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”

8.1. Introducción

- ***Problema de optimización:*** hallar un $\boldsymbol{\theta} \in \Theta$ que minimice una ***función de pérdida*** o ***función coste*** $\mathcal{L} : \Theta \to \mathbb{R}$:

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \Theta} \mathcal{L}(\boldsymbol{\theta}) \tag{1}$$

- ***Espacio paramétrico:*** $\Theta \subseteq \mathbb{R}^D$; D es el número de variables
- ***Optimización continua:*** no consideramos la discreta
- ***Maximizar*** una ***función de puntuación*** o ***de recomensa*** $R(\boldsymbol{\theta})$ equivale a minimizar $\mathcal{L}(\boldsymbol{\theta}) = -R(\boldsymbol{\theta})$
- ***Función objetivo:*** función a minimizar o maximizar
- ***Solver:*** algoritmo para hallar un óptimo del objetivo

8.1.1. Optimización local vs global

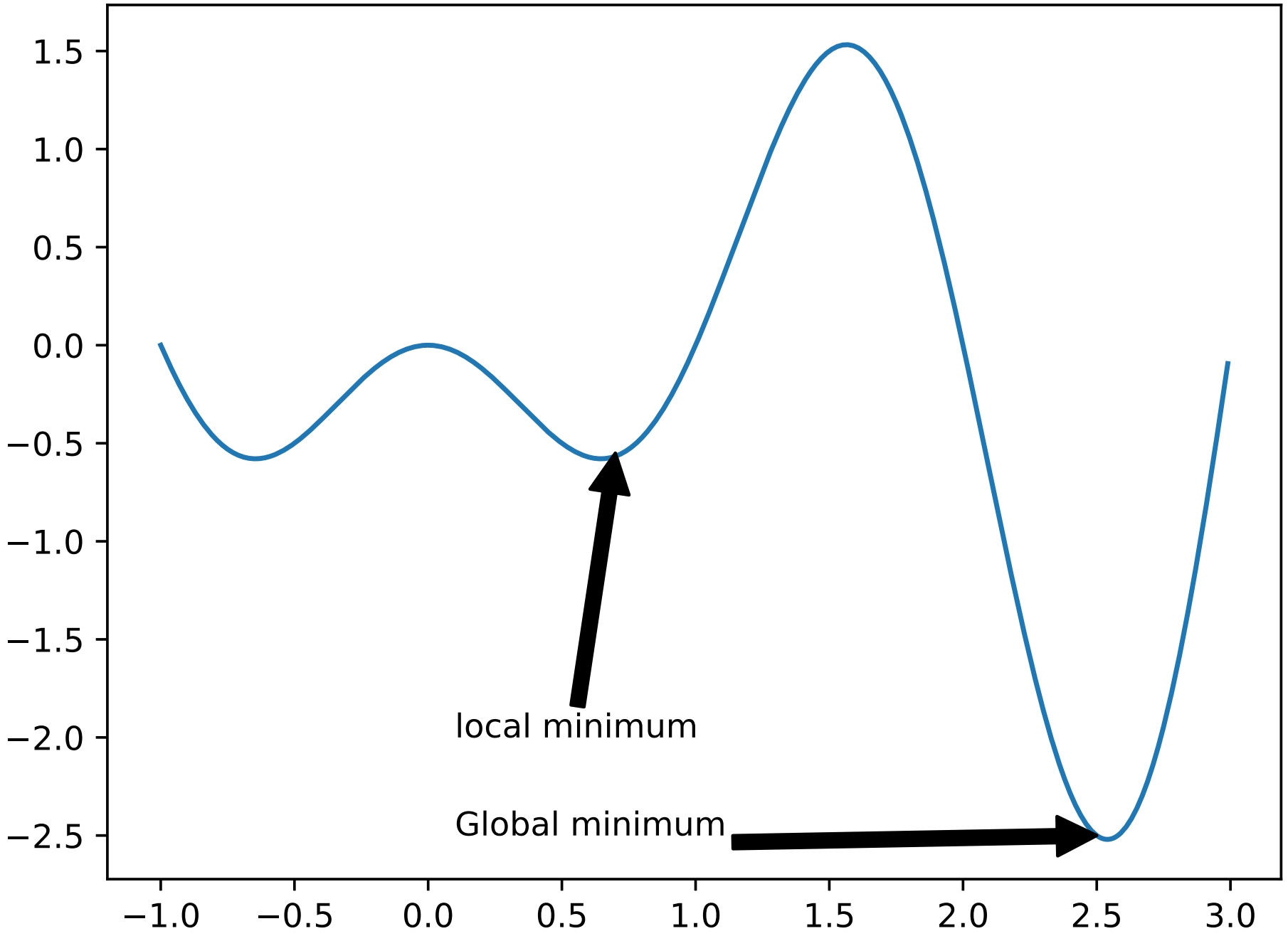
- ***Optimización global:*** hallar un ***óptimo global***
- ***Óptimo local:*** $\boldsymbol{\theta}^*$ es un ***mínimo local*** si

$$\exists \delta > 0, \quad \forall \boldsymbol{\theta} \in \Theta \text{ s.t. } \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta, \quad \mathcal{L}(\boldsymbol{\theta}^*) \leq \mathcal{L}(\boldsymbol{\theta}) \tag{2}$$

- ***Mínimo local (plano):*** quizás rodeado de otros mínimos locales
- ***Mínimo local estricto:***

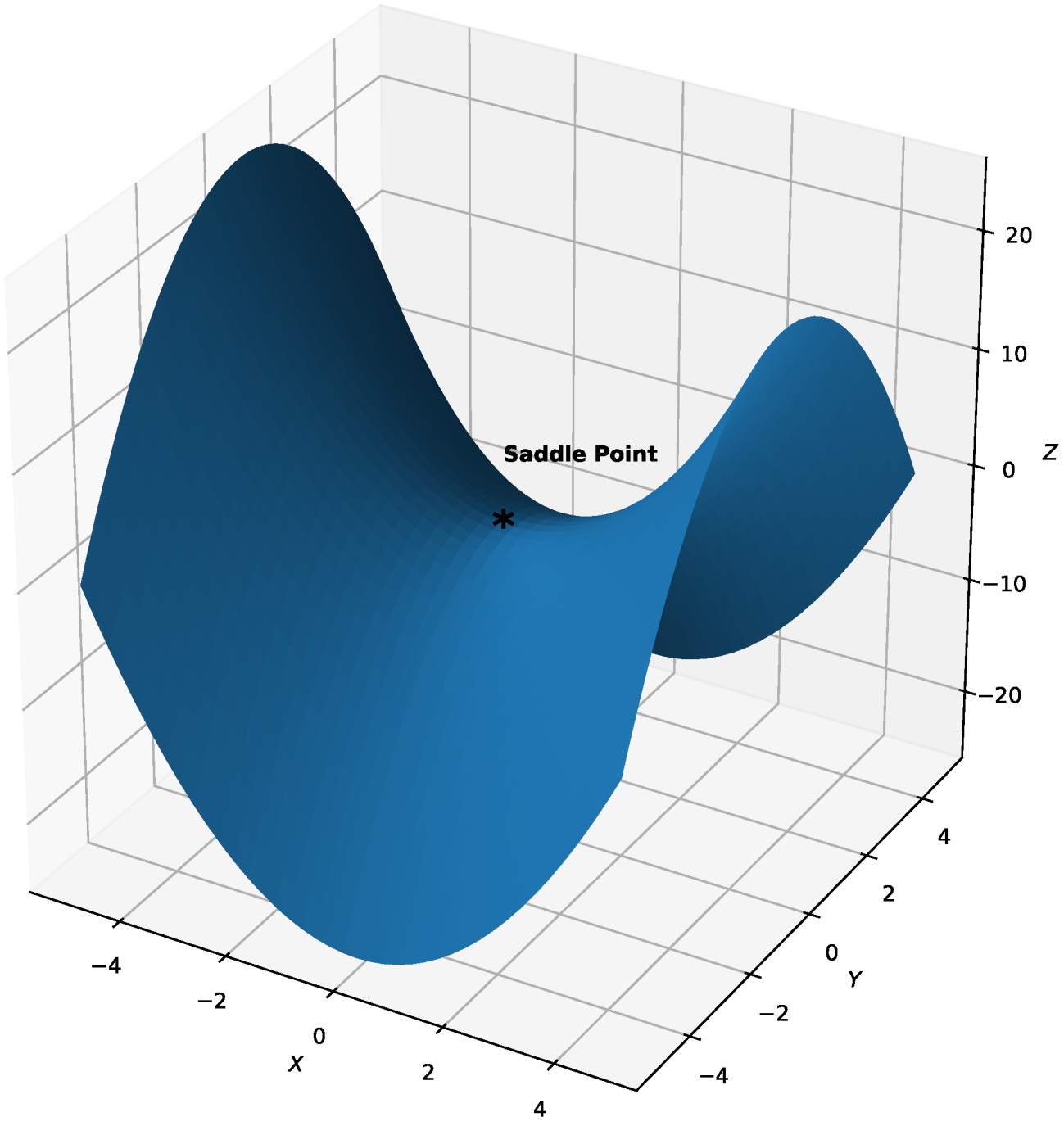
$$\exists \delta > 0, \quad \forall \boldsymbol{\theta} \in \Theta, \boldsymbol{\theta} \neq \boldsymbol{\theta}^* : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| < \delta, \quad \mathcal{L}(\boldsymbol{\theta}^*) < \mathcal{L}(\boldsymbol{\theta}) \tag{3}$$

- ***Algoritmo globalmente convergente:*** garantiza su convergencia a un ***punto estacionario*** del objetivo (óptimo local, global o punto de silla) a partir de cualquier de cualquier punto de partida


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
local minimum
Global minimum

8.1.1.1. Condiciones de optimalidad local

- Sean $\mathcal{L}(\boldsymbol{\theta})$ un objetivo doblemente diferenciable, $\boldsymbol{g}(\boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta})$ el vector gradiente y $\mathbf{H}(\boldsymbol{\theta}) = \nabla^2 \mathcal{L}(\boldsymbol{\theta})$ la matriz Hessiana
- Sean $\boldsymbol{\theta}^* \in \mathbb{R}^D$ un punto, $\boldsymbol{g}^* = \boldsymbol{g}(\boldsymbol{\theta}^*)|_{\boldsymbol{\theta}^*}$ el gradiente en dicho punto y $\mathbf{H}^* = \mathbf{H}(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}$ la correspondiente Hessiana
- ***Condición necesaria:*** Si $\boldsymbol{\theta}^*$ es un mínimo local, $\boldsymbol{g}^* = \mathbf{0}$ (i.e., $\boldsymbol{\theta}^*$ debe ser un ***punto estacionario***) y $\mathbf{H}^*$ es semi-definida positiva
- ***Condición suficiente:*** Si $\boldsymbol{g}^* = \mathbf{0}$ y $\mathbf{H}^*$ es definida positiva, $\boldsymbol{\theta}^*$ es un mínimo local
- *Necesidad de gradiente nulo:* si fuese no nulo, el objetivo se minora a pequeña distancia en la dirección del negativo del gradiente
- *Insuficiencia del gradiente nulo:* un punto estacionario puede ser un mínimo o máximo local, o un ***punto de silla*** (ver figura sig.)
- *Suficiencia de la Hessiana (semi-)definida positiva:* el objetivo no decrece (crece, si no es semi) en el entorno de $\boldsymbol{\theta}^*$


Saddle Point
20
10
0
−10
−20
Z
−4
−2
0
2
4
X
4
2
0
−2
−4
Y

8.1.2. Optimización con o sin restricciones

- ***Optimización sin restricciones:*** debemos hallar cualquier valor del espacio paramétrico Θ que minimice la pérdida

- ***Optimización con restricciones:***

$$\boldsymbol{\theta}^* \in \underset{\boldsymbol{\theta} \in \mathcal{C}}{\arg\min}\ \mathcal{L}(\boldsymbol{\theta}) \tag{4}$$

donde $\mathcal{C} \subseteq \Theta$ es el ***conjunto de soluciones posibles:***

$$\mathcal{C} = \{\boldsymbol{\theta} :\ g_j(\boldsymbol{\theta}) \leq 0 : j \in \mathcal{I},\ h_k(\boldsymbol{\theta}) = 0 : k \in \mathcal{E}\} \in \mathbb{R}^D \tag{5}$$

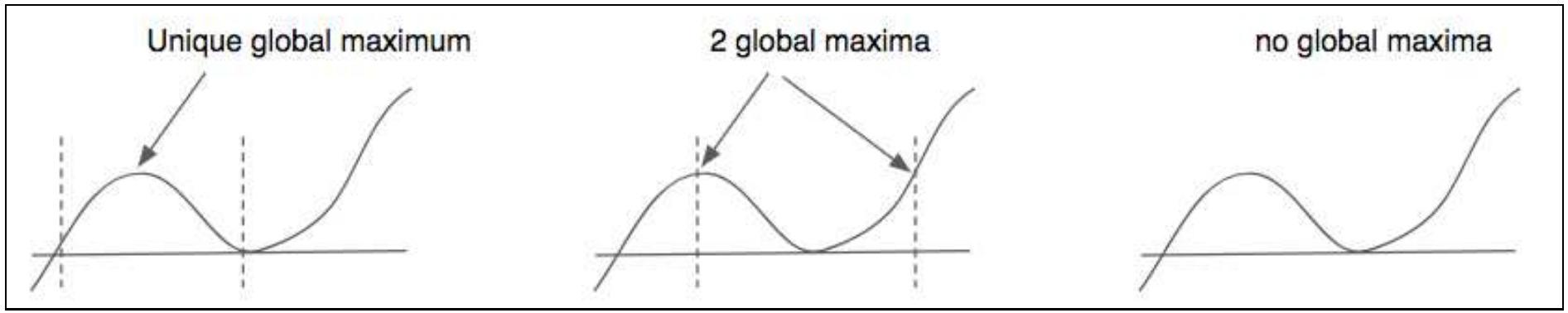
con las ***restricciones de desigualdad***

$$g_j(\boldsymbol{\theta}) \leq 0 : j \in \mathcal{I} \tag{6}$$

y las ***restricciones de igualdad***

$$h_k(\boldsymbol{\theta}) = 0 : k \in \mathcal{E} \tag{7}$$

- La adición de restricciones puede cambiar el número de óptimos:



- Si añadimos muchas restricciones, el conjunto de soluciones posibles puede reducirse mucho, hasta el vacío, y complicar el ***problema de hallar una solución posible*** (sin importar su coste)
- Una estrategia usual para resolver problemas con restricciones consiste en eliminar una o más restricciones, añadiendo un término al objetivo por cada restricción eliminada que lo penalice en función del grado de incumplimiento de la restricción
- Otra estrategia usual para (tratar de) resolver problemas con restricciones consiste en ignorar una o más restricciones de desigualdad y comprobar si la solución hallada las cumple

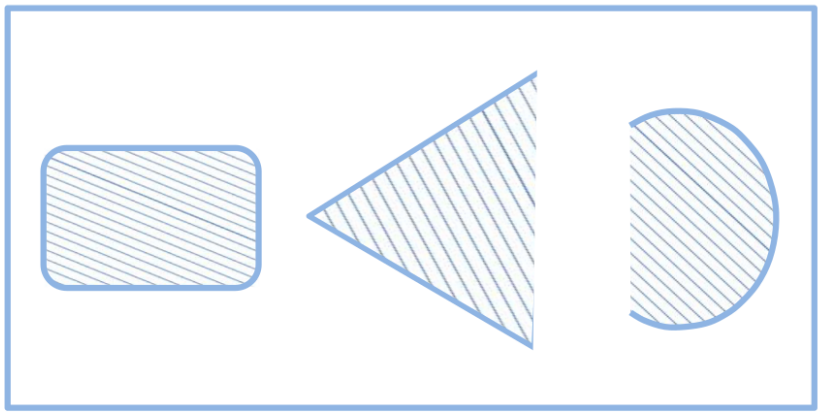
8.1.3. Optimización convexa vs no-convexa

- ***Optimización convexa:*** problema de conjunto de soluciones convexo y función objetivo convexa; cada óptimo local es global

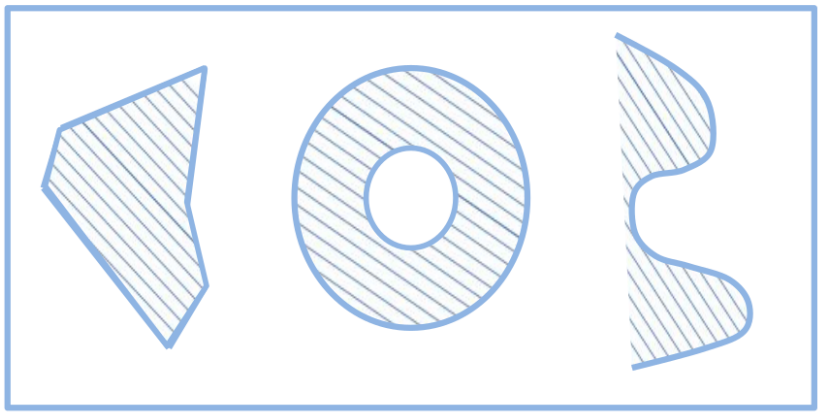
8.1.3.1. Conjuntos convexos

- ***Conjunto convexo:*** $\mathcal{S}$ es convexo si, para todo $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}$:

$$\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}' \in \mathcal{S} \quad \text{para todo} \quad \lambda \in [0,1] \tag{8}$$



