{L}$ es convexa, ϕ también lo es y sí se puede

8.2.3. Ratios de convergencia

- Queremos algoritmos que converjan rápidamente a un óptimo
- Descenso por gradiente converge con ***ratio lineal*** en problemas convexos con gradiente acotado por una constante de Lipschitz
- ***Ratio de convergencia:*** $\mu \in (0, 1)$ tal que

$$|\mathcal{L}(\boldsymbol{\theta}_{t+1} - \mathcal{L}(\boldsymbol{\theta}_*))| \leq \mu |\mathcal{L}(\boldsymbol{\theta}_t) - \mathcal{L}(\boldsymbol{\theta}_*)| \tag{39}$$

- El ratio puede derivarse explícitamente en algunos problemas

▸ *Ratio de convergencia de un objetivo cuadrático:*

▹ Consideremos un objetivo cuadrático

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^t \mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}^t\boldsymbol{\theta} + c \quad \text{con} \quad \mathbf{A} \succ 0 \tag{40}$$

▹ Aplicamos descenso por gradiente con búsqueda lineal exacta

▹ Se puede ver que el ratio de convergencia es

$$\mu = \left(\frac{\lambda_{\text{max}}(\mathbf{A}) - \lambda_{\text{min}}(\mathbf{A})}{\lambda_{\text{max}}(\mathbf{A}) + \lambda_{\text{min}}(\mathbf{A})}\right)^2 = \left(\frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1}\right)^2 \tag{41}$$

donde el número de condición de $\mathbf{A}$,

$$\kappa(\mathbf{A}) = \frac{\lambda_{\text{max}}(\mathbf{A})}{\lambda_{\text{min}}(\mathbf{A})} \tag{42}$$

mide la curvatura del objetivo (respecto a un bol simétrico)

8.7. Optimización acotada

- ***Optimización acotada o MM (majorize-minimize):*** clase de algoritmos de optimización de gran interés en ML
  - ***Algoritmo expectation-maximization (EM):*** caso especial de algoritmo MM muy usado en ML

8.7.1. El algoritmo general

- ***Objetivo:*** maximizar $\mathrm{LL}(\boldsymbol{\theta})$

- ***Función sustituta (surrogate):*** construimos una función cota inferior del objetivo, $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$, que lo iguala en un $\boldsymbol{\theta}^t$ dado:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) \leq \mathrm{LL}(\boldsymbol{\theta}) \quad \text{y} \quad Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \mathrm{LL}(\boldsymbol{\theta}^t) \tag{137}$$

  - Si se cumplen ambas condiciones, decimos que ***minoriza*** LL

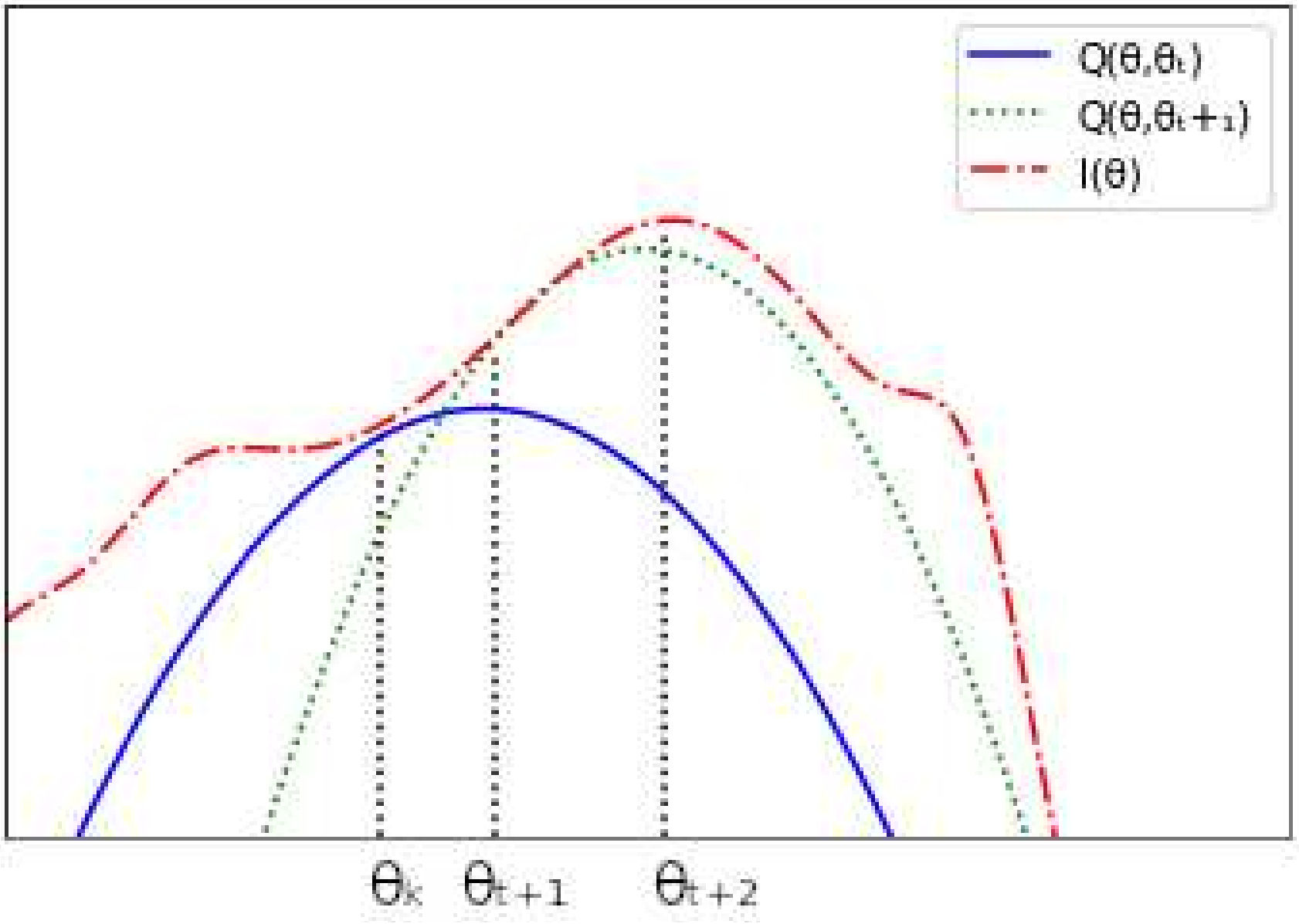
- ***Algoritmo majorize-minimize (MM):*** para $t = 0, 1, \ldots$

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \; Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) \tag{138}$$

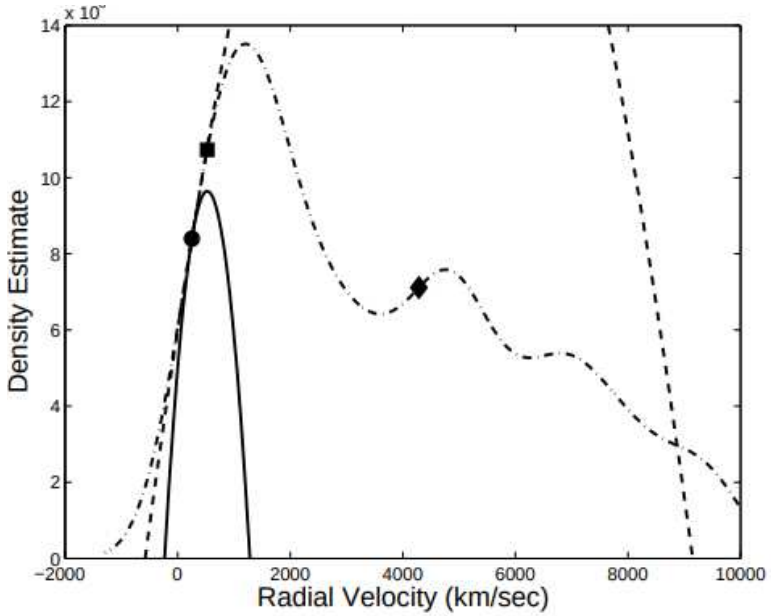
  - Si $\boldsymbol{\theta}^{t+1}$ se escoge tal que $Q(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$:

$$\mathrm{LL}(\boldsymbol{\theta}^{t+1}) \geq Q(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \mathrm{LL}(\boldsymbol{\theta}^t) \tag{139}$$

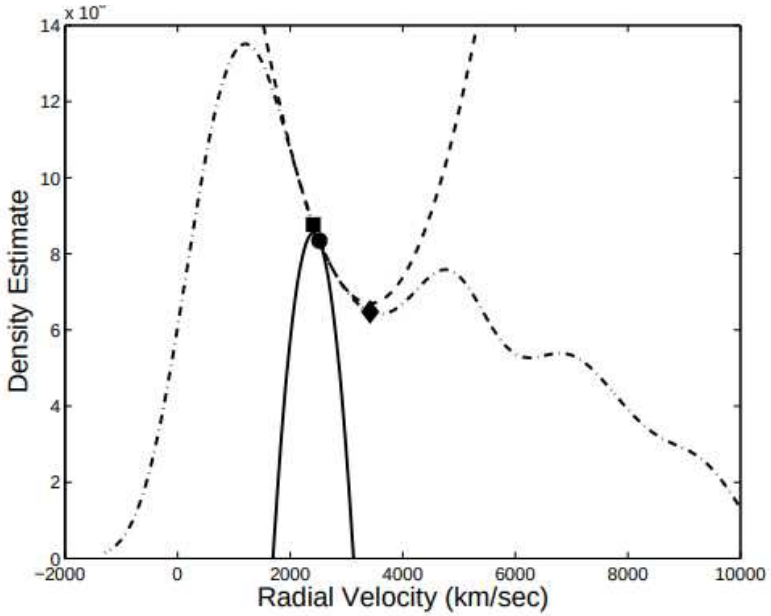
- *Ejemplo:* $Q(\theta, \theta^t)$ toca $\mathrm{LL}(\theta)$ en θ^t; su maximización da lugar a θ^{t+1} y $Q(\theta, \theta^{t+1})$ toca $\mathrm{LL}(\theta)$ en θ^{t+1}; su maximización da lugar a θ^{t+2}, etc.