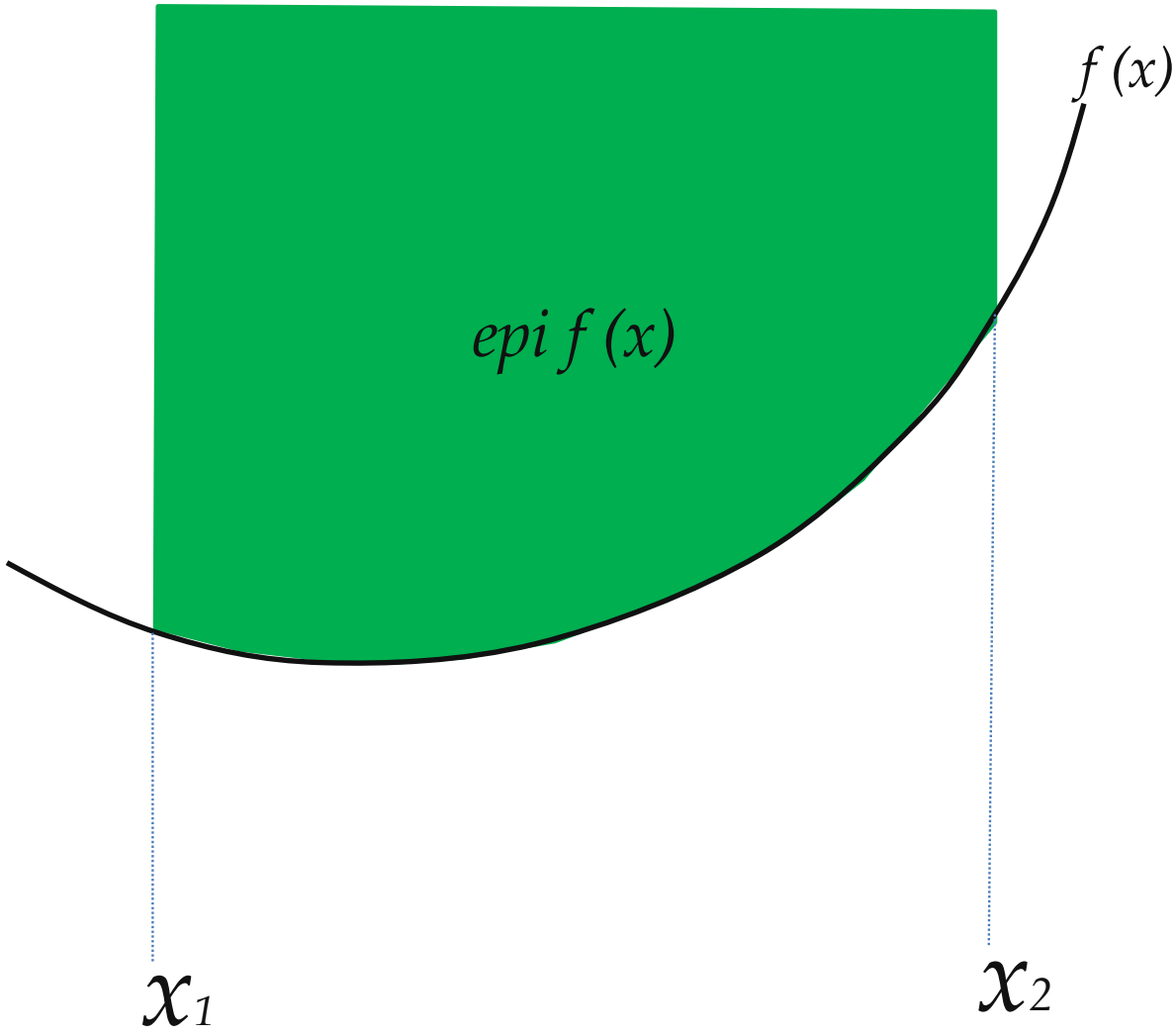
Convex



Not Convex

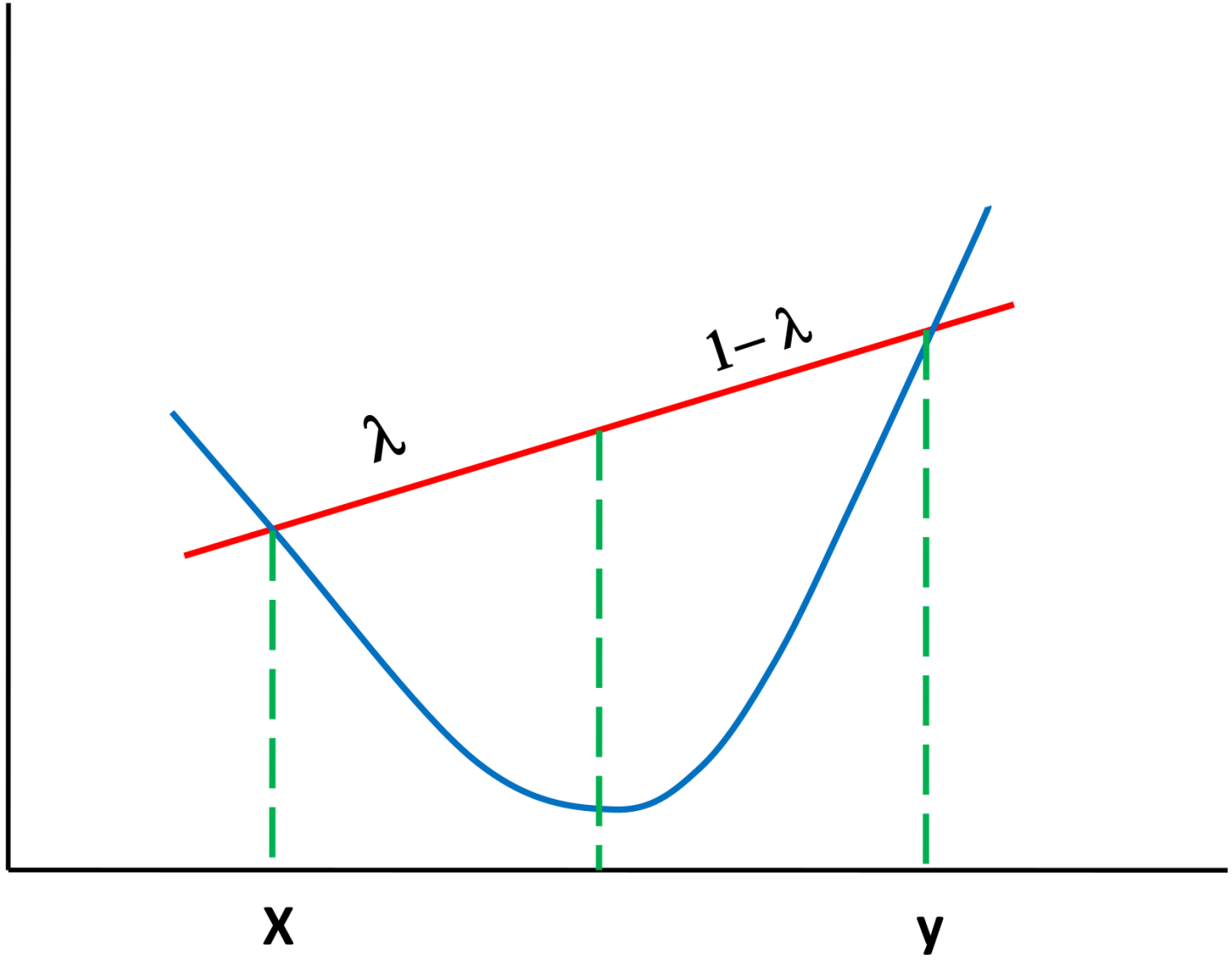
8.1.3.2. Funciones convexas

- ***Función convexa:*** f es convexa si su epigrafo (puntos en f o encima) es conjunto convexo

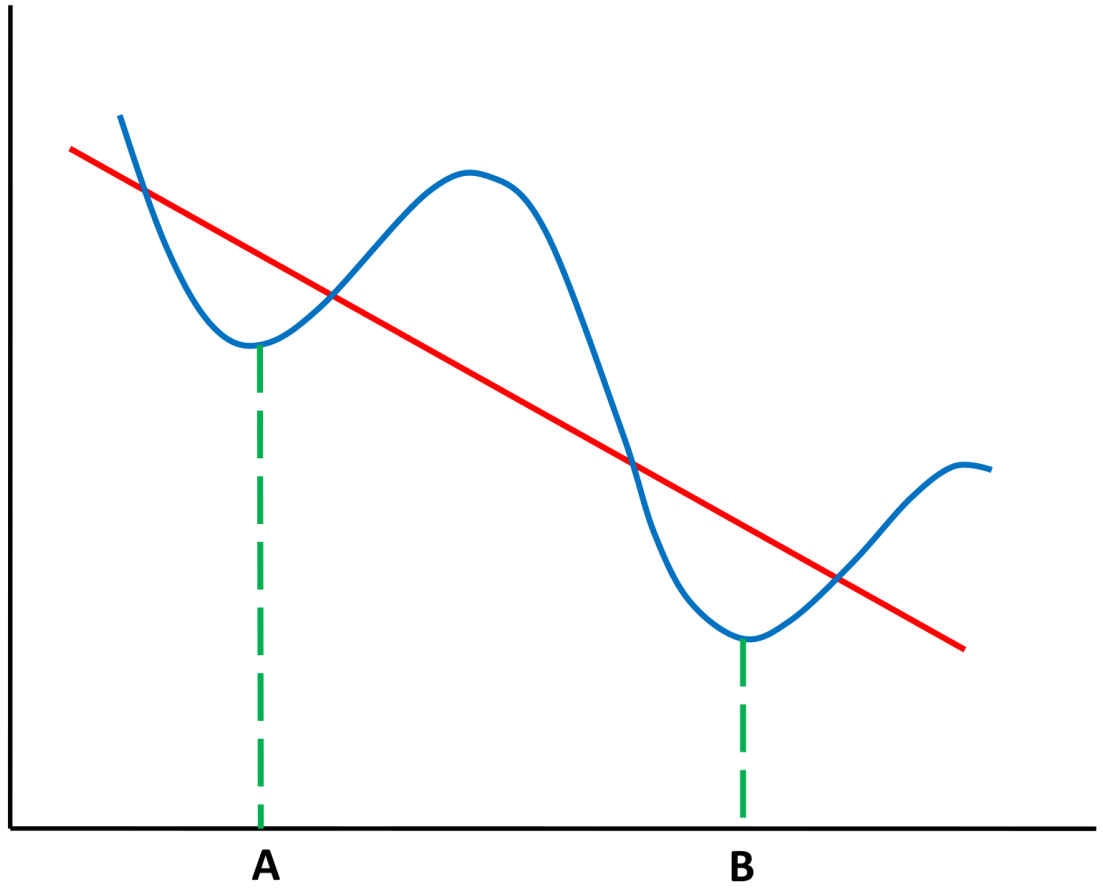


- Alternativamente, f es ***convexa*** si se define sobre un conjunto convexo $\mathcal{S}$ y, para todo $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}$ y $\lambda \in [0, 1]$:

$$f(\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}) \tag{9}$$



- f es ***estrictamente convexa*** si la desigualdad es estricta
- f es ***cóncava*** si $-f$ es convexa
- f es ***estrictamente cóncava*** si $-f$ es estrictamente convexa
- Ejemplo de función que no es convexa ni cóncava:



- Ejemplos de funciones convexas 1d:

$$x^2 \tag{10}$$

$$e^{ax} \tag{11}$$

$$-\log x \tag{12}$$

$$x^a, a > 1, x > 0 \tag{13}$$

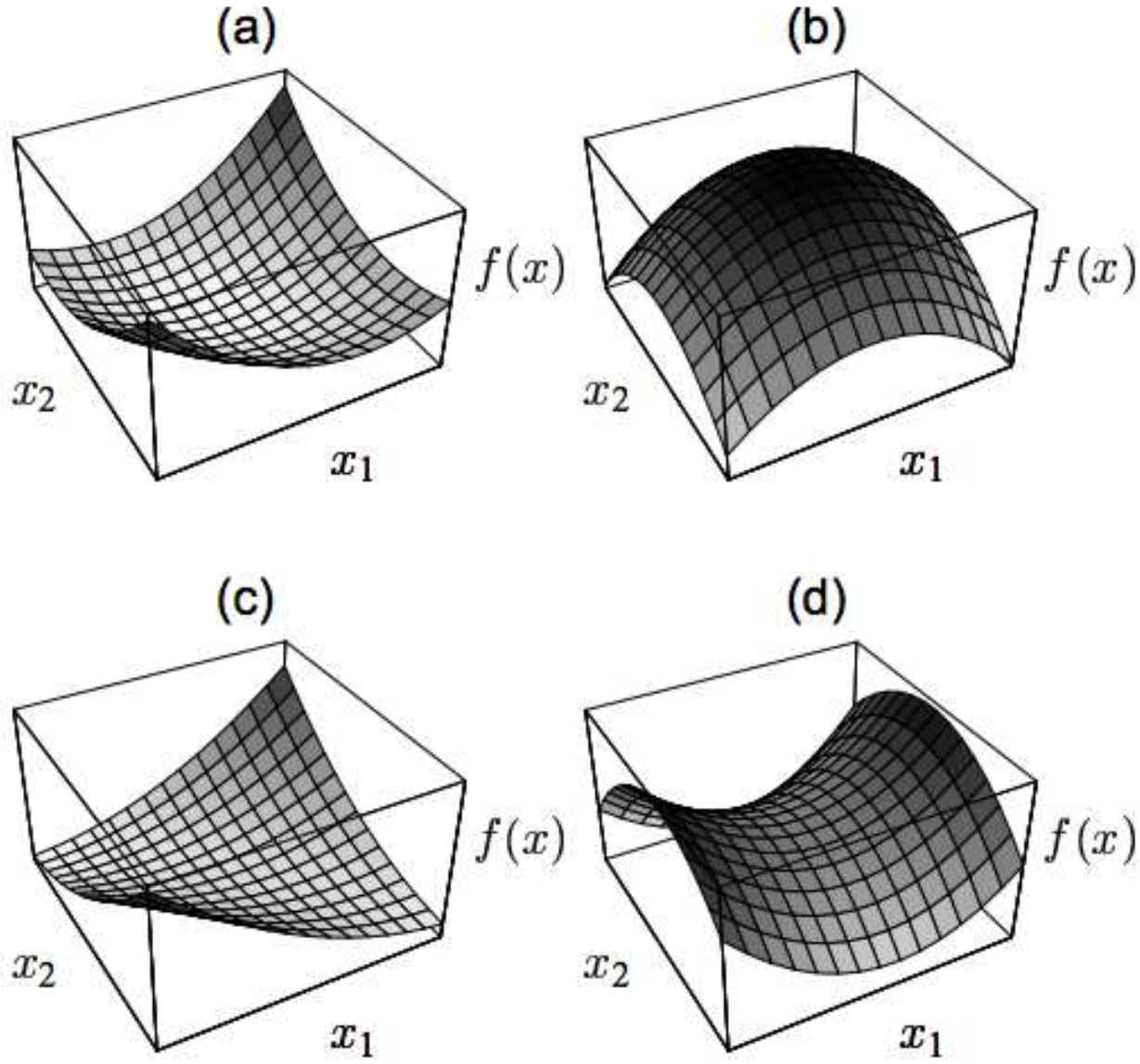
$$|x|^a, a \geq 1 \tag{14}$$

$$x \log x, x > 0 \tag{15}$$

8.1.3.3. Caracterización de funciones convexas

- Intuitivamente, una función convexa tiene forma de bol
- ***Teorema 8.1.1:***
  - Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función doblemente diferenciable
  - f es convexa sii su Hessiana $\mathbf{H} = \nabla^2 f(\boldsymbol{x})$ es semi-definida positiva en el dominio de f
  - f es estrictamente convexa sii $\mathbf{H}$ es definida positiva

- *Ejemplo:* forma cuadrática $f(\boldsymbol{x}) = \boldsymbol{x}^t \mathbf{A} \boldsymbol{x}$ en 2d
  - a) $\mathbf{A}$ +; b) -; c) semi+; d) indefinida



8.1.3.4. Funciones fuertemente convexas

- f ***fuertemente convexa con*** $m > 0$ sii para todo $\boldsymbol{x}, \boldsymbol{y}$ del dominio:

$$(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y}))^t(\boldsymbol{x} - \boldsymbol{y}) \geq m\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \tag{16}$$

- Si f es fuertemente convexa, también es estrictamente convexa

- Si f es dos veces continuamente diferenciable en $\mathbb{R}^n$, f es fuertemente convexa con $m > 0$ sii

$$\nabla^2 f(\boldsymbol{x}) \succeq m\mathbf{I} \quad \text{para todo } \boldsymbol{x} \tag{17}$$

  esto es, $\nabla^2 f(\boldsymbol{x}) - m\mathbf{I}$ es semi-definida positiva

  - Equivalentemente, f es fuertemente convexa con $m > 0$ sii el mínimo valor propio de $\nabla^2 f(\boldsymbol{x})$ es al menos m para todo $\boldsymbol{x}$

- ***Caso de la recta real:*** $f : \mathbb{R} \to \mathbb{R}$
  - $\nabla^2 f(x)$ es la segunda derivada, $f''(x)$
  - f es convexa sii $f''(x) \geq 0$ para todo x
  - f es estrictamente convexa si $f''(x) > 0$ para todo x
  - f es fuertemente convexa sii $f''(x) \geq m > 0$ para todo x