- ***Similitud con el método de Newton:*** si Q es una cota inferior cuadrática, el MM se asemeja al método de Newton, pues ajusta y optimiza una aproximación cuadrática del objetivo repetidamente
  - ***Diferencia:*** MM garantiza un mejora del objetivo en cada iteración, incluso si no es convexo, pero Newton no
  - *Ejemplo:* a la izquierda Newton se "pasa de largo" buscando un máximo y a la derecha se va a un mínimo



(a) Overshooting.



(b) Seeking the wrong root.

8.7.2. El algoritmo EM

- ***Algoritmo expectation maximization (EM):*** algoritmo de optimización acotada para calcular el estimador MLE o MAP de modelos probabilísticos con ***datos perdidos*** o ***variables ocultas***

  - ***Notación:*** para cada dato n, $\boldsymbol{y}_n$ denota su parte observada y $\boldsymbol{z}_n$ su parte perdida u oculta

  - ***Algoritmo EM básico:*** repetir los siguientes dos pasos

    - ***Paso E (expectation):*** estimación de datos perdidos

    - ***Paso M (maximization):*** cálculo del MLE o MAP a partir de los datos completos

  - ***Convergencia:*** veremos que el paso E calcula una función sustituta del objetivo, por lo que el EM converge a un màximo local

8.7.2.1. Cota inferior

- ***Objetivo:*** maximizar la log-verosimilitud de los datos observados

$$\mathrm{LL}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta}) \tag{140}$$

$$= \sum_{n=1}^{N} \log \left[\sum_{\boldsymbol{z}_n} p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta}) \right] \tag{141}$$

  - ***Dificultad:*** difícil de optimizar a causa del logaritmo delante del sumatorio

- ***Evidence lower bound (ELBO):*** dado un conjunto de distribuciones arbitrarias sobre cada $\boldsymbol{z}_n$, $q_n(\boldsymbol{z}_n)$, la ***desigualdad de Jensen*** (sección 6.2.4) permite construir una función $\text{Ł}(\boldsymbol{\theta}, q_{1:N})$ cota inferior de la log-verosimilitud marginal o evidencia:

$$\text{LL}(\boldsymbol{\theta}) = \log p(\boldsymbol{y}_{1:N} \mid \boldsymbol{\theta}) \tag{142}$$

$$= \sum_{n=1}^{N} \log \left[\sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \frac{p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \right] \tag{143}$$

$$\geq \sum_{n=1}^{N} \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \tag{144}$$

$$= \sum_{n} \underbrace{\mathbb{E}_{q_n}[\log p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})] + \mathbb{H}(q_n)}_{\text{Ł}(\boldsymbol{\theta}, q_n \mid \boldsymbol{y}_n)} \tag{145}$$

$$= \sum_{n} \text{Ł}(\boldsymbol{\theta}, q_n) \triangleq \text{Ł}(\boldsymbol{\theta}, \{q_n\}) = \text{Ł}(\boldsymbol{\theta}, q_{1:N}) \tag{146}$$

8.7.2.2. Paso E

- ***Paso E:*** cálculo de $q_n^* = p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta})$ y, así, $\text{Ł}(\boldsymbol{\theta}, q_n^*) = \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta})$

$$\text{Ł}(\boldsymbol{\theta}, q_n) = \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \tag{147}$$

$$= \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta}) p(\boldsymbol{y}_n \mid \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} \tag{148}$$

$$= \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log \frac{p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta})}{q_n(\boldsymbol{z}_n)} + \sum_{\boldsymbol{z}_n} q_n(\boldsymbol{z}_n) \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta}) \tag{149}$$

$$= -\,\mathbb{KL}(q_n(\boldsymbol{z}_n) \,\|\, p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta})) + \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta}) \tag{150}$$

$$\overset{q_n = q_n^*}{=} \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta}) \tag{151}$$

pues $\mathbb{KL}(q_n(\boldsymbol{z}_n) \,\|\, p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta})) = 0$ sii $q_n \triangleq q_n^* = p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta})$

- ***Función sustituta:*** dado que $Ł(\boldsymbol{\theta}, \{q_n^*\}) = \mathrm{LL}(\boldsymbol{\theta})$, la función

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = Ł(\boldsymbol{\theta}, \{q_n^* = p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta}^t)\}) \tag{152}$$

  es ELBO por Jensen,

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) \leq \mathrm{LL}(\boldsymbol{\theta}) \tag{153}$$

  y toca $\mathrm{LL}(\boldsymbol{\theta})$ en $\boldsymbol{\theta}^t$ al tomar $\{q_n^* = p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta}^t)\}$,

$$Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t) = \mathrm{LL}(\boldsymbol{\theta}^t) \tag{154}$$

- ***Paso E aproximado:*** si el cálculo de $q_n^* = p(\boldsymbol{z}_n \mid \boldsymbol{y}_n, \boldsymbol{\theta})$ es muy costoso, podemos emplear una aproximación a la misma y la Q, aunque menos ajustada, sigue siendo ELBO
  - ***Aproximación directa:*** comprobamos que la LL no decrece; cosa quizás sencilla si solo consideramos distribuciones delta
  - ***EM variacional:*** EM generalizado en marco Bayesiano

8.7.2.3. Paso M

- ***Expected complete data log likelihood:*** el paso M maximiza $Ł(\boldsymbol{\theta}, \{q_n^t\})$ con respecto a $\boldsymbol{\theta}$, donde las $\{q_n^t\}$ son las distribuciones halladas en el paso E de la iteración t; ahora bien, como los términos de entropía $\mathbb{H}(q_n)$ no dependen de $\boldsymbol{\theta}$, podemos ignorarlos,

$$\mathrm{LL}^t(\boldsymbol{\theta}) = \sum_n \mathbb{E}_{q_n^t(\boldsymbol{z}_n)}[\log p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})] \tag{155}$$

  - ***Caso familia exponencial:*** si la probabilidad conjunta pertenece a la familia exponencial, no necesitamos $\{q_n^t\}$; bastan estadísticos suficientes esperados, $\mathbb{E}[\mathcal{T}(\boldsymbol{y}_n, \boldsymbol{z}_n)]$,

$$\mathrm{LL}^t(\boldsymbol{\theta}) = \sum_n \mathbb{E}[\mathcal{T}(\boldsymbol{y}_n, \boldsymbol{z}_n)^t \boldsymbol{\theta} - A(\boldsymbol{\theta})] \tag{156}$$

$$= \sum_n (\mathbb{E}[\mathcal{T}(\boldsymbol{y}_n, \boldsymbol{z}_n)]^t - A(\boldsymbol{\theta})) \tag{157}$$

- ***Paso M:*** maximización de la log-verosimilitud completa esperada

$$\boldsymbol{\theta}^{t+1} = \arg\max_{\boldsymbol{\theta}} \sum_n \mathbb{E}_{q_n^t(\boldsymbol{z}_n)}[\log p(\boldsymbol{y}_n, \boldsymbol{z}_n \mid \boldsymbol{\theta})] \tag{158}$$

  - ***Caso familia exponencial:*** se resuelve en forma cerrada

8.7.3. Ejemplo: EM para un GMM

8.7.3.1. Paso E

- ***Responsability:*** el paso E calcula la ***responsabilidad*** del clúster k en la generación del dato n, según la estimación actual de los parámetros $\boldsymbol{\theta}^{(t)}$,

$$r_{nk}^{(t)} = p^*(z_n = k \mid \boldsymbol{y}_n, \boldsymbol{\theta}^{(t)}) \tag{159}$$

$$= \frac{\pi_k^{(t)}\, p(\boldsymbol{y}_n \mid \boldsymbol{\theta}_k^{(t)})}{\sum_{k'} \pi_{k'}^{(t)}\, p(\boldsymbol{y}_n \mid \boldsymbol{\theta}_{k'}^{(t)})} \tag{160}$$

8.7.3.2. Paso M

- ***Log-verosimilitud completa esperada:*** versión ponderada de la LL para la Gaussiana multivariada; sea $z_{nk} = \mathbb{1}(z_n = k)$,

$$\mathrm{LL}^t(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_n \log p(z_n \mid \boldsymbol{\pi}) + \log p(\boldsymbol{y}_n \mid z_n, \boldsymbol{\theta})\right] \tag{161}$$

$$= \mathbb{E}\left[\sum_n \log\left(\prod_k \pi_k^{z_{nk}}\right) + \log\left(\prod_k \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}\right)\right] \tag{162}$$

$$= \sum_n \sum_k \mathbb{E}[z_{nk}] \log \pi_k + \sum_n \sum_k \mathbb{E}[z_{nk}] \log \mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{163}$$

$$= \sum_n \sum_k r_{nk}^{(t)} \log(\pi_k)$$

$$-\frac{1}{2}\sum_n \sum_k r_{nk}^{(t)} \left[\log|\boldsymbol{\Sigma}_k| + (\boldsymbol{y}_n - \boldsymbol{\mu}_k)^t \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}_k)\right] + \text{const} \tag{164}$$

- ***Solución cerrada:*** sea $r_k^{(t)} \triangleq \sum_n r_{nk}(t)$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{1}{r_k^{(t)}} \sum_n r_{nk}(t) \boldsymbol{y}_n \tag{165}$$