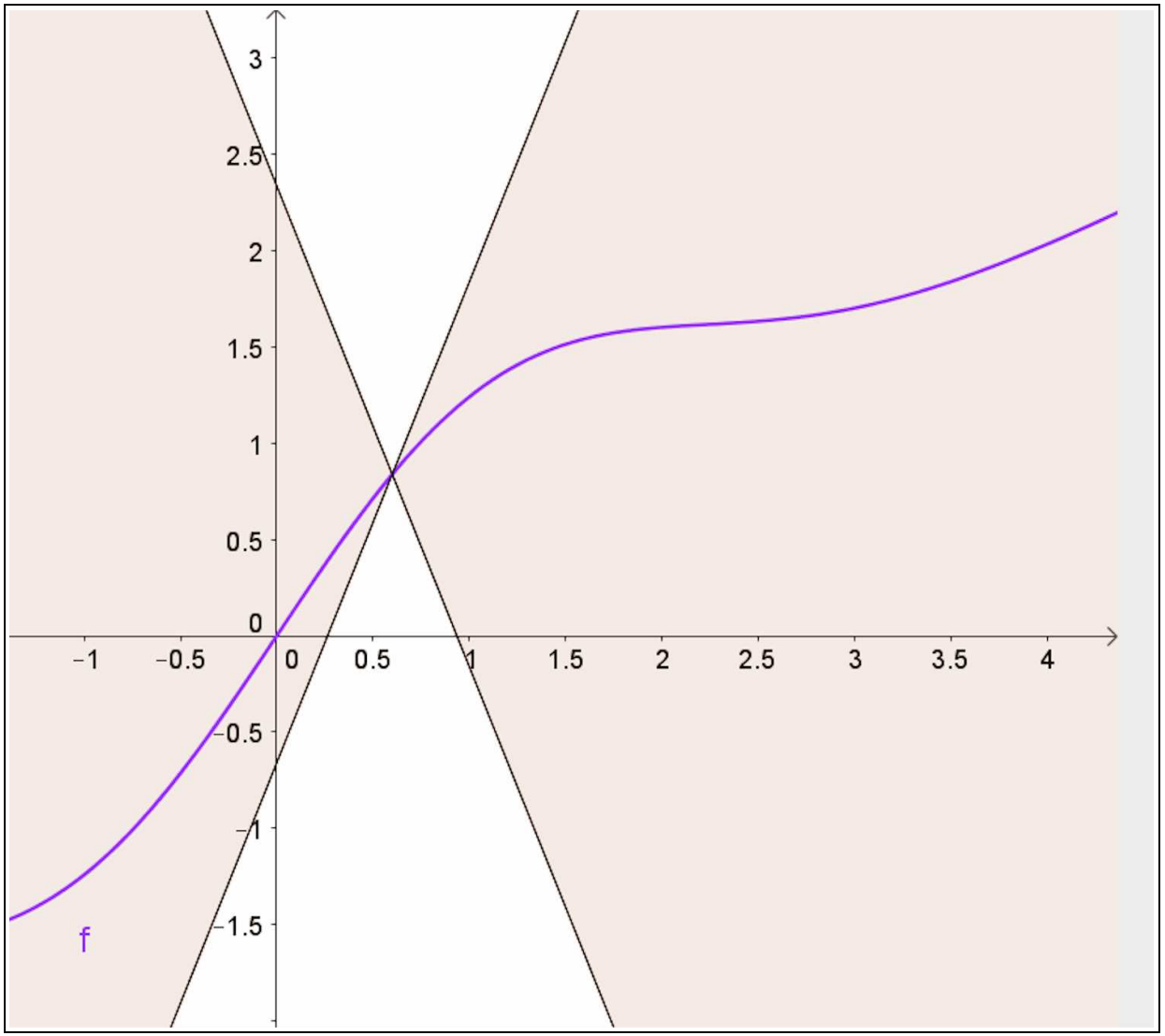
8.1.4. Optimización suave vs no-suave

- ***Optimización suave:*** el objetivo y las restricciones son funciones continuamente diferenciables
  - ***Constante de Lipschitz:*** mide el grado de suavidad de una función suave; en $\mathbb{R}$, es una constante $L \geq 0$ tq para todo x_1, x_2:

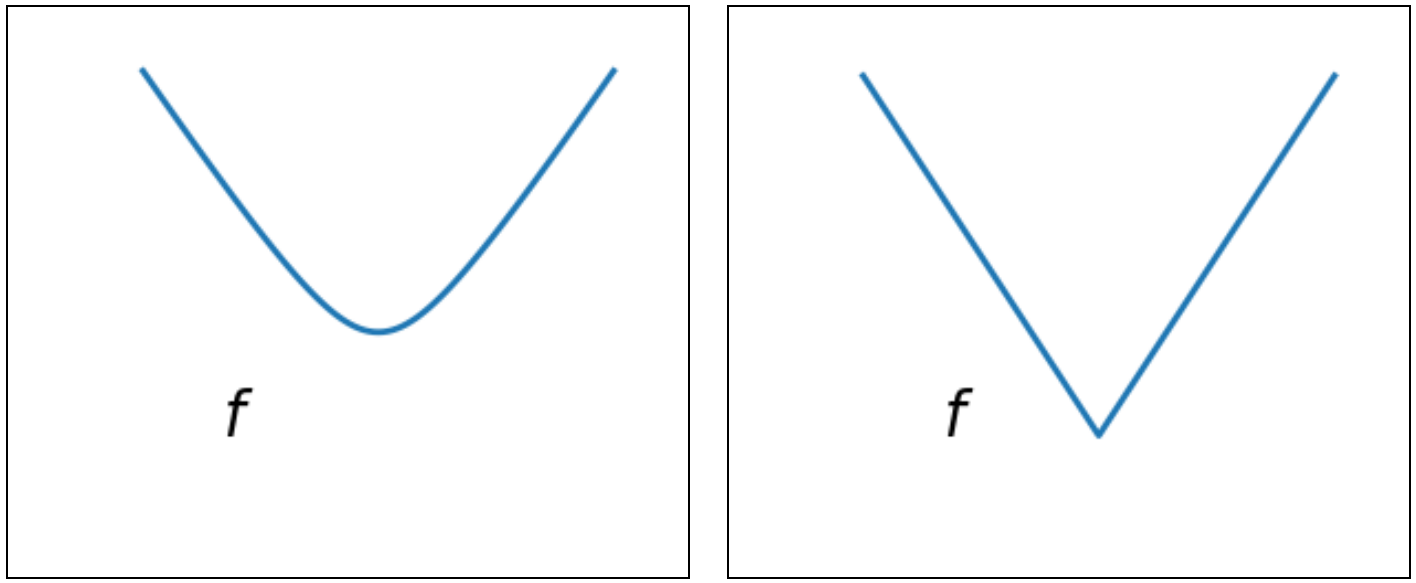
$$|f(x_1) - f(x_2)| \leq L|x_1 - x_2| \tag{18}$$

  - f no puede variar más de L si la entrada varía en una unidad
  - Generalizable a entradas vectoriales con una norma apropiada
- ***Optimización no suave:*** existen al menos algunos puntos en los que el gradiente de la función objetivo o las restricciones no están bien definidos

- *Interpretación de la constante de Lipschitz:* existe un cono doble cuyo origen se mueve a lo largo de f de manera que la gráfica completa siempre queda fuera del cono



- *Ejemplos de función suave y no-suave (con una discontinuidad):*



- ***Objetivo compuesto:*** en ocasiones, el objetivo $\mathcal{L}(\boldsymbol{\theta})$ se divide en una parte suave $\mathcal{L}_s(\boldsymbol{\theta})$ y otra no-suave $\mathcal{L}_r(\boldsymbol{\theta})$ (r de *rough*),

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_s(\boldsymbol{\theta}) + \mathcal{L}_r(\boldsymbol{\theta}) \tag{19}$$

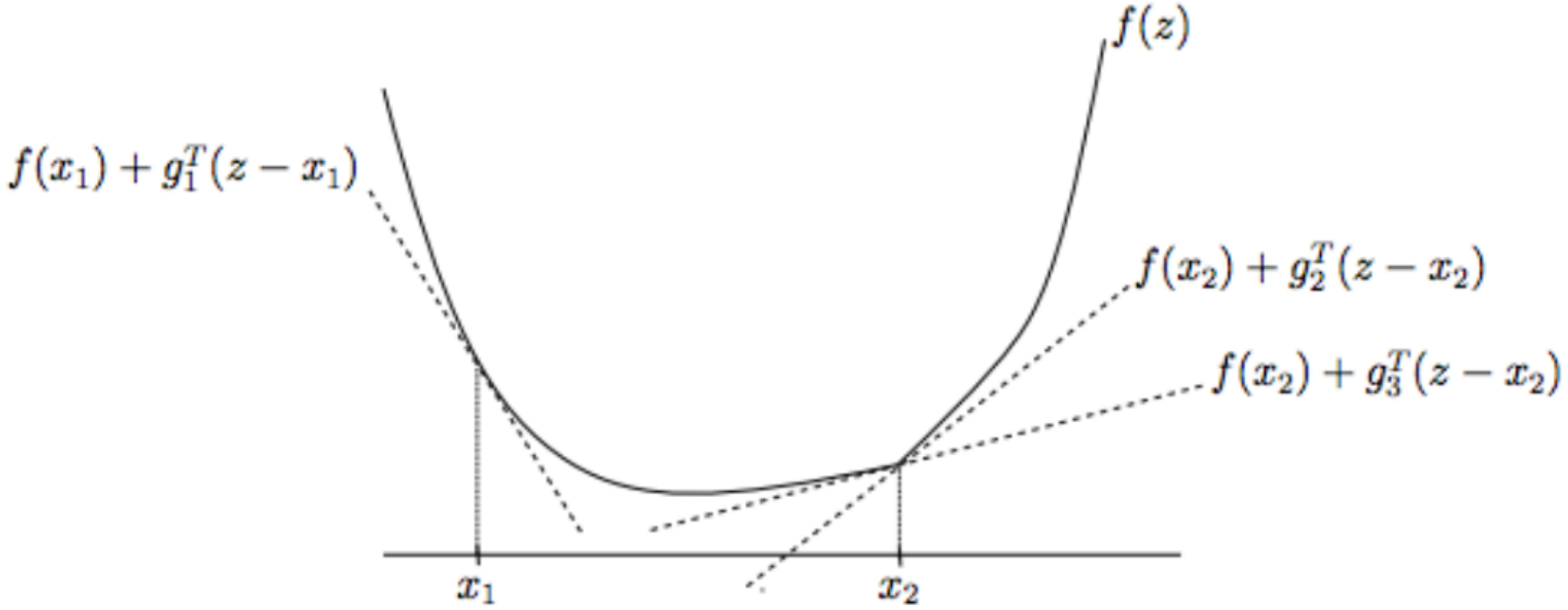
  - En aprendizaje automático, $\mathcal{L}_s(\boldsymbol{\theta})$ suele ser la pérdida empírica y $\mathcal{L}_r(\boldsymbol{\theta})$ un regularizador como la norma ℓ_1 de $\boldsymbol{\theta}$

8.1.4.1. Subgradientes

- Sea $f : \mathbb{R}^n \to \mathbb{R}$ un función convexa; decimos que $\boldsymbol{g} \in \mathbb{R}^n$ es un ***subgradiente*** de f (en $\boldsymbol{x}$ para todo $\boldsymbol{z}$) si:

$$f(\boldsymbol{z}) \geq f(\boldsymbol{x}) + \boldsymbol{g}^t(\boldsymbol{z} - \boldsymbol{x}) \tag{20}$$

- *Ejemplo:* f tiene un único subgradiente en x_1 (la derivada de f en x_1), g_1, pero múltiples subgradientes en x_2, como g_2 y g_3

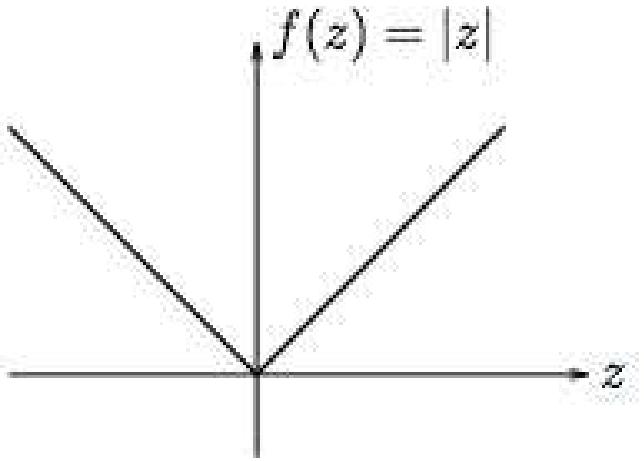


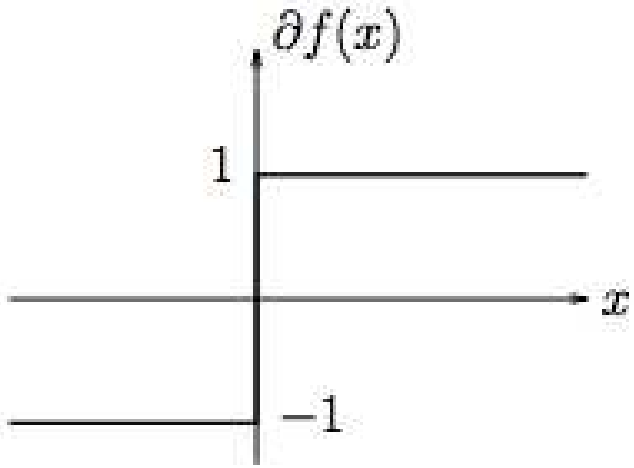
- Una función f es ***subdiferenciable*** en $\boldsymbol{x}$ si tiene al menos un subgradiente en $\boldsymbol{x}$

- El conjunto de tales subgradientes es el ***subdiferencial*** de f en $\boldsymbol{x}$ y se denota como $\partial f(\boldsymbol{x})$

- *Ejemplo:* $f(x) = |x|$

$$\partial f(x) = \begin{cases} \{-1\} & \text{si } x < 0 \\ [-1, 1] & \text{si } x = 0 \\ \{+1\} & \text{si } x > 0 \end{cases} \tag{21}$$





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

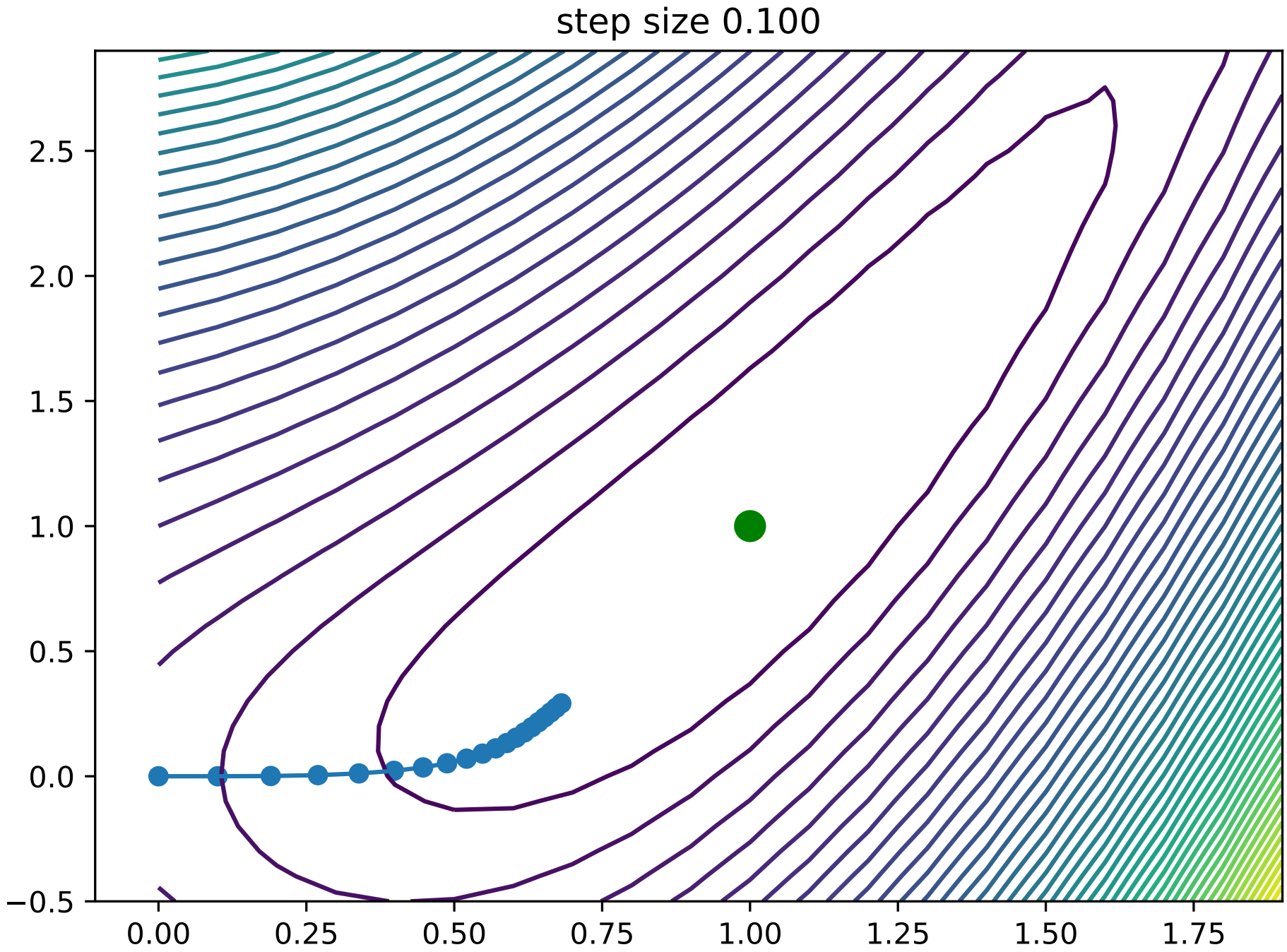
- ***Learning rate schedule:*** secuencia de tamaños de paso $\{\eta_t\}$

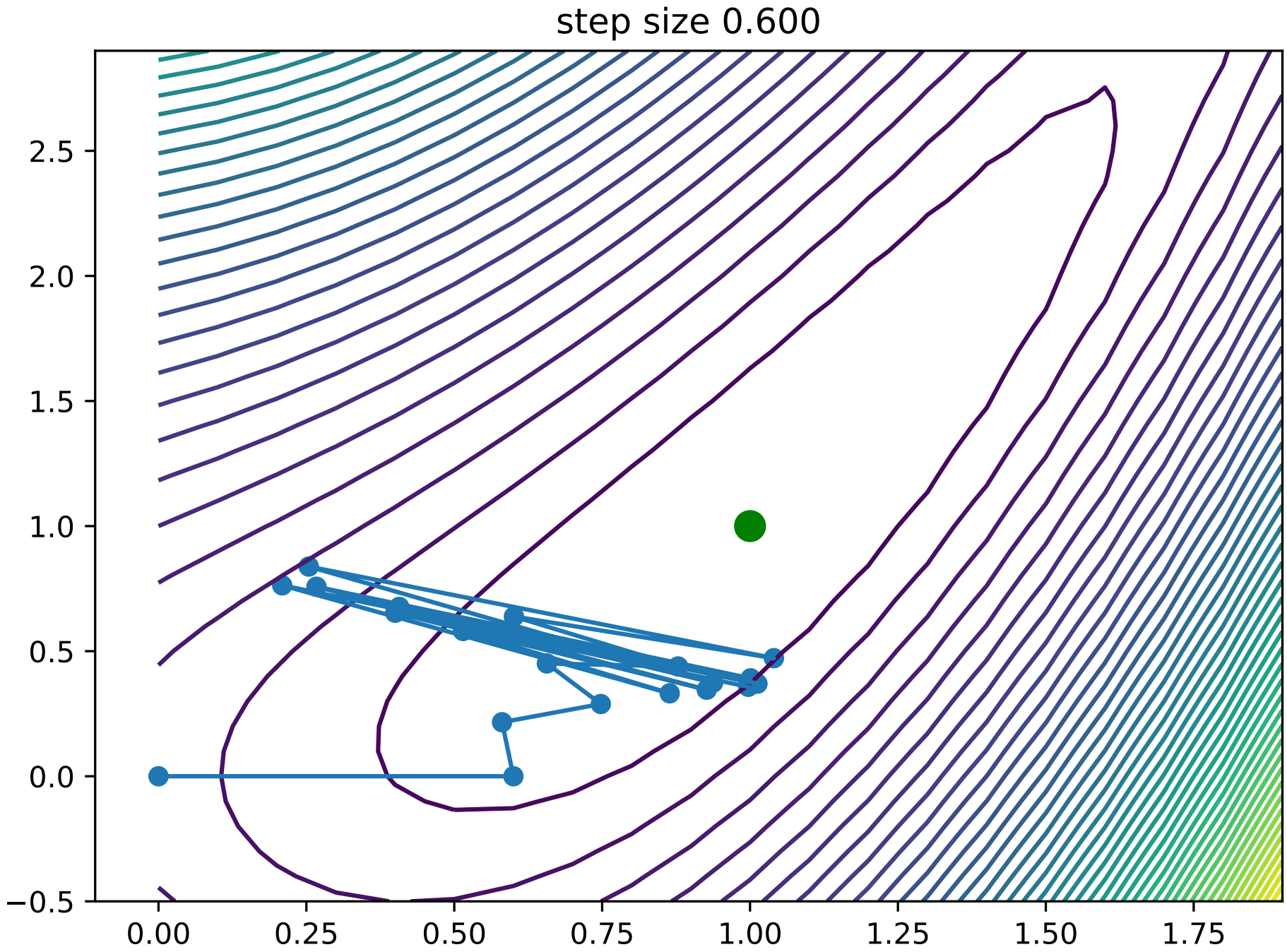
8.2.2.1. Tamaño de paso constante

- 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


step size 0.100
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75


step size 0.600
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

- En algunos casos podemos derivar una cota superior teórica para el máximo tamaño de paso que podemos usar

- *Ejemplo:* con un objetivo cuadrático

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^t \mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}^t\boldsymbol{\theta} + c \quad \text{con} \quad \mathbf{A} \succeq \mathbf{0} \tag{28}$$

  descenso por gradiente será globalmente convergente sii

$$\eta < \frac{2}{\lambda_{\max}(\mathbf{A})} \tag{29}$$

  donde $\lambda_{\max}(\mathbf{A})$ mide la pendiente más pronunciada

- Más generalmente, si L es la constante de Lipschitz del gradiente, la convergencia queda garantizada si

$$\eta < \frac{2}{L} \tag{30}$$

  aunque L es desconocida en la práctica, por lo que usualmente necesitamos adaptar el tamaño de paso

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

- *Ejemplo de búsqueda lineal exacta:*
  - Consideramos una pérdida cuadrática

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^t \mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}^t\boldsymbol{\theta} + c \tag{33}$$

  - Derivada de ϕ:

$$\frac{d\phi(\eta)}{d\eta} = \frac{d}{d\eta}\left[\frac{1}{2}(\boldsymbol{\theta} + \eta\boldsymbol{d})^t \mathbf{A}(\boldsymbol{\theta} + \eta\boldsymbol{d}) + \boldsymbol{b}^t(\boldsymbol{\theta} + \eta\boldsymbol{d}) + c\right] \tag{34}$$

$$= \boldsymbol{d}^t \mathbf{A}(\boldsymbol{\theta} + \eta\boldsymbol{d}) + \boldsymbol{d}^t\boldsymbol{b} \tag{35}$$

$$= \boldsymbol{d}^t(\mathbf{A}\boldsymbol{\theta} + \boldsymbol{b}) + \eta\boldsymbol{d}^t\mathbf{A}\boldsymbol{d} \tag{36}$$

  - Igualando a cero obtenemos:

$$\eta = -\frac{\boldsymbol{d}^t(\mathbf{A}\boldsymbol{\theta} + \boldsymbol{b})}{\boldsymbol{d}^t\mathbf{A}\boldsymbol{d}} \tag{37}$$

- ***Búsqueda lineal aproximada*** emplea algún método eficiente que garantice una reducción suficiente de la función objetivo

  - ***Método de backtracking Armijo:*** partiendo del η actual o uno grande, lo reduce mediante un factor $0 < \beta < 1$ en cada paso hasta cumplir la condición de ***Armijo-Goldstein***

$$\mathcal{L}(\boldsymbol{\theta}_t + \eta\, \boldsymbol{d}_t) \leq \mathcal{L}(\boldsymbol{\theta}_t) + c\, \eta\, \boldsymbol{d}_t^t \nabla \mathcal{L}(\boldsymbol{\theta}_t) \tag{38}$$

  donde $c \in (0, 1)$ es una constante, típicamente $c = 10^{-4}$

  ↦ En la práctica, la inicialización de la búsqueda y cómo hacer el backtracking pueden afectar significativamente el rendimiento

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

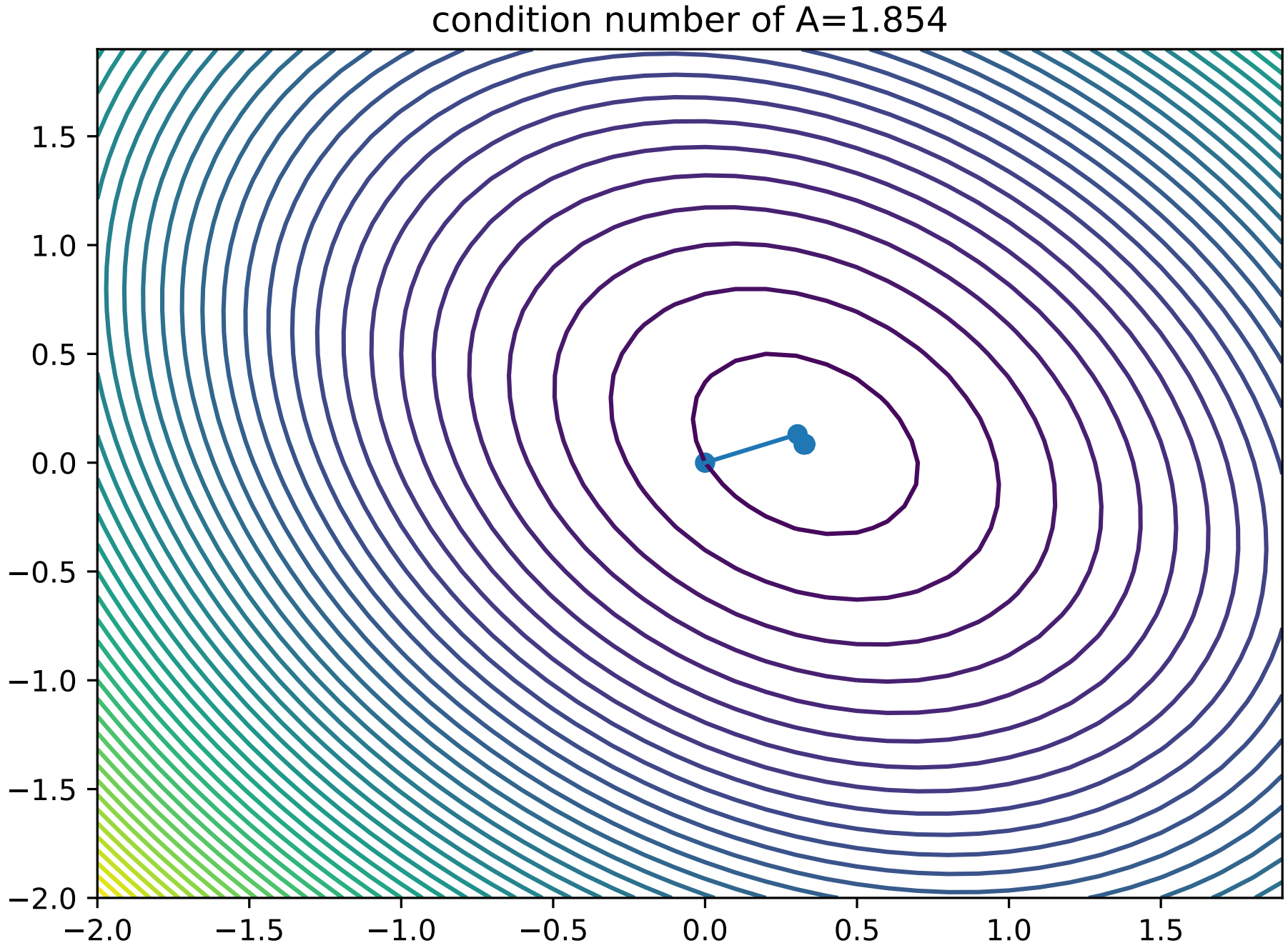
$$\mu = \left(\frac{\lambda_{\text{max}}(\mathbf{A}) - \lambda_{\text{min}}(\mathbf{A})}{\lambda_{\text{max}}(\mathbf{A}) + \lambda_{\text{min}}(\mathbf{A})}\right)^2 = \left(\frac{\kappa(\mathbf{A}) - 1}{\kappa(\mathbf{A}) + 1}\right)^2 \tag{41}$$

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

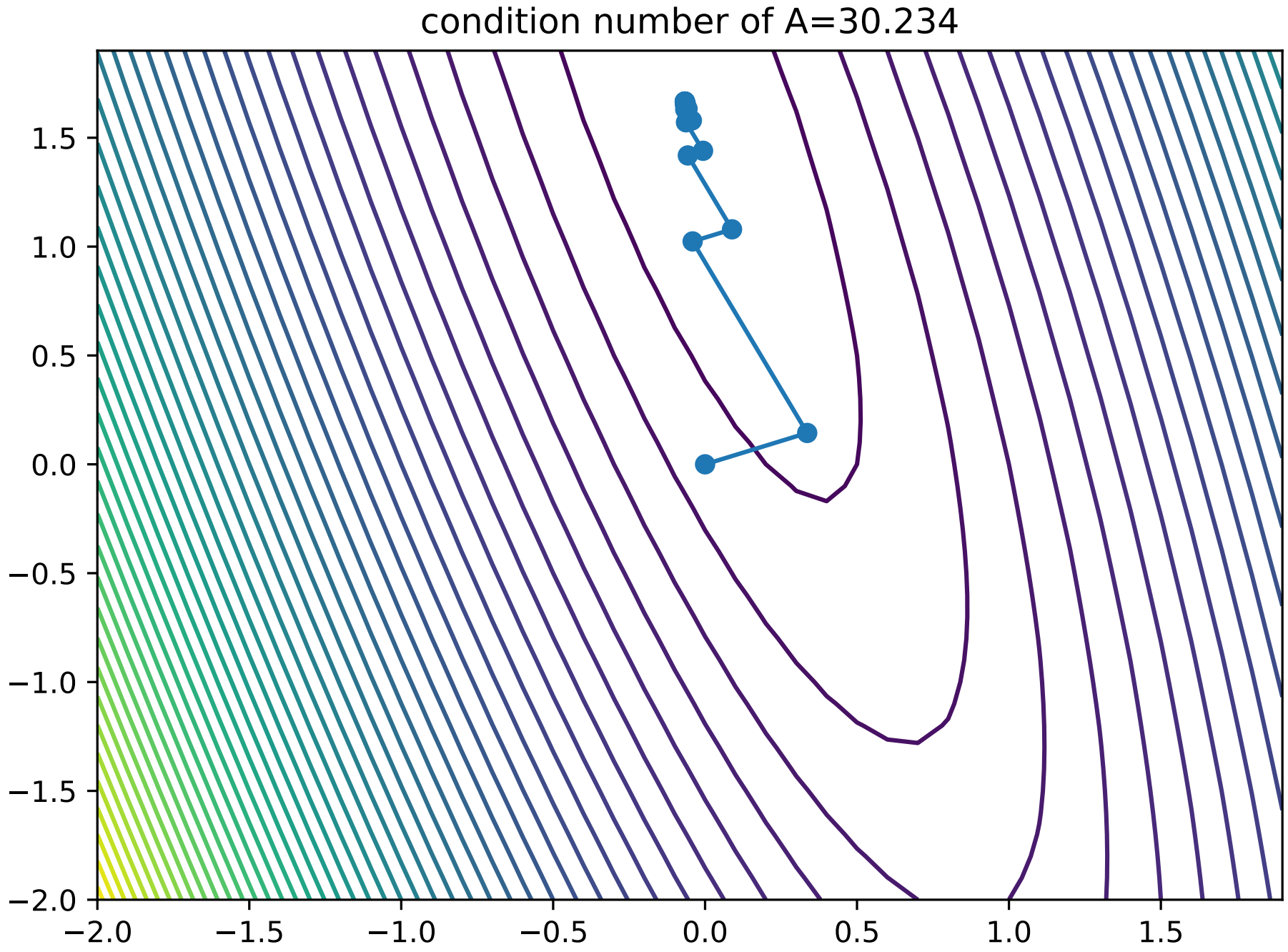
$$\kappa(\mathbf{A}) = \frac{\lambda_{\text{max}}(\mathbf{A})}{\lambda_{\text{min}}(\mathbf{A})} \tag{42}$$

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

▷ *Ejemplo:* $\mathbf{A} = [20, 5; 5, 16], \quad b = [-14; -6], \quad c = 10, \quad \kappa(\mathbf{A}) = 1.8541$



▷ *Ejemplo:* $\mathbf{A} = [20, 5; 5, 2]$, $b = [-14; -6]$, $c = 10$, $\kappa(\mathbf{A}) = 30.234$



- ***Objetivos no cuadráticos:***
  - Con objetivos complejos, el objetivo será por lo general localmente cuadrático alrededor de un óptimo local, por lo que el ratio dependerá del número de condición de la Hessiana en el mismo
  - En general podremos acelerar la convergencia con un objetivo sustituto que aproxime la Hessiana del objetivo en cada paso
- ***Descenso por gradiente conjugado*** es un método alternativo que trata de aliviar el comportamiento ineficiente (en zig-zag) de descenso por gradiente con búsqueda lineal exacta

8.2.4. Momentum

- Descenso por gradiente puede moverse muy lentamente en regiones llanas de la pérdida, por lo que se suelen aplicar heurísticos para acelerar la convergencia

- En física clásica, la ***cantidad de movimiento***, ***movimiento lineal***, ***ímpetu*** o ***momentum*** es el producto de la masa de un cuerpo por su velocidad instantánea

- Se conserva en un sistema cerrado

8.2.4.1. Momentum

- ***Momentum:*** heurístico que acelera el movimiento en direcciones previamente buenas y lo frena en las que el gradiente ha cambiado súbitamente, como una bola pesada rodando montaña abajo:

$$\boldsymbol{m}_t = \beta \, \boldsymbol{m}_{t-1} + \boldsymbol{g}_{t-1} \tag{43}$$

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \, \boldsymbol{m}_t \tag{44}$$

donde $\boldsymbol{m}_t$ es el momentum y $0 < \beta < 1$; usualmente $\beta = 0.9$

- Con $\beta = 0$, coincide con descenso por gradiente:

$$\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} - \eta_t \, \boldsymbol{g}_{t-1} \tag{45}$$

- Se observa que $\boldsymbol{m}_t$ viene a ser una media móvil de gradientes ponderados exponencialmente (EWMA):

$$\boldsymbol{m}_t = \beta \boldsymbol{m}_{t-1} + \boldsymbol{g}_{t-1} \tag{46}$$

$$= \beta^2 \boldsymbol{m}_{t-2} + \beta \boldsymbol{g}_{t-2} + \boldsymbol{g}_{t-1} \tag{47}$$

$$= \cdots = \sum_{\tau=0}^{t-1} \beta^\tau \boldsymbol{g}_{t-\tau-1} \tag{48}$$

- Si todos los gradientes pasados son iguales, digamos $\boldsymbol{g}$:

$$\boldsymbol{m}_t = \boldsymbol{g} \sum_{\tau=0}^{t-1} \beta^\tau \tag{49}$$

- El factor de escalado es una seria geométrica de suma infinita:

$$1 + \beta + \beta^2 + \cdots = \sum_{i=0}^{\infty} \beta^i = \frac{1}{1-\beta} \tag{50}$$

- Luego, en el límite, multiplicamos $\boldsymbol{g}$ por $\frac{1}{1-\beta}$; por 10 si $\beta = 0.9$

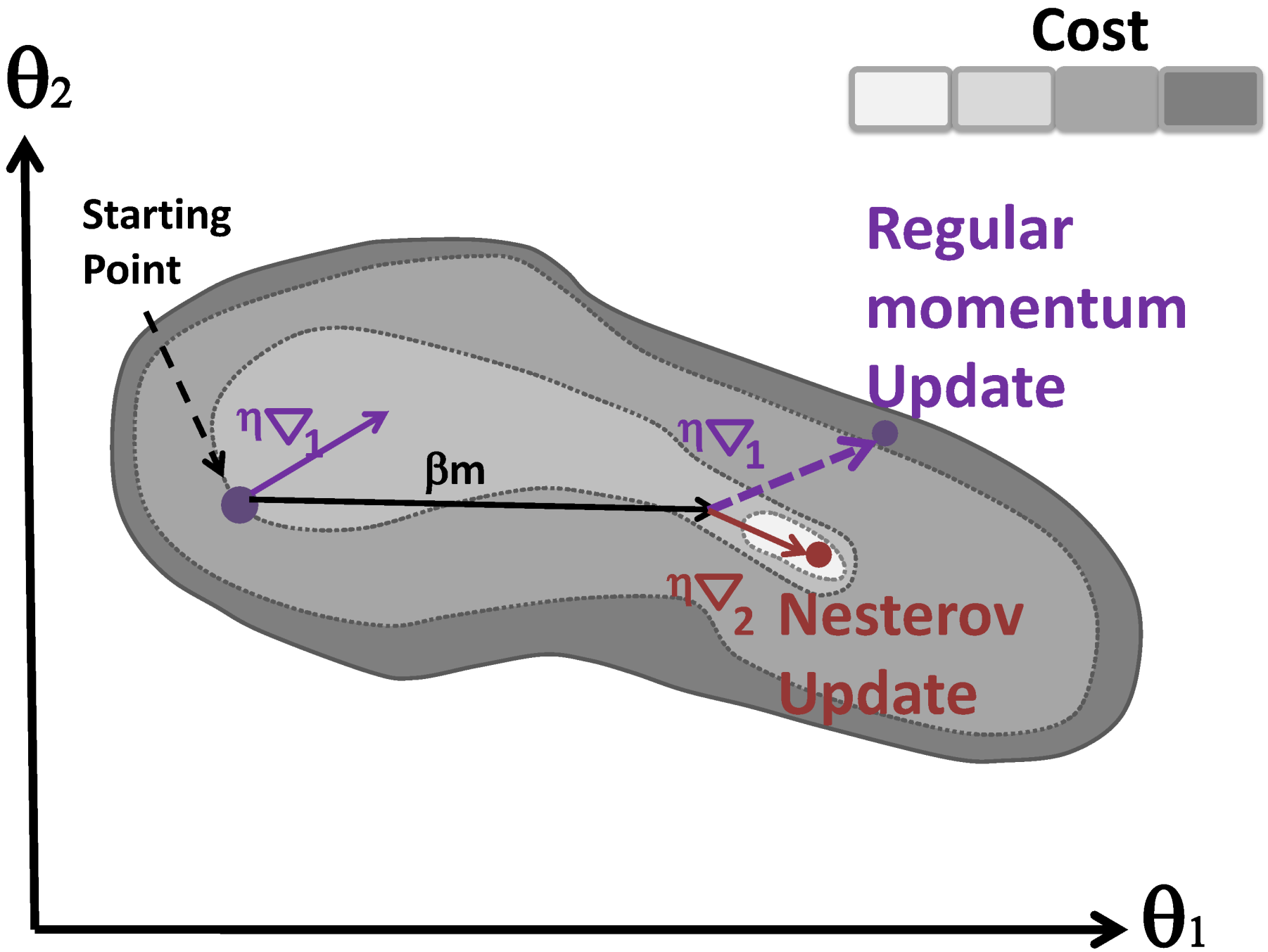
8.2.4.2. Momentum Nesterov

- Momentum estándar oscila al final del valle por no frenar bastante
- El ***gradiente acelerado de Nesterov*** modifica descenso por gradiente añadiendo un paso de extrapolación

$$\tilde{\boldsymbol{\theta}}_{t+1} = \boldsymbol{\theta}_t + \beta_t(\boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1}) \tag{51}$$

$$\boldsymbol{\theta}_{t+1} = \tilde{\boldsymbol{\theta}}_{t+1} - \eta_t \nabla \mathcal{L}(\tilde{\boldsymbol{\theta}}_{t+1}) \tag{52}$$