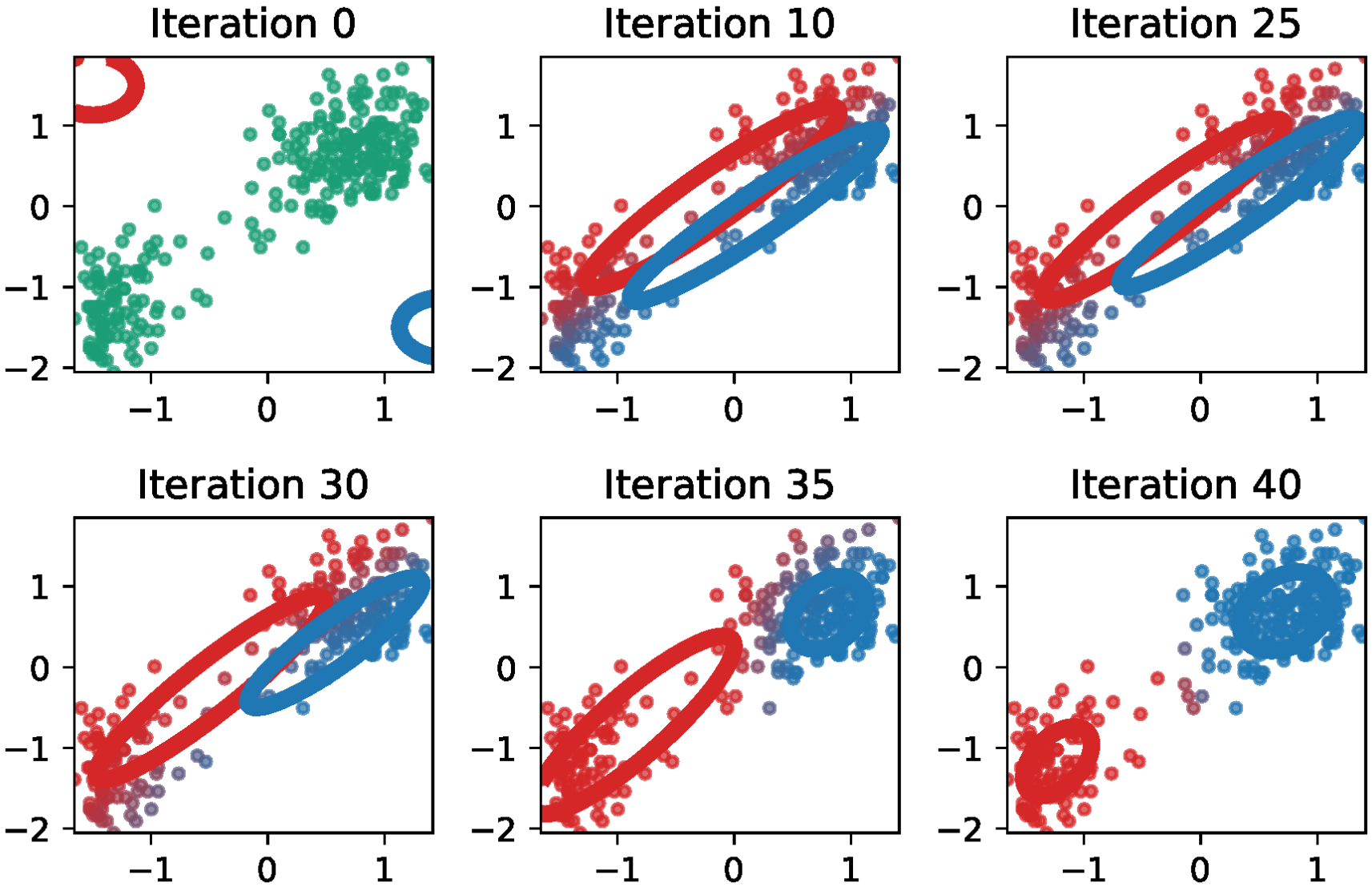
$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{r_k^{(t)}} \sum_n r_{nk}(t) (\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(t+1)})^t \tag{166}$$

$$= \frac{1}{r_k^{(t)}} \left(\sum_n r_{nk}(t) \boldsymbol{y}_n \boldsymbol{y}_n^t \right) - \boldsymbol{\mu}_k^{(t+1)} (\boldsymbol{\mu}_k^{(t+1)})^t \tag{167}$$

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_n r_{nk}(t) = \frac{r_k^{(t)}}{N} \tag{168}$$

8.7.3.3. Ejemplo

- ***Old Faithful:*** GMM ajustado con el EM a datos 2d del géiser Old Faithful (minutos siguiente erupción vs duración; estandarizados)



8.7.3.4. Estimación MAP

- ***Problema del colapso de la varianza:*** si $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$ y $\boldsymbol{\mu}_k$ se asigna a un único punto, $\boldsymbol{y}_n$, su verosilimilitud diverge con $\sigma_k \to 0$

$$\mathcal{N}(\boldsymbol{y}_n \mid \boldsymbol{\mu}_k = \boldsymbol{y}_n, \sigma_k^2 \mathbf{I}) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^0 \tag{169}$$

- ***Estimación MAP:*** maximiza la log-verosimilitud completa esperada más un log-prior

$$\begin{aligned} \mathrm{LL}^t(\boldsymbol{\theta}) = & \left[\sum_n \sum_k r_{nk}^{(t)} \log \pi_k + \sum_n \sum_k r_{nk}^{(t)} \log p(\boldsymbol{y}_n \mid \boldsymbol{\theta}_k) \right] \\ & + \log p(\boldsymbol{\pi}) + \sum_k \log p(\boldsymbol{\theta}_k) \end{aligned} \tag{170}$$