- El paso de extrapolación añadido actúa a modo de "mirada al futuro" (*look-ahead*) para amortiguar oscilaciones


Cost
θ2
Starting
Point
Regular
momentum
Update
η∇1
βm
η∇1
η∇2
Nesterov
Update
θ1

- El gradiente acelerado de Nesterov se suele llamar ***momentum Nesterov*** porque se puede ver como una variante del momentum estándar:

$$\boldsymbol{m}_{t+1} = \beta \boldsymbol{m}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{\theta}_t + \beta \boldsymbol{m}_t) \tag{53}$$

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \boldsymbol{m}_{t+1} \tag{54}$$

- Si el momentum actual apunta en la buena dirección, el gradiente en $\boldsymbol{\theta}_t + \beta \boldsymbol{m}_t$ será más preciso que en $\boldsymbol{\theta}_t$, por lo que se acelerará la convergencia (con valores apropiados de β y η_t)

8.3. Métodos de segundo orden

- Los métodos de primer orden aprovechan que el gradiente es fácil de calcular y almacenar, pero no modelan la curvatura del espacio, por lo que convergen lentamente

- En contraste con los métodos de primer orden, los de segundo orden tienen en cuenta la curvatura del espacio, típicamente mediante la Hessiana, con el fin de acelerar la convergencia

8.3.1. Método de Newton

- El ***método de Newton*** es un clásico de segundo orden:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{H}_t^{-1} \boldsymbol{g}_t \tag{55}$$

donde la Hessiana

$$\mathbf{H}_t \triangleq \nabla^2 \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_t} = \nabla^2 \mathcal{L}(\boldsymbol{\theta}_t) = \mathbf{H}(\boldsymbol{\theta}_t) \tag{56}$$

se asume definida positiva

- Intuitivamente, $\mathbf{H}_t^{-1}$ "deshace" la curvatura del espacio en $\boldsymbol{\theta}_t$, facilitando así un descenso por gradiente rápido

▸ *Algoritmo de Newton:*

1. Inicializar $\boldsymbol{\theta}_0$
2. Para $t = 1, 2, \ldots$ (hasta convergencia):
3. Evaluar $\boldsymbol{g}_t = \nabla \mathcal{L}(\boldsymbol{\theta}_t)$
4. Evaluar $\mathbf{H}_t = \nabla^2 \mathcal{L}(\boldsymbol{\theta}_t)$
5. Hallar $\boldsymbol{d}_t$ tal que $\mathbf{H}_t \boldsymbol{d}_t = -\boldsymbol{g}_t$
6. Hallar η_t por búsqueda lineal en $\boldsymbol{d}_t$
7. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta_t \boldsymbol{d}_t$

▸ *Derivación del algoritmo de Newton:*

- Aproximación Taylor de segundo orden de $\mathcal{L}(\boldsymbol{\theta})$ en $\boldsymbol{\theta}_t$:

$$\mathcal{L}_{\text{quad}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^t(\boldsymbol{\theta} - \boldsymbol{\theta}_t) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_t)\mathbf{H}_t(\boldsymbol{\theta} - \boldsymbol{\theta}_t) \tag{57}$$

- El mínimo de $\mathcal{L}_{\text{quad}}$ se halla en:

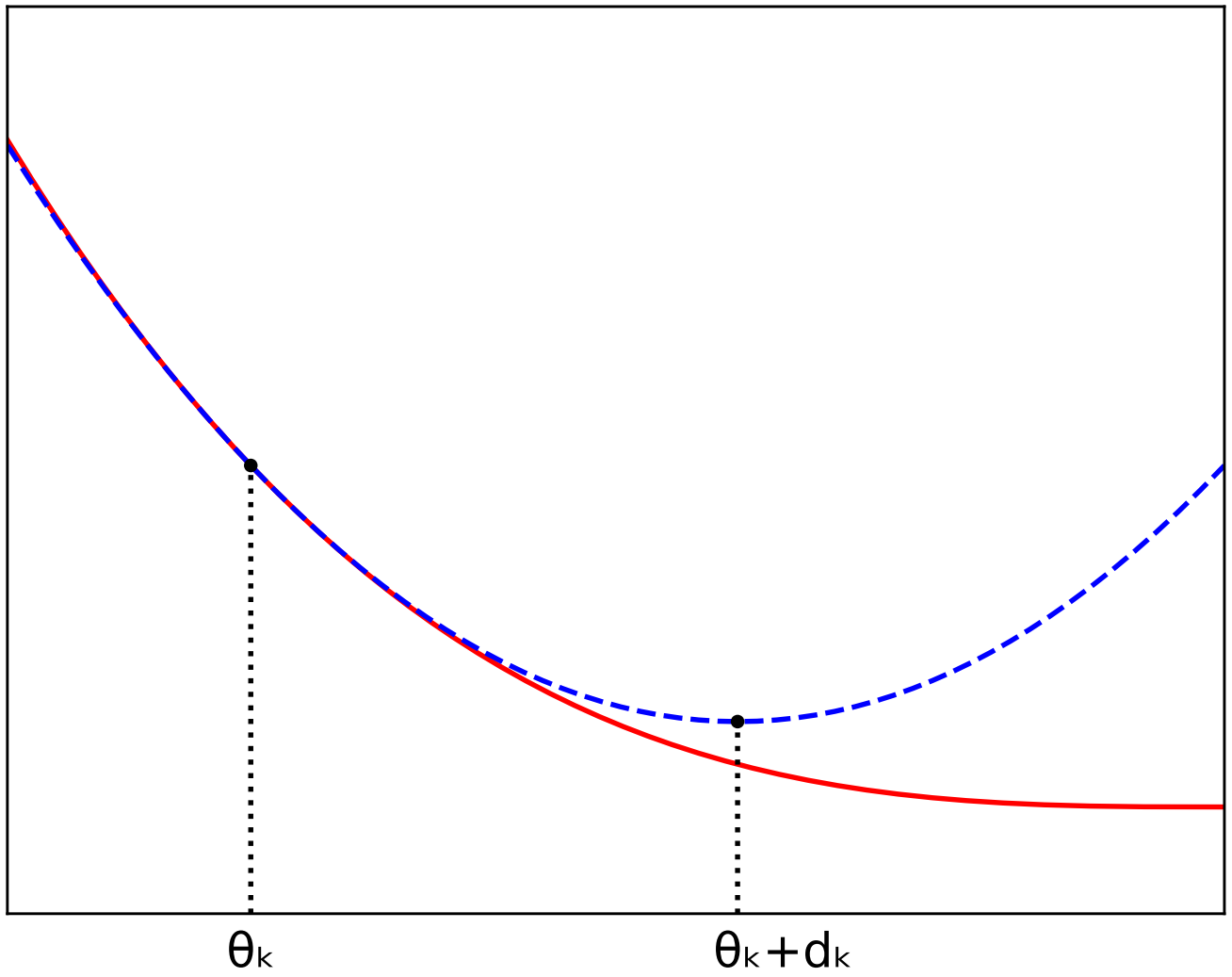
$$\boldsymbol{\theta} = \boldsymbol{\theta}_t - \mathbf{H}_t^{-1}\boldsymbol{g}_t \tag{58}$$

- Así pues, si la aproximación cuadrática es buena, debemos escoger la dirección de descenso:

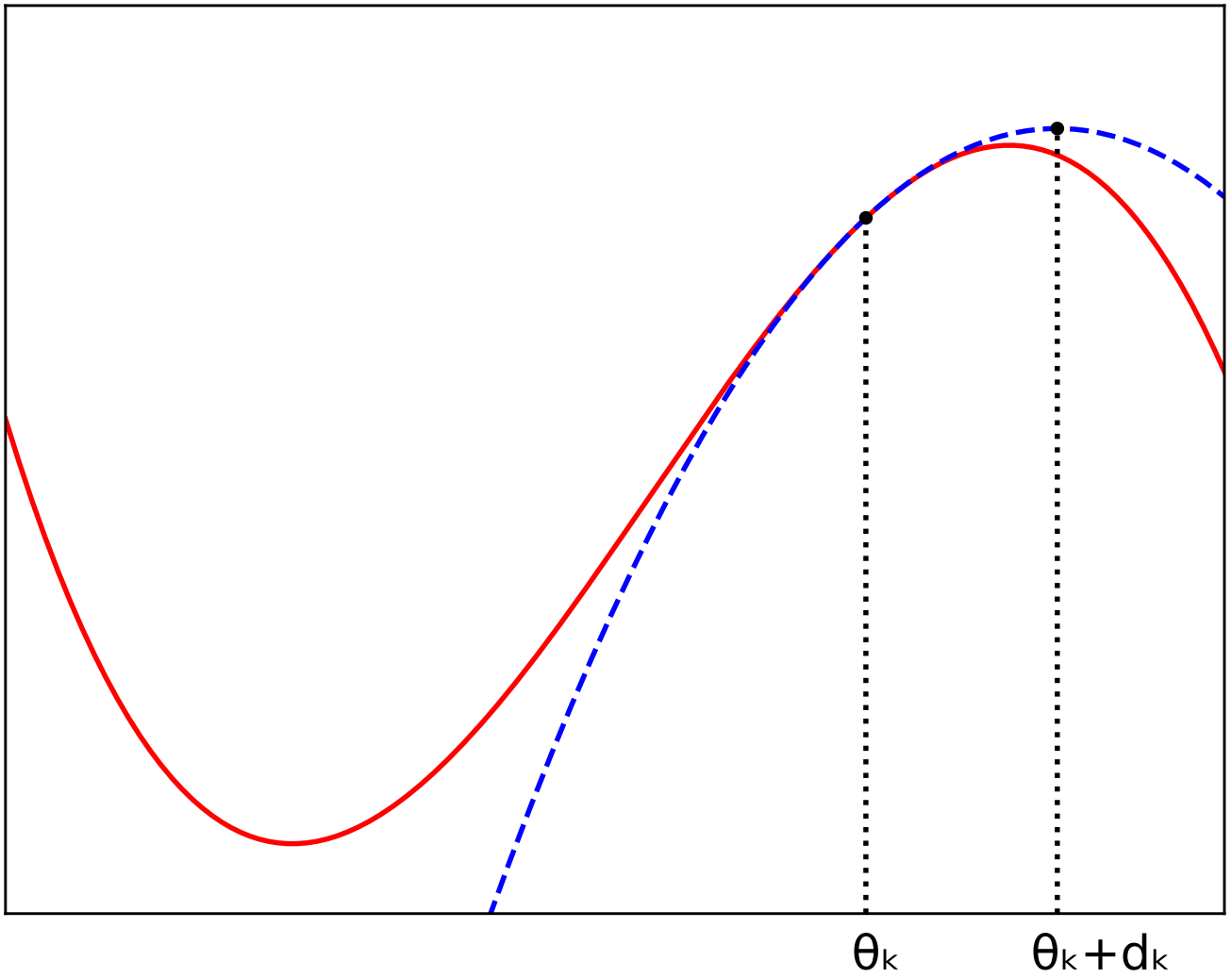
$$\boldsymbol{d}_t = -\mathbf{H}_t^{-1}\boldsymbol{g}_t \tag{59}$$

- Aunque Newton estándar usa $\eta_t = 1$, conviene emplear búsqueda lineal para hallar el mejor tamaño de paso

- *Ejemplo:* Newton para minimizar una función 1d
  - La curva sólida es $\mathcal{L}(x)$ y la discontinua su aproximación de segundo orden, $\mathcal{L}_{\text{quad}}(\theta)$, en θ_k; el paso de Newton, d_k, es lo que añadimos a θ_k para minimizar $\mathcal{L}_{\text{quad}}(\theta)$



- *Ejemplo:* Newton aplicado a una función no convexa
  - Ajustamos una función cuadrática alrededor de θ_k y nos movemos a un punto estacionario $\theta_k + d_k$ que es un máximo local de f, no un mínimo (ya que la Hessiana no es definida positiva)



8.3.2. BFGS y otros métodos quasi-Newton

- ***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):*** método popular que, a partir de una estimación inicial, $\mathbf{B}_0$, típicamente $\mathbf{B}_0 = \mathbf{I}$, aplica actualizaciones sucesivas de rango dos:

$$\mathbf{B}_{t+1} = \mathbf{B}_t + \frac{\boldsymbol{y}_t \boldsymbol{y}_t^t}{\boldsymbol{y}_t^t \boldsymbol{s}_t} - \frac{(\mathbf{B}_t \boldsymbol{s}_t)(\mathbf{B}_t \boldsymbol{s}_t)^t}{\boldsymbol{s}_t^t \mathbf{B}_t \boldsymbol{s}_t} \tag{60}$$

donde

$$\boldsymbol{s}_t = \boldsymbol{\theta}_t - \boldsymbol{\theta}_{t-1} \tag{61}$$

$$\boldsymbol{y}_y = \boldsymbol{g}_t - \boldsymbol{g}_{t-1} \tag{62}$$

- ***Condiciones de Wolfe:*** $\mathbf{B}_{t+1}$ es definida positiva si $\mathbf{B}_0$ lo es y η_t se escoge con búsqueda lineal cumpliendo ***Armijo***

$$\mathcal{L}(\boldsymbol{\theta}_t + \eta\, \boldsymbol{d}_t) \leq \mathcal{L}(\boldsymbol{\theta}_t) + c\,\eta\, \boldsymbol{d}_t^t \nabla \mathcal{L}(\boldsymbol{\theta}_t) \tag{63}$$

así como la condición de curvatura

$$\nabla \mathcal{L}(\boldsymbol{\theta}_t + \eta_t \boldsymbol{d}_t) \geq c_2\, \eta_t\, \boldsymbol{d}_t^t\, \nabla \mathcal{L}(\boldsymbol{\theta}_t) \qquad (0 < c < c_2 < 1) \tag{64}$$

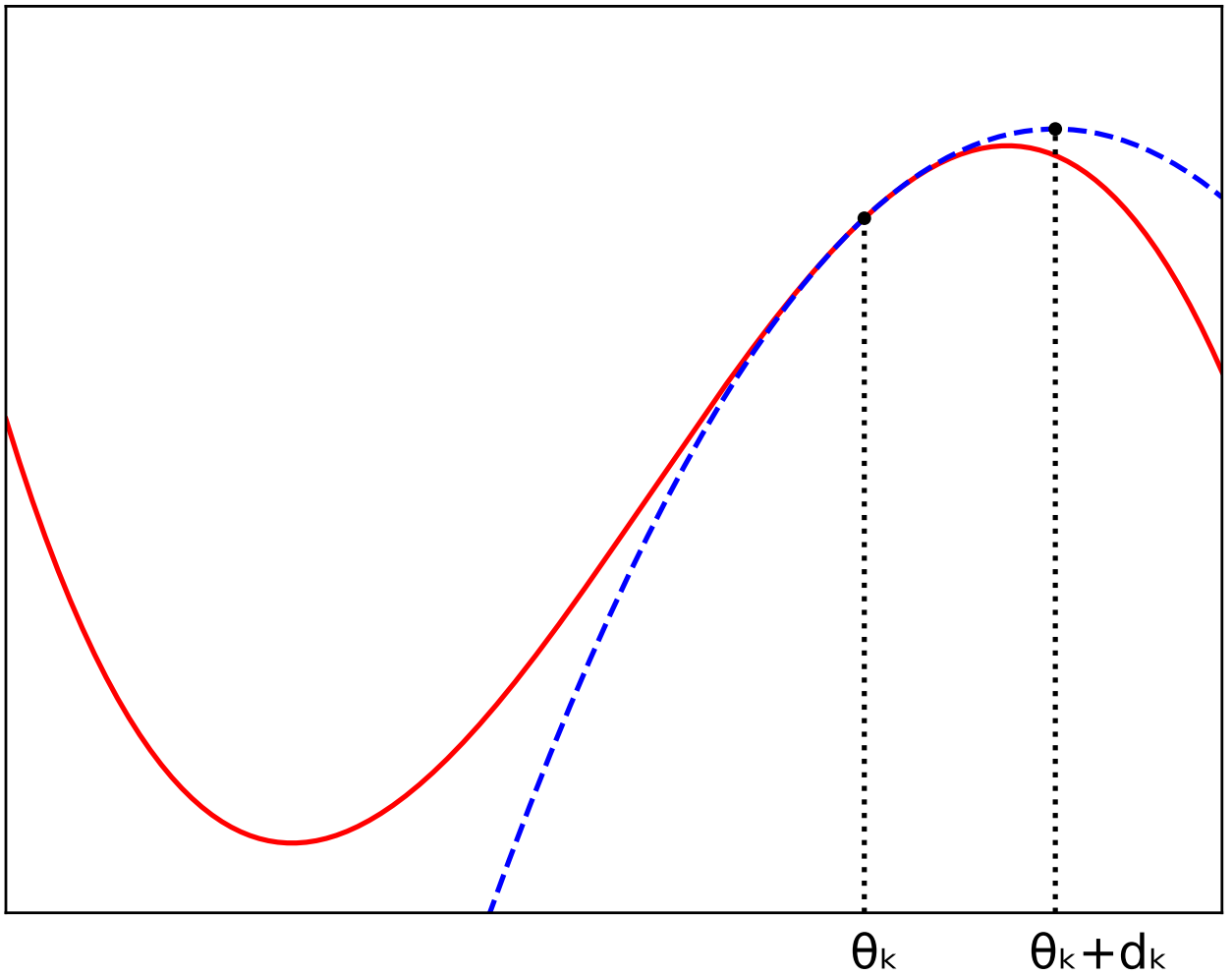
- Alternativamente, BFGS puede actualizar iterativamente una aproximación a la inversa de la Hessiana, $\mathbf{C}_t \approx \mathbf{H}^{-1}$ con

$$\mathbf{C}_{t+1} = \left(\mathbf{I} - \frac{\boldsymbol{s}_t \boldsymbol{y}_t^t}{\boldsymbol{y}_t^t \boldsymbol{s}_t}\right) \mathbf{C}_t \left(\mathbf{I} - \frac{\boldsymbol{y}_t \boldsymbol{s}_t^t}{\boldsymbol{y}_t^t \boldsymbol{s}_t}\right) + \frac{\boldsymbol{s}_t \boldsymbol{s}_t^t}{\boldsymbol{y}_t^t \boldsymbol{s}_t} \tag{65}$$

- ***Limited memory BFGS (L-BFGS):*** no almacena $\mathbf{B}_t$ explícitamente, sino los M ($5 \leq M \leq 20$) pares $(\boldsymbol{s}_t, \boldsymbol{y}_t)$ más recientes para aproximar $\mathbf{H}_t^{-1}\boldsymbol{g}_t$ con una secuencia de productos escalares

8.3.3. Métodos en regiones de confianza

- Si el objetivo no es convexo, la Hessiana puede no ser definida positiva y $\boldsymbol{d}_t = -\mathbf{H}_t^{-1}\boldsymbol{g}_t$ puede no ser una dirección de descenso



▸ *Optimización en región de confianza:*

- Si la aproximación cuadrática de Newton no es válida, no es buena idea fijar una dirección y hallar una distancia óptima

- Alternativamente, podemos asumir que hay una ***región de confianza*** $\mathcal{R}_t$ alrededor de $\boldsymbol{\theta}_t$ donde el objetivo se aproxima bien con un modelo $M_t(\boldsymbol{\delta})$, $\boldsymbol{\delta} = \boldsymbol{\theta} - \boldsymbol{\theta}_t$, y hallar una dirección óptima en $\mathcal{R}_t$

$$\boldsymbol{\delta}_t^* = \arg\min_{\boldsymbol{\delta} \in \mathcal{R}_t} M_t(\boldsymbol{\delta}) \tag{66}$$

- Usualmente, asumimos que el modelo es cuadrático

$$M_t(\boldsymbol{\delta}) = \mathcal{L}(\boldsymbol{\theta}_t) + \boldsymbol{g}_t^t \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^t \mathbf{H}_t \boldsymbol{\delta} \tag{67}$$

con gradiente $\boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_t}$ y Hessiana $\mathbf{H}_t = \nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})|_{\boldsymbol{\theta}_t}$

- Además, se suele asumir que $\mathcal{R}_t$ es una bola de radio r,

$$\mathcal{R}_t = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r\} \tag{68}$$

▷ Así, el problema restringido se convierte en uno no restringido:

$$\boldsymbol{\delta}_t^* = \arg\min_{\boldsymbol{\delta}} \; M(\boldsymbol{\delta}) + \lambda \|\boldsymbol{\delta}\|_2^2 \tag{69}$$

$$= \arg\min_{\boldsymbol{\delta}} \; \boldsymbol{g}^t \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^t (\mathbf{H} + \lambda \mathbf{I}) \boldsymbol{\delta} \tag{70}$$

donde $\lambda > 0$ es un multiplicador de Langrange dependiente de r

▷ ***Regularización de Tikhonov:*** resuelve el problema con

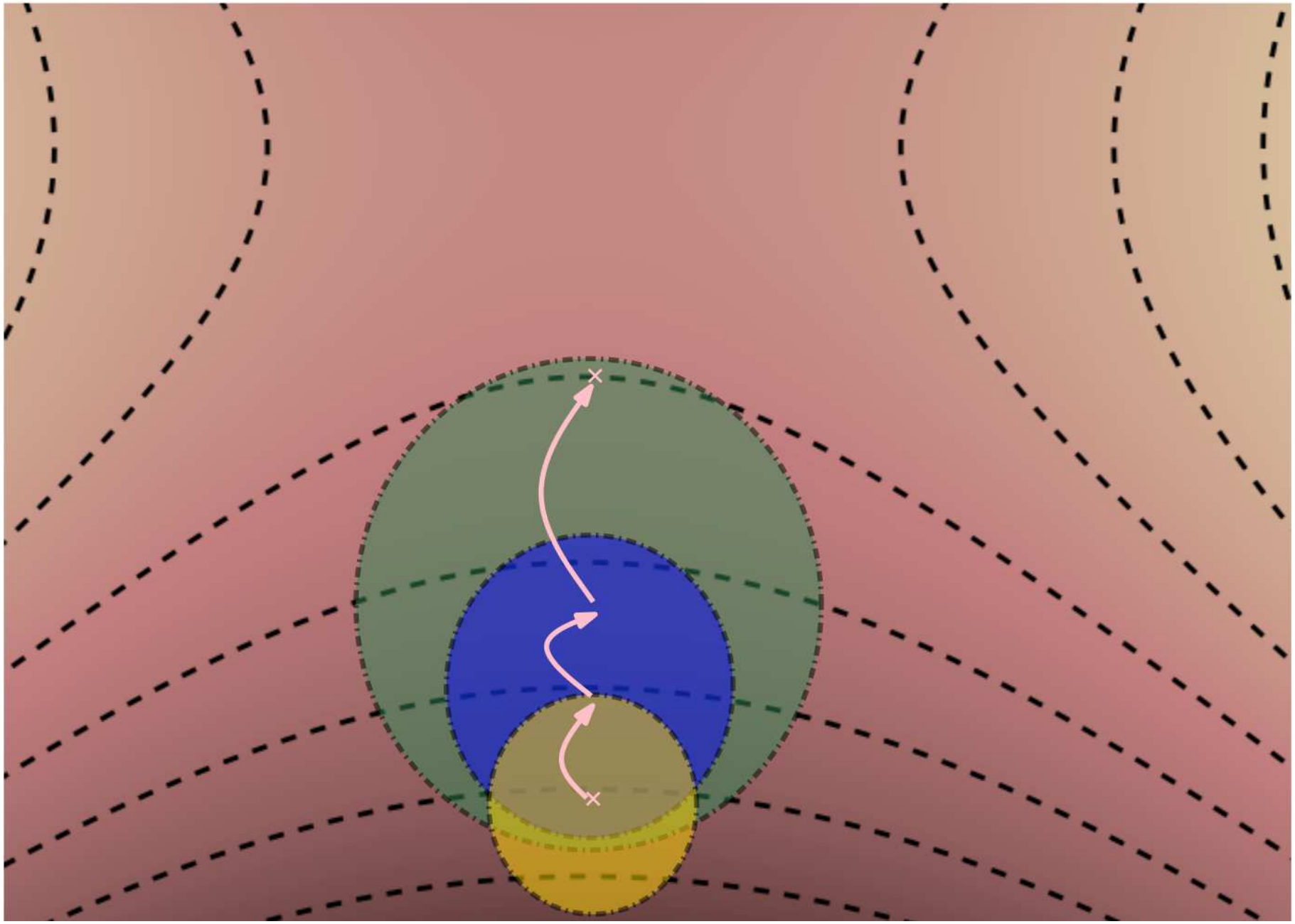
$$\boldsymbol{\delta}^* = -(\mathbf{H} + \lambda \mathbf{I})^{-1} \boldsymbol{g}_t \tag{71}$$

↦ La adición de un $\lambda \mathbf{I}$ suficientemente grande a $\mathbf{H}$ garantiza que la matriz resultante sea definida positiva

↦ Con $\lambda \to 0$ converge a Newton

↦ Con un λ suficientemente grande, positiviza los valores propios negativos (y hace λ los nulos)

- Obj.: líneas discontinuas; aproximaciones cuadráticas: círculos



8.4. Descenso por gradiente estocástico

- ***Optimización estocástica:*** minimizamos el valor esperado del objetivo con respecto a cierta variable aleatoria $\boldsymbol{z}$ añadida

$$\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{q(\boldsymbol{z})}[\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z})] \tag{72}$$

- ***Descenso por gradiente estocástico (SGD):***
  - En la iteración t observamos $\mathcal{L}_t(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{z}_t)$, con $\boldsymbol{z}_t \sim q$
  - Podemos hallar un estimador insesgado del gradiente de $\mathcal{L}$
  - Si $q(\boldsymbol{z})$ no depende de $\boldsymbol{\theta}$, podemos usar $\boldsymbol{g}_t = \nabla_{\boldsymbol{\theta}} \mathcal{L}_t(\boldsymbol{\theta}_t)$
  - ***Stochastic gradient descent (SGD):***

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \nabla \mathcal{L}(\boldsymbol{\theta}_t, \boldsymbol{z}_t) = \boldsymbol{\theta}_t - \eta_t \boldsymbol{g}_t \tag{73}$$

  - Si $\boldsymbol{g}_t$ es insesgado, converge a un punto estacionario

8.4.1. Aplicación a problemas de sumas finitas

- ***Problema de sumas finitas:*** p.e. minimizar el riesgo empírico

$$\mathcal{L}(\boldsymbol{\theta}_t) = \frac{1}{N}\sum_{n=1}^{N} \ell(\boldsymbol{y}_n, f(\boldsymbol{x}_n; \boldsymbol{\theta}_t)) = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n(\boldsymbol{\theta}_t) \tag{74}$$

- Gradiente del riesgo empírico:

$$\boldsymbol{g}_t = \frac{1}{N}\sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta}_t) = \frac{1}{N}\sum_{n=1}^{N} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{y}_n, f(\boldsymbol{x}_n; \boldsymbol{\theta}_t)) \tag{75}$$

- ***Minibatch:*** $\mathcal{B}_t$, muestreo de $B \ll N$ muestras en la iteración t, a partir del cual obtenemos un estimador insesgado de $\boldsymbol{g}_t$

$$\boldsymbol{g}_t \approx \frac{1}{|\mathcal{B}_t|}\sum_{n\in\mathcal{B}_t} \nabla_{\boldsymbol{\theta}} \mathcal{L}_n(\boldsymbol{\theta}_t) = \frac{1}{|\mathcal{B}_t|}\sum_{n\in\mathcal{B}_t} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{y}_n, f(\boldsymbol{x}_n; \boldsymbol{\theta}_t)) \tag{76}$$

8.4.2. Ejemplo: SGD para ajustar regresión lineal

- Objetivo de regresión lineal:

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{n=1}^{N}(\boldsymbol{x}_n^t\boldsymbol{\theta} - y_n)^2 = \frac{1}{2N}\|\mathbf{X}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2 \tag{77}$$

- Gradiente:

$$\boldsymbol{g}_t = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{\theta}_t^t\boldsymbol{x}_n - y_n)\boldsymbol{x}_n \tag{78}$$

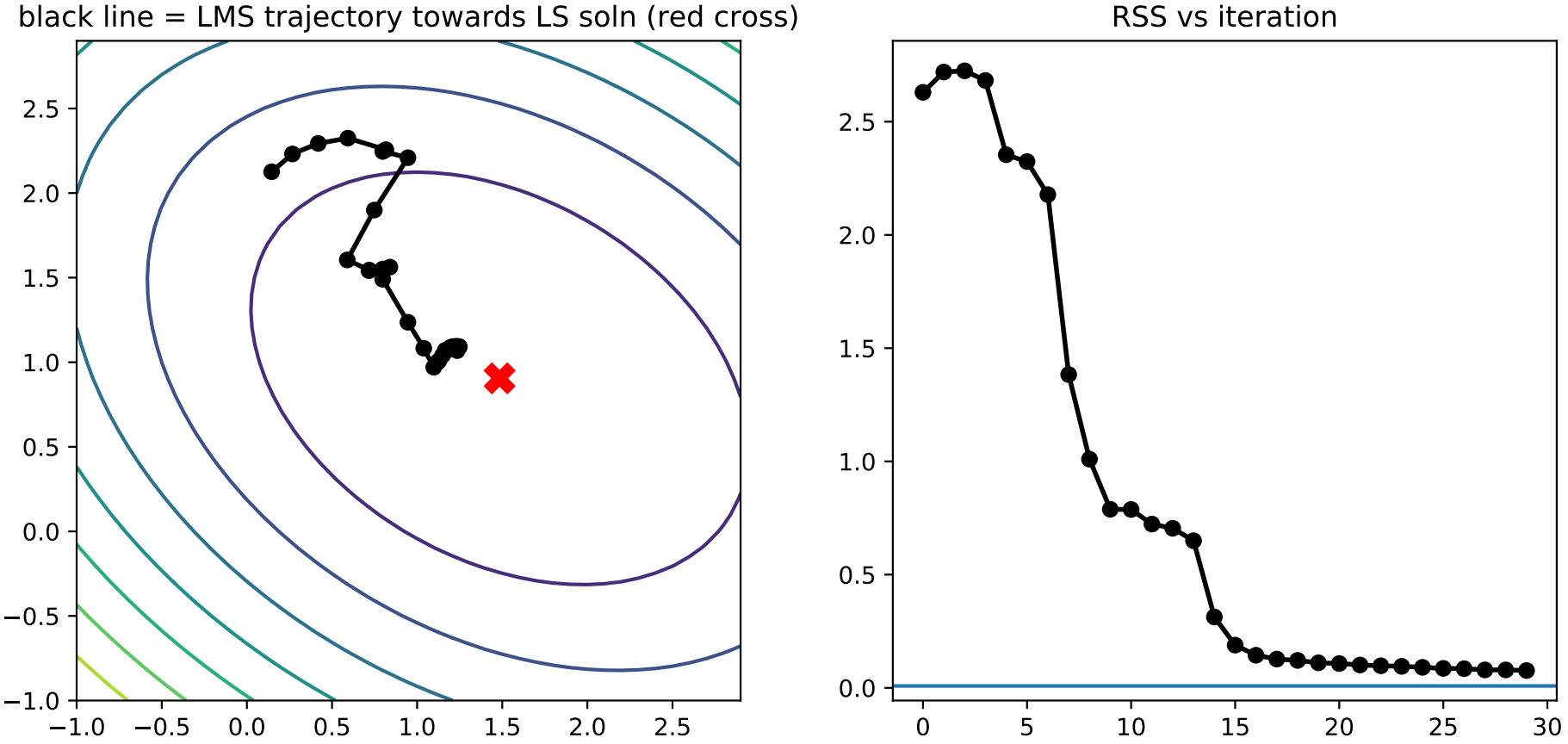
- SGD con un minibatch de talla $B = 1$:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t(\boldsymbol{\theta}_t^t\boldsymbol{x}_n - y_n)\boldsymbol{x}_n \tag{79}$$

  donde n es el índice de la muestra escogida en la iteración t

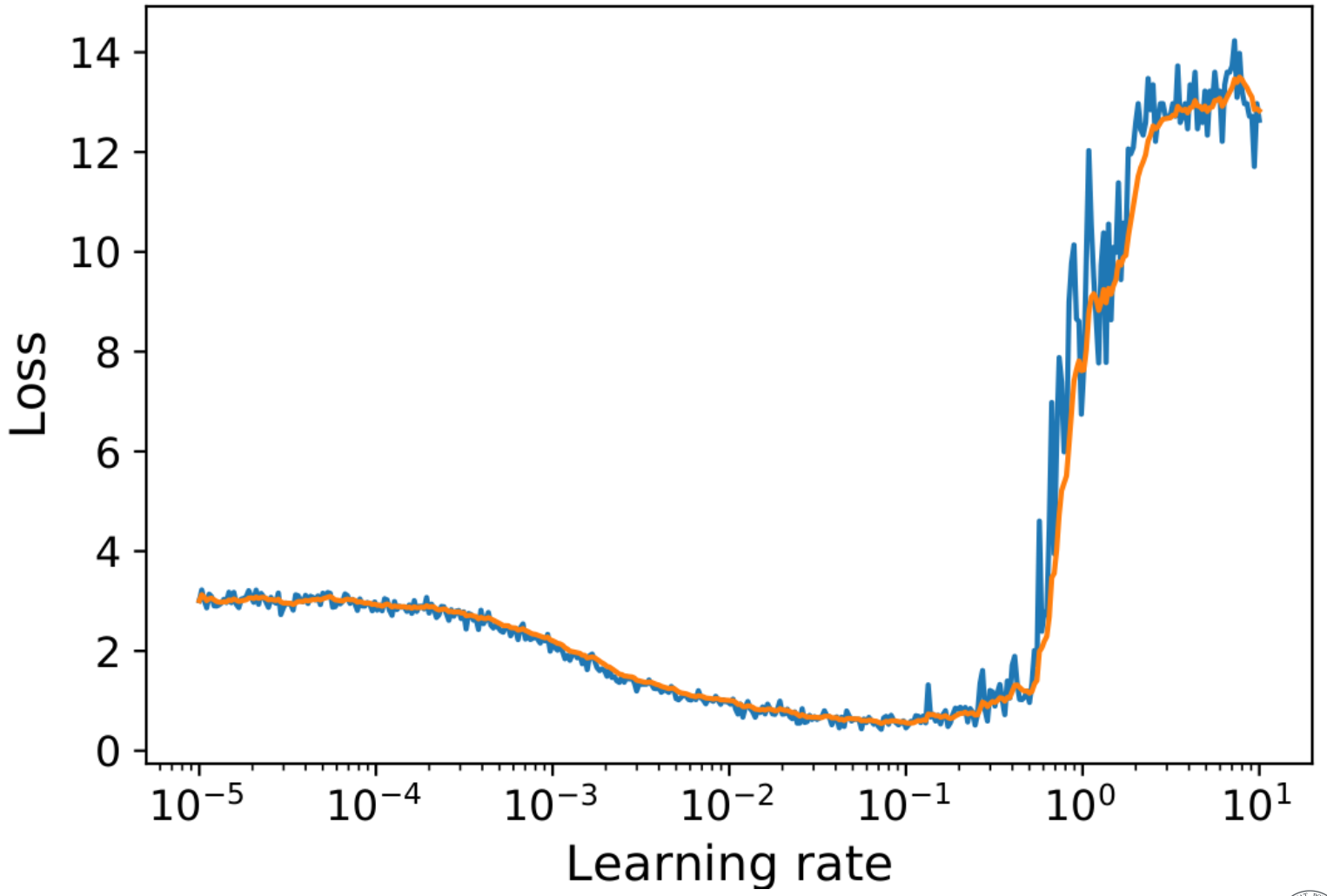
- El algoritmo completo se conoce como ***mínimos cuadrados (LMS, least mean squares)***, ***regla delta*** o ***Widrow-Hoff***