- ***Algoritmo EM:***
  - ***Paso E:*** igual que para el MLE
  - ***Paso M para los coeficientes:*** con prior ***Dirichlet,*** $\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha})$, conjugada de la categórica,

$$\tilde{\pi}_k^{(t+1)} = \frac{r_k^{(t)} + \alpha_k - 1}{N + \sum_k \alpha_k - K} \tag{171}$$

    $\mapsto$ Coincide con el MLE con un prior uniforme, $\alpha_k = 1$

- ***Paso M para las componentes:*** con prior ***Normal-Inverse-Wishart,*** conjugada de la Gaussiana multivariada,

$$p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \mathrm{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k \mid \breve{\boldsymbol{m}}, \breve{\kappa}, \breve{\nu}, \breve{\mathbf{S}}) \tag{172}$$

  - Con $\breve{\kappa} = 0$, las $\boldsymbol{\mu}_k$ no se regularizan, por lo que el prior solo afecta a las $\boldsymbol{\Sigma}_k$ y los estimadores MAP son:

$$\tilde{\boldsymbol{\mu}}_k^{(t+1)} = \hat{\boldsymbol{\mu}}_k^{(t+1)} \tag{173}$$

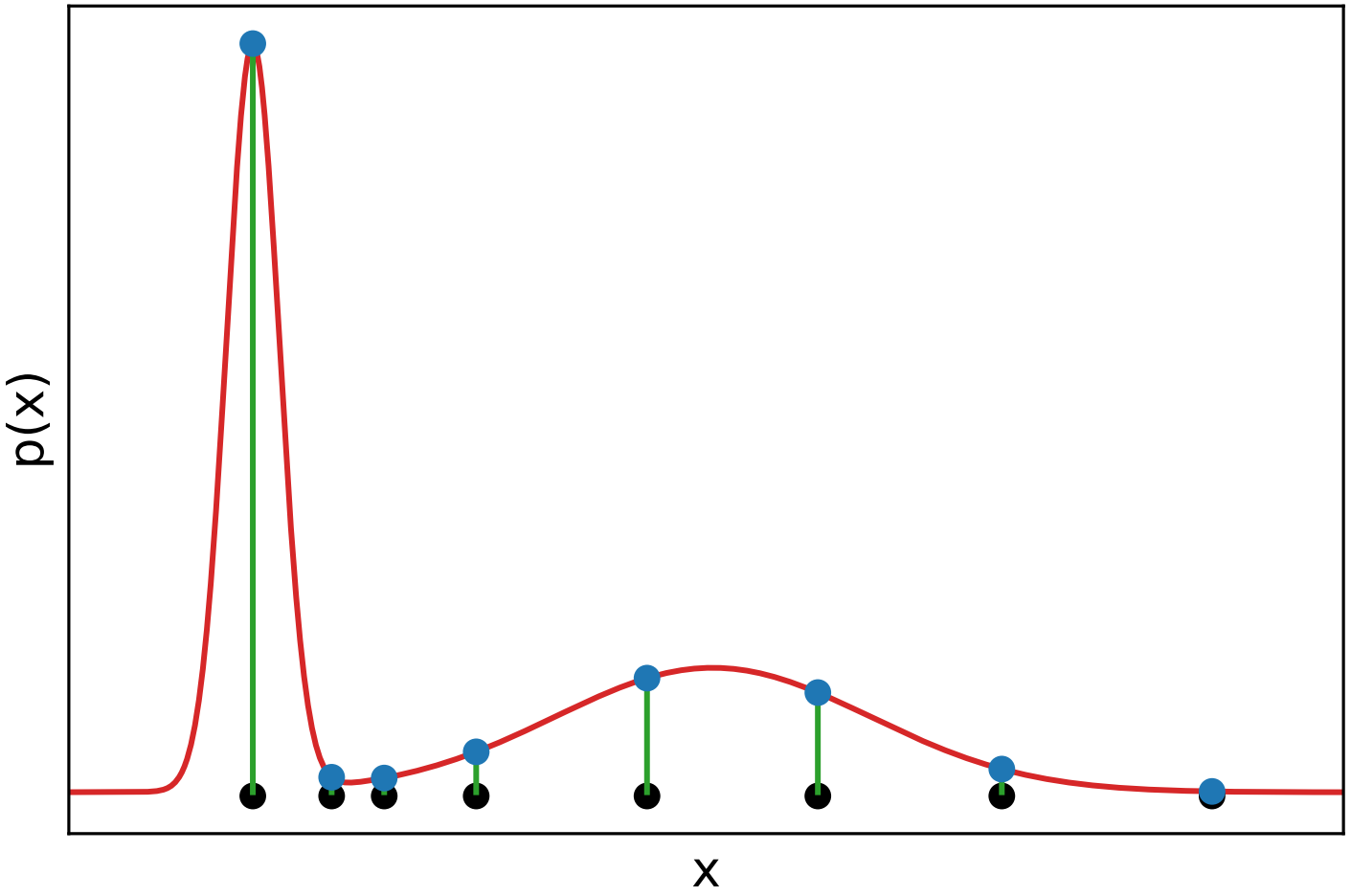
$$\tilde{\boldsymbol{\Sigma}}_k^{(t+1)} = \frac{\breve{\mathbf{S}} + \hat{\boldsymbol{\Sigma}}_k^{(t+1)}}{\breve{\nu} + r_k^{(t)} + D + 2} \tag{174}$$

  - ***Covarianza a priori:*** si $s_d = \frac{1}{N}\sum_n (x_{nd} - \bar{x}_d)^2$ es la varianza global en la dimensión d, una posibilidad consiste en usar

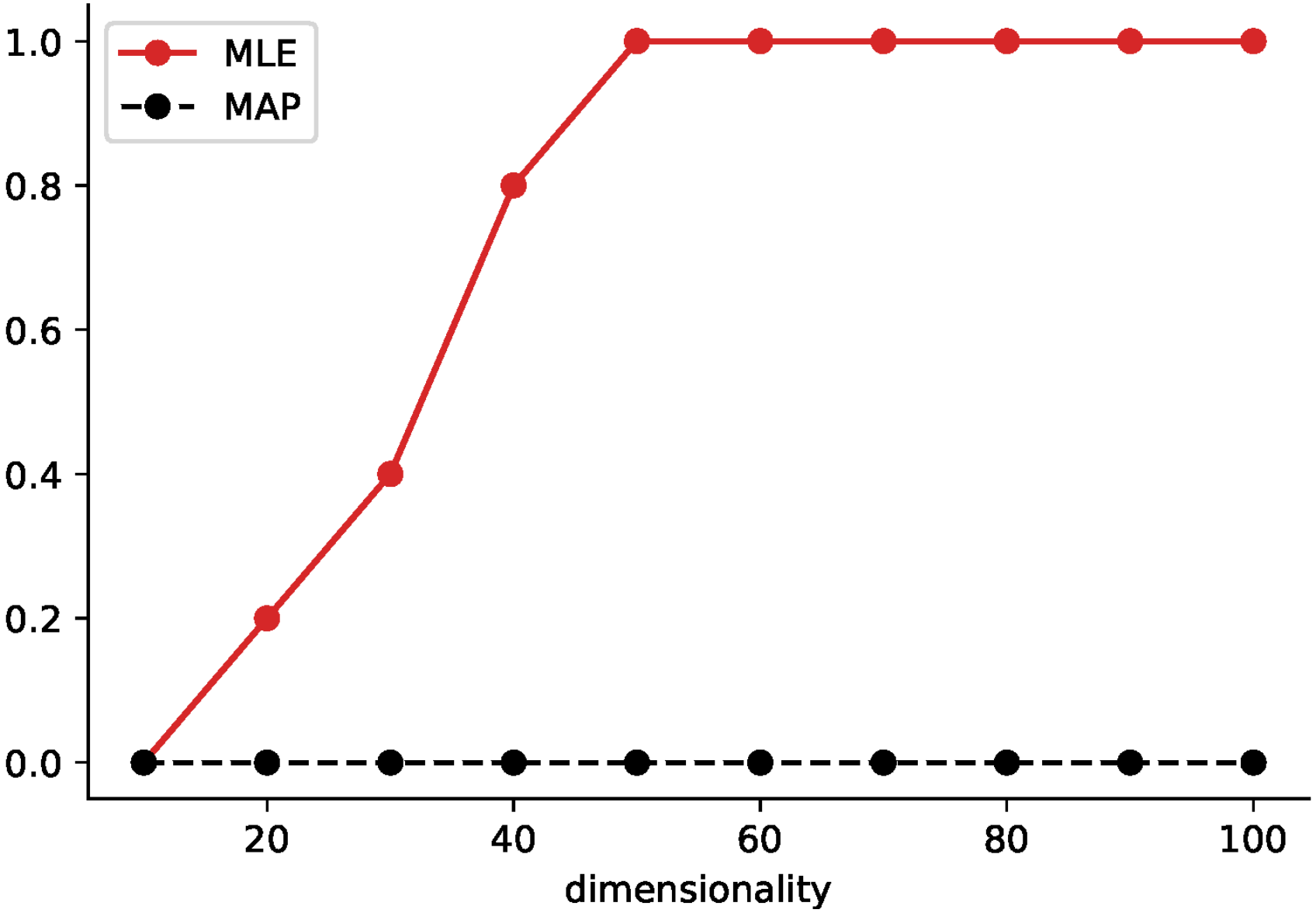
$$\breve{\mathbf{S}} = \frac{1}{K^{1/D}} \mathrm{diag}(s_1^2, \ldots, s_D^2) \tag{175}$$

  - El hiperparámetro $\breve{\nu}$ controla la fuerza del prior; una elección usual es el prior propio más débil: $\breve{\nu} = D + 1$