- *Ejemplo:* empezamos en $\boldsymbol{\theta}_0 = (-0.5, 2)$ y SGD converge en unas 26 iteraciones ($\|\boldsymbol{\theta}_{26} - \boldsymbol{\theta}_{25}\|_2^2 < 0.01$)



8.4.3. Elección del tamaño de paso

- Efecto del tamaño de paso en la pérdida de una red neuronal entrenada con SGD en FashionMNIST: la red queda subajustada si es pequeño; si es grande, la red se inestabiliza y desajusta



- Para escoger un tamaño de paso adecuado, podemos empezar con un valor pequeño e incrementarlo paulatinamente, evaluando cada valor probado con unos pocos minibatches; finalmente nos quedamos con el valor de menor pérdida, o mejor un poco menor

- En lugar de escoger un tamaño de paso fijo, podemos usar una secuencia de tamaños de paso ajustados a lo largo del tiempo

- ***Condiciones de Robbins-Monro:*** condición suficiente para garantizar la convergencia de SGD en función de $\{\eta_t\}$

$$\eta_t \to 0 \tag{80}$$

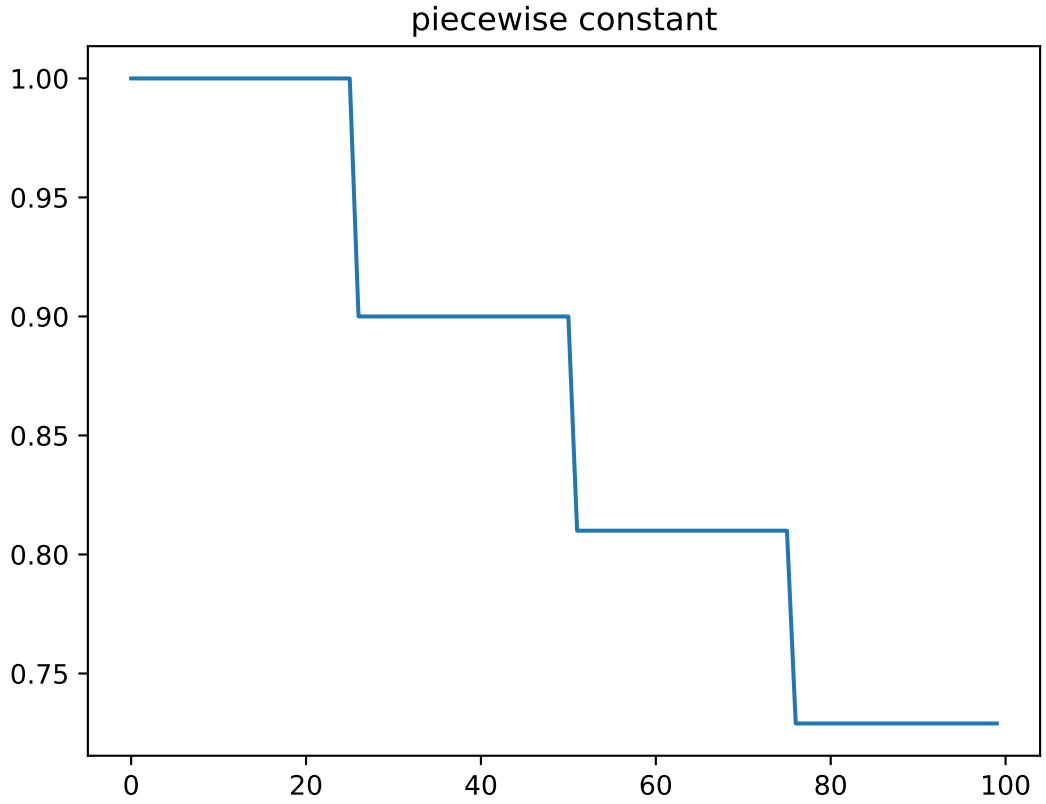
y

$$\frac{\sum_{t=1}^{\infty} \eta_t^2}{\sum_{t=1}^{\infty} \eta_t} \to 0 \tag{81}$$

- ***Caída escalonada:*** el factor de aprendizaje se reajusta en cada punto (umbral) temporal de un conjunto predefinido

$$\eta_t = \eta_i \qquad \text{si} \qquad t_i \leq t \leq t_{i+1} \tag{82}$$

  - *Ejemplo:* $\eta_i = \eta_0 \gamma^i$ con $\eta_0 = 1$ y $\gamma = 0.9$ cada 20 iteraciones

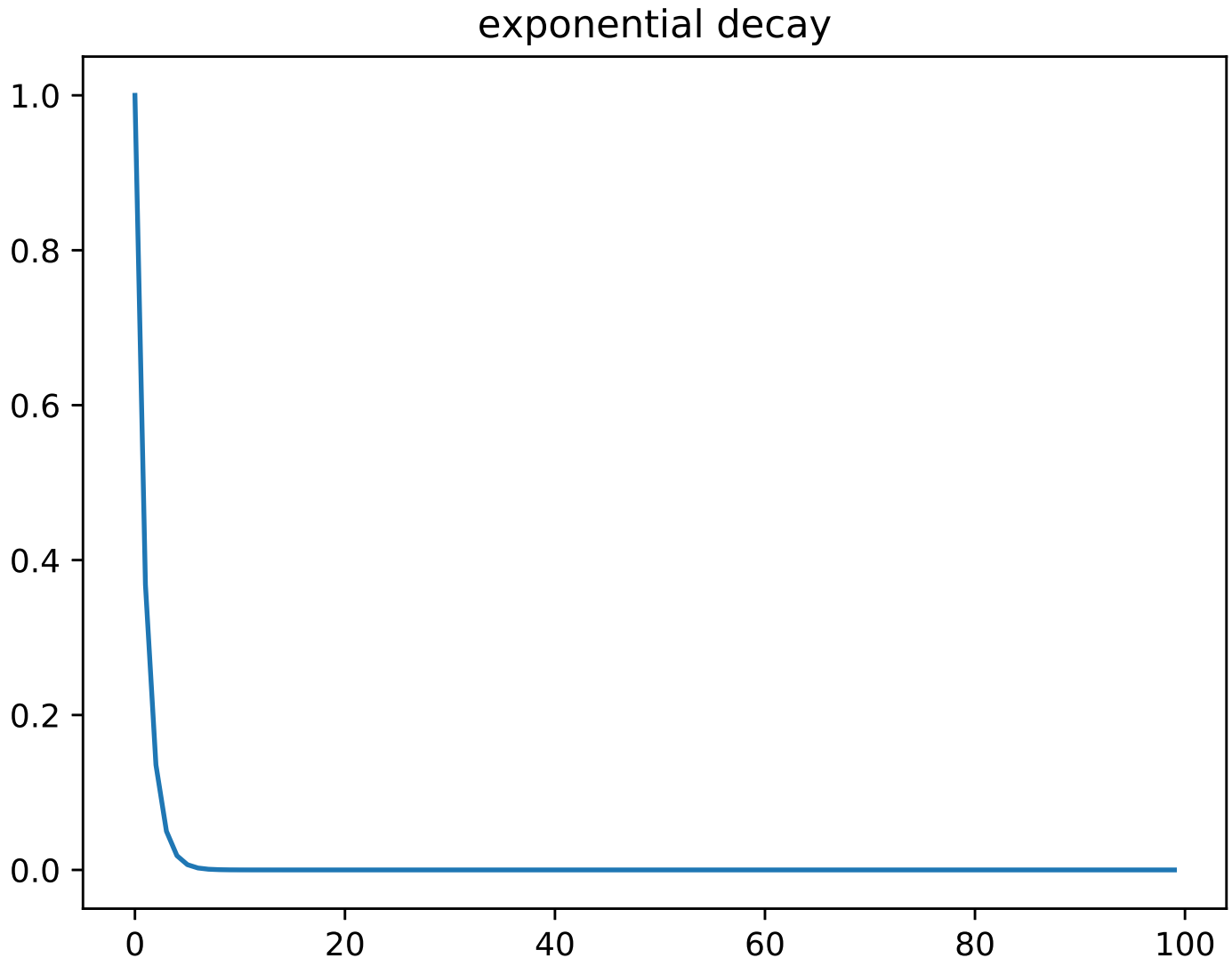


  - ***Reduce-on-plateau:*** los umbrales de tiempo se establecen dinámicamente, cuando la pérdida en train o validación se estanca

▶ *Caída exponencial:*

$$\eta_t = \eta_0 e^{-\lambda t} \tag{83}$$

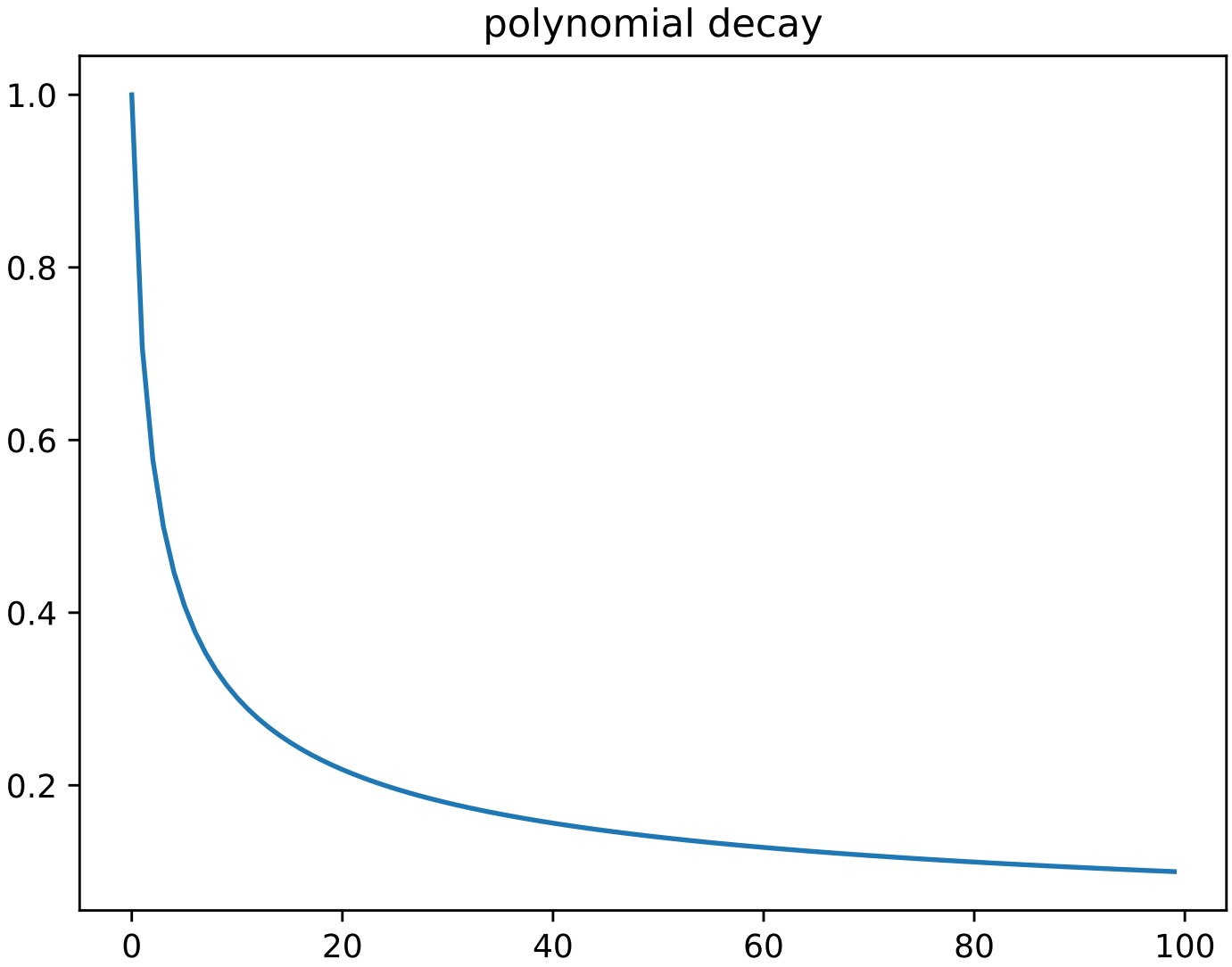
▷ Caída exponencial es típicamente demasiado rápida:



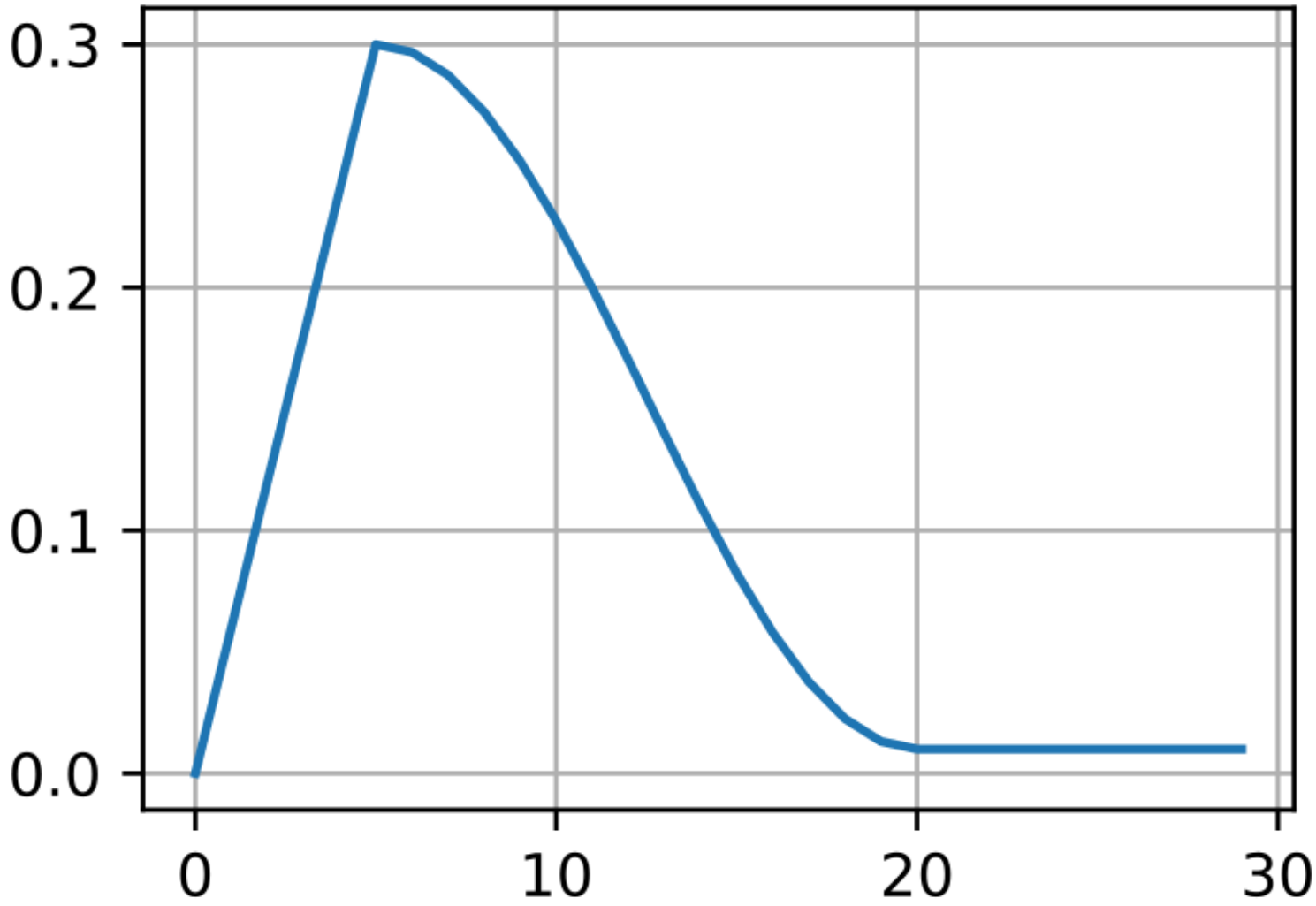
- ***Caída polinómica:***

$$\eta_t = \eta_0(\beta t + 1)^{-\alpha} \tag{84}$$

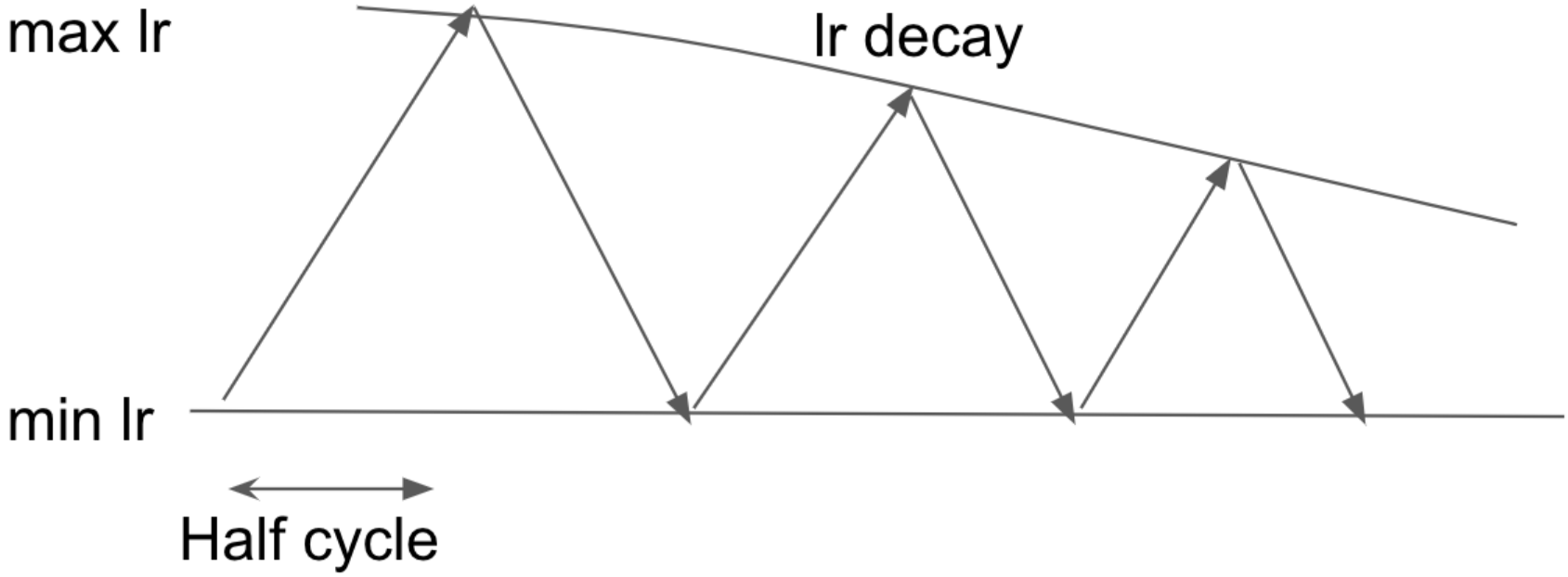
  - ***Square-root schedule:*** $\alpha = 0.5 \quad \beta = 1 \quad \rightsquigarrow \quad \eta_t = \eta_0 \frac{1}{\sqrt{t+1}}$