- *Ejemplo:* mixtura de $K = 2$ componentes a ajustar con $N = 100$ datos sintéticos en D dimensiones ($D = 1$ en la gráfica)
  - ↦ La primera componente es un pico estrecho (con $\sigma_1 \approx 0$) centrado en un único dato x_1



▷ *Ejemplo (cont.):* fracción del número de veces (de 5 intentos) que el EM presenta problemas numéricos con MLE y MAP

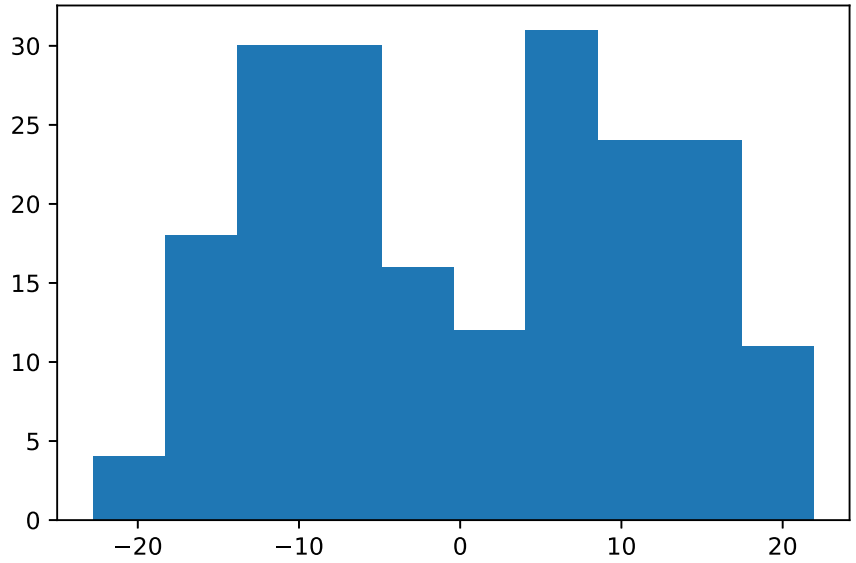


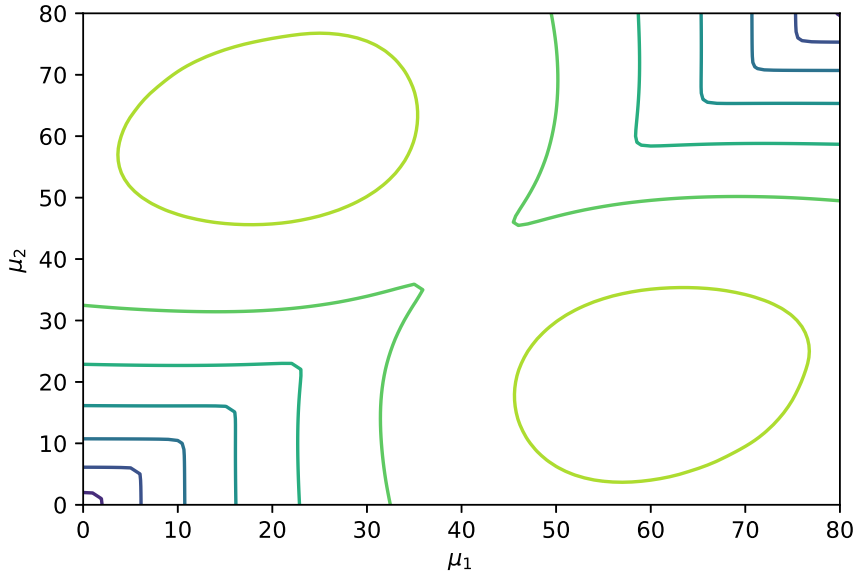
8.7.3.5. Noconvexidad de la NLL

- ***Noconvexidad de la NLL:*** la log-verosimilitud de una mixtura suele tener múltiples modas, esto es, más de un óptimo global

$$\mathrm{LL}(\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \sum_{z_n=1}^{K} p(\boldsymbol{y}_n, z_n \mid \boldsymbol{\theta}) \tag{176}$$

  - *Ejemplo:* 200 puntos de una mixtura de 2 Gaussianas 1d con $\pi_k = 0.5$, $\sigma_k = 5$, $\mu_1 = -10$ y $\mu_2 = 10$; y verosimilitud $p(\mathcal{D} \mid \mu_1, \mu_2)$





    - ↦ ***Label switching problem:*** 2 óptimos cambiando etiquetas

▶ *Complejidad del problema: label switching problem*

- ▷ Es difícil establecer el número de modas pues, aunque potencialmente hay $K!$ etiquetados distintos, muchos picos pueden reducirse al mezclarse con otros cercanos

- ▷ ***Número de modas exponencial:*** en cualquier caso, puede haber un número de modas exponencial con K, por lo que el problema es NP-duro

- ▷ ***Óptimo local:*** únicamente podemos aspirar a encontrar un buen óptimo local, por lo general posible con una buena inicialización