- ***Calentamiento del factor de aprendizaje:*** inicialmente, incrementamos rápidamente el factor de aprendizaje para reducirlo a continuación gradualmente (y que quede en una región “plana”)
  - *Ejemplo:* calentamiento lineal y enfriamiento coseno



- ***Factor de aprendizaje cíclico:*** lo incrementamos y decrementamos múltiples veces para escapar de mínimos locales
  - *Ejemplo:* mínimo y máximo basados en un calentamiento inicial; medio-ciclo en función del número de re-inicios deseado



- ***SGD con re-inicios calientes:*** aprovecha los modelos obtenidos tras cada enfriamiento para construir un modelo ensamble
- ***Búsqueda lineal:*** si la varianza del gradiente es pequeña, podemos hallar el factor de aprendizaje con búsqueda lineal y Armijo

8.4.4. Promediado iterativo

- ***Promediado iterativo*** o ***Polyak-Ruppert:*** para reducir la varianza de los estimadores producidos por SGD

$$\bar{\boldsymbol{\theta}}_t = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{\theta}_i = \frac{1}{t}\boldsymbol{\theta}_t + \frac{t-1}{t}\bar{\boldsymbol{\theta}}_{t-1} \tag{85}$$

- Ratio de convergencia asintótico óptimo entre algoritmos SGD
- En regresión lineal, equivale a regularización ℓ_2
- ***Stochastic Weight Averaging (SWA):*** variante orientada a la búsqueda de soluciones que generalicen mejor

8.4.6. SGD precondicionado

- ***SGD precondicionado:*** introduce una ***matriz de precondicionado*** o ***precondicionador*** $\mathbf{M}_t$, típicamente definida positiva:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta_t \mathbf{M}_t^{-1} \boldsymbol{g}_t \tag{86}$$

- Por simplicidad computacional, se suelen escoger preacondicionares diagonales, posiblemente sin información de segundo orden

8.4.6.1. AdaGrad

- **AdaGrad *(adaptive gradient):*** propuesto para objetivos convexos y gradientes con muchos elementos nulos

$$\theta_{t+1,d} = \theta_{t,d} - \eta_t \frac{1}{\sqrt{s_{t,d} + \epsilon}} g_{t,d} \tag{87}$$

donde $\epsilon > 0$ y $s_{t,d}$ es la suma de gradientes cuadráticos

$$s_{t,d} = \sum_{i=1}^{t} g_{i,d}^2 \tag{88}$$

- En notación vectorial:

$$\Delta \boldsymbol{\theta}_t = -\eta_t \frac{1}{\sqrt{\boldsymbol{s}_t + \epsilon}} \boldsymbol{g}_t \tag{89}$$

- Visto como SGD precondicionado, equivale a tomar:

$$\mathbf{M}_t = \mathrm{diag}(\boldsymbol{s}_t + \epsilon) \tag{90}$$

- Así, resulta un ejemplo de ***factor de aprendizaje adaptativo***, si bien el η_t global también debe fijarse, típicamente con $\eta_t = \eta_0$

8.4.6.2. RMSPROP y ADADELTA

- **RMSPROP:** en lugar de sumar gradientes cuadráticos pasados, emplea una media móvil ponderada exponencialmente (EWMA):

$$s_{t+1,d} = \beta s_{t,d} + (1-\beta) g_{t,d}^2 \tag{91}$$

con $\beta \approx 0.9$ para dar más peso a los gradientes recientes

- RMS viene de *"root mean square"* ya que:

$$\sqrt{s_{t,d}} \approx \mathrm{RMS}(\boldsymbol{g}_{1:t,d}) = \sqrt{\frac{1}{t}\sum_{\tau=1}^{t} g_{\tau,d}^2} \tag{92}$$

- La actualización RMSPROP es como la de ADAGRAD:

$$\Delta\boldsymbol{\theta}_t = -\eta_t \frac{1}{\sqrt{\boldsymbol{s}_t + \epsilon}} \boldsymbol{g}_t \tag{93}$$

- **ADADELTA:** variante de RMSPROP que también incluye una EWMA de las actualizaciones $\boldsymbol{\delta}_t$ como sigue:

$$\Delta\boldsymbol{\theta}_t = -\eta_t \frac{\sqrt{\boldsymbol{\delta}_{t-1} + \epsilon}}{\sqrt{\boldsymbol{s}_t + \epsilon}} \boldsymbol{g}_t \tag{94}$$

donde

$$\boldsymbol{\delta}_t = \beta \boldsymbol{\delta}_{t-1} + (1-\beta)(\Delta\boldsymbol{\theta}_t)^2 \tag{95}$$

- ADADELTA elimina, hasta cierto punto, la necesidad de ajustar el factor de aprendizaje η_t, por lo que se suele fijar a uno

8.4.6.3. ADAM

- **ADAM *("adaptative moment estimation"):*** calcula una EWMA de gradientes (momentum) y gradientes cuadráticos (RMSPROP)

$$\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\boldsymbol{g}_t \tag{96}$$

$$\boldsymbol{s}_t = \beta_2 \boldsymbol{s}_{t-1} + (1-\beta_2)\boldsymbol{g}_t^2 \tag{97}$$

  y lleva a cabo la actualización:

$$\Delta\boldsymbol{\theta}_t = -\eta_t \frac{1}{\sqrt{\boldsymbol{s}_t} + \epsilon}\boldsymbol{m}_t \tag{98}$$

- Valores estándar: $\beta_1 = 0.9$, $\beta_2 = 0.999$ y $\epsilon = 10^{-6}$
- Suele tomarse $\eta_t = 0.001$

- Con $\boldsymbol{m}_0 = \boldsymbol{s}_0 = \mathbf{0}$, los estimadores iniciales están sesgados hacia valores pequeños, por lo que se recomienda corregir sesgos con:

$$\hat{\boldsymbol{m}}_t = \frac{\boldsymbol{m}_t}{1 - \beta_1^t} \tag{99}$$

$$\hat{\boldsymbol{s}}_t = \frac{\boldsymbol{s}_t}{1 - \beta_2^t} \tag{100}$$

- Con $\beta_1 = 0$ y sin corrección de sesgo, coincide con RMSPROP

8.4.6.4. Problemas con los factores adaptativos

- ***Factor de aprendizaje adaptativo:*** factor de aprendizaje que varía con el tiempo, especialmente cuando empleamos métodos de escalado diagonal, del tipo $\eta_0 \mathbf{M}_t^{-1}$
- Un problema común es el de ajustar bien el factor base, η_0
- Los métodos EWMA en un contexto estocástico producen gradientes ruidosos, por lo que otro problema común es la posible no convergencia incluso en problemas convexos
- Debido a la importancia de estas técnicas, se vienen proponiendo diversas soluciones para abordar estos problemas como, por ejemplo, AMSGRAD, PADAM y YOGI

8.4.6.5. Matrices de precondicionado no diagonales

- La aceleración de SGD con matrices de precondicionado diagonales puede no ser efectiva con parámetros muy correlacionados
- ***Full-matrix Adagrad:*** usa la matriz de precondicionado

$$\mathbf{M}_t = \left[(\mathbf{G}_t \mathbf{G}_t^t)^{\frac{1}{2}} + \epsilon \mathbf{I}_D \right]^{-1} \tag{101}$$

donde

$$\mathbf{G}_t = [\boldsymbol{g}_t, \ldots, \boldsymbol{g}_1] \tag{102}$$

siendo $\boldsymbol{g}_i = \nabla_{\boldsymbol{\phi}} c(\boldsymbol{\phi}_i)$ es el gradiente en el paso i

- ***Algoritmo shampoo:*** aproxima $\mathbf{M}$ con una matriz diagonal por bloques y explota su estructura para invertirla eficientemente

8.5. Optimización con restricciones

- ***Optimización con restricciones:***

$$\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \mathcal{L}(\boldsymbol{\theta}) \tag{103}$$

donde $\mathcal{C} \subseteq \Theta$ es el ***conjunto de soluciones posibles:***

$$\mathcal{C} = \{\boldsymbol{\theta} : \ g_j(\boldsymbol{\theta}) \leq 0 : j \in \mathcal{I}, \ h_k(\boldsymbol{\theta}) = 0 : k \in \mathcal{E}\} \in \mathbb{R}^D \tag{104}$$

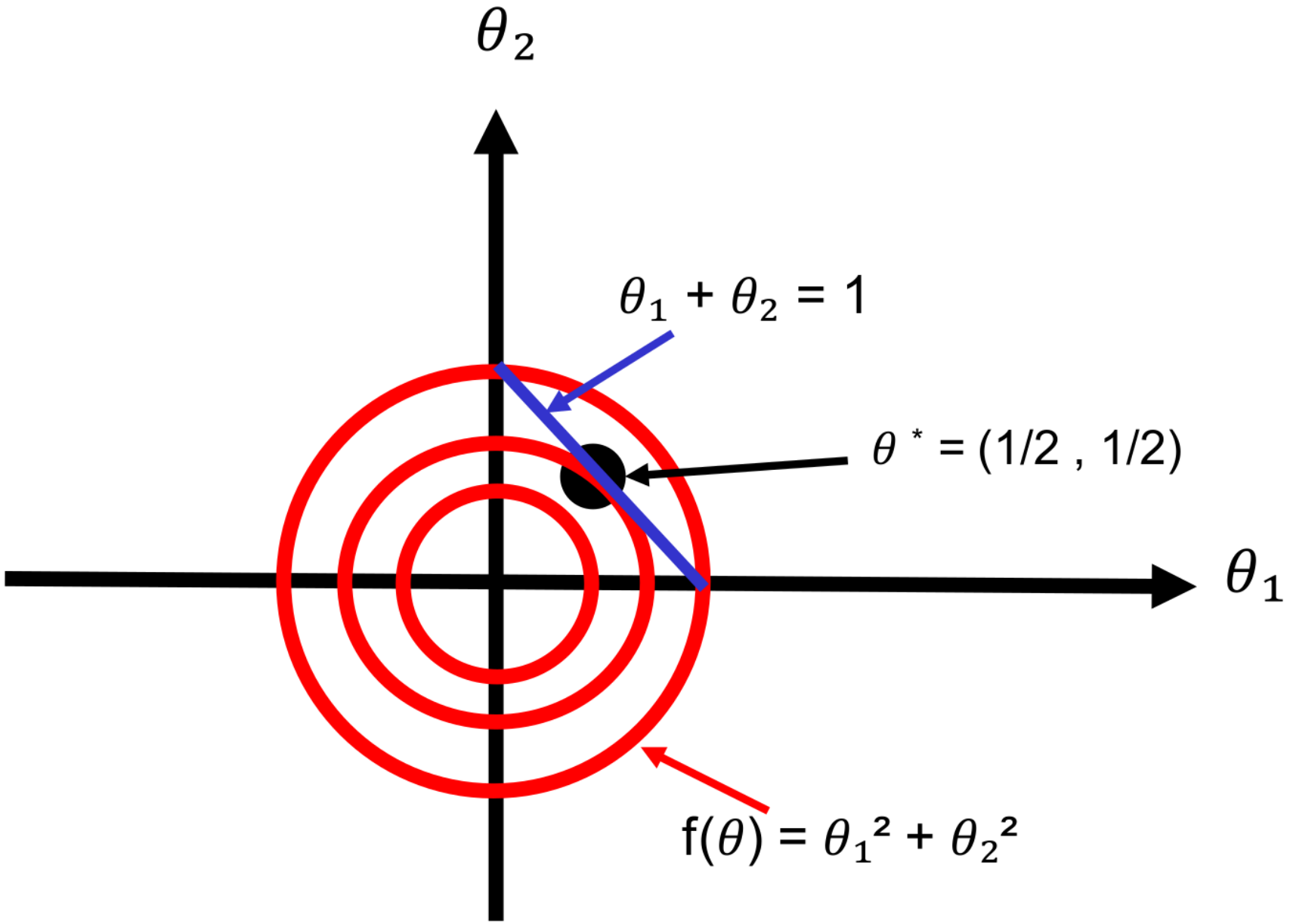
con las ***restricciones de desigualdad***

$$g_j(\boldsymbol{\theta}) \leq 0 : j \in \mathcal{I} \tag{105}$$

y las ***restricciones de igualdad***

$$h_k(\boldsymbol{\theta}) = 0 : k \in \mathcal{E} \tag{106}$$

- *Ejemplo:* $\mathcal{L}(\boldsymbol{\theta}) = \theta_1^2 + \theta_2^2, \quad h(\boldsymbol{\theta}) = 1 - \theta_1 - \theta_2 = 0, \quad \boldsymbol{\theta}^* = (0.5, 0.5)$



8.5.1. Multiplicadores de Lagrange

- Asumimos una única restricción y de igualdad, $h(\boldsymbol{\theta}) = 0$
- $\nabla h(\boldsymbol{\theta})$ ***es ortogonal a la superfície de la restricción***
  - Consideremos otro punto cercano en la superfície, $\boldsymbol{\theta} + \boldsymbol{\epsilon}$
  - Si aproximamos h alrededor de $\boldsymbol{\theta}$ con Taylor de primer orden:

$$h(\boldsymbol{\theta} + \boldsymbol{\epsilon}) \approx h(\boldsymbol{\theta}) + \boldsymbol{\epsilon}^t \nabla h(\boldsymbol{\theta}) \tag{107}$$

  - Como $\boldsymbol{\theta}$ y $\boldsymbol{\theta} + \boldsymbol{\epsilon}$ están en la superfície, $h(\boldsymbol{\theta}) = h(\boldsymbol{\theta} + \boldsymbol{\epsilon})$ y

$$\boldsymbol{\epsilon}^t \nabla h(\boldsymbol{\theta}) \approx 0 \tag{108}$$

  - Dado que $\boldsymbol{\epsilon}$ es paralelo a la superfície, $\nabla h(\boldsymbol{\theta})$ es ortogonal

- ***Si θ^* es un punto en la superfície de la restricción que minimiza $\mathcal{L}(\theta)$, existe un $\lambda^* \in \mathbb{R}$ tal que $\nabla\mathcal{L}(\theta^*) = \lambda^*\nabla h(\theta^*)$***

  - Ya hemos visto que $\nabla h(\boldsymbol{\theta}^*)$ es ortogonal a la superfície

  - $\nabla\mathcal{L}(\boldsymbol{\theta})$ es ortogonal a la superfície en $\boldsymbol{\theta}^*$ ya que, si no, $\mathcal{L}(\boldsymbol{\theta})$ decrecería moviéndonos a pequeña distancia por la superfície

  - Como $\nabla h(\boldsymbol{\theta})$ y $\nabla\mathcal{L}(\boldsymbol{\theta})$ son ortogonales a la superfície en $\boldsymbol{\theta}^*$, son paralelos (o anti-paralelos) y existe un $\lambda^* \in \mathbb{R}$ tal que

$$\nabla\mathcal{L}(\boldsymbol{\theta}^*) = \lambda^*\nabla h(\boldsymbol{\theta}^*) \tag{109}$$

- ***Multiplicador de Lagrange:*** es la constante λ^*; puede ser positiva, negativa, o cero si $\nabla\mathcal{L}(\boldsymbol{\theta}^*) = 0$

- ***Lagrangiana:*** convierte (109) en objetivo

$$L(\boldsymbol{\theta}, \lambda) \triangleq \mathcal{L}(\boldsymbol{\theta}) + \lambda h(\boldsymbol{\theta}) \tag{110}$$

- ***Punto crítico:*** punto estacionario de la Lagrangiana

$$\nabla_{\boldsymbol{\theta},\lambda} L(\boldsymbol{\theta}, \lambda) = \mathbf{0} \quad \Leftrightarrow \quad \lambda \nabla_{\boldsymbol{\theta}} h(\boldsymbol{\theta}) = \nabla \mathcal{L}(\boldsymbol{\theta}), \quad h(\boldsymbol{\theta}) = 0 \tag{111}$$

- ***Múltiples restricciones de igualdad:*** $m > 1$ restricciones

$$L(\boldsymbol{\theta}, \boldsymbol{\lambda}) \triangleq \mathcal{L}(\boldsymbol{\theta}) + \sum_{i=1}^{m} \lambda_i h_i(\boldsymbol{\theta}) \tag{112}$$

  - $D + m$ ecuaciones com $D + m$ incógnitas: podemos emplear optimización sin restricciones para hallar un punto estacionario

8.5.1.1. Ejemplo 2D cuadrático con una restricción

- Minimización de $\mathcal{L}(\boldsymbol{\theta}) = \theta_1^2 + \theta_2^2$ sujeto a $\theta_1 + \theta_2 = 1$
- Lagrangiana:

$$L(\theta_1, \theta_2, \lambda) = \theta_1^2 + \theta_2^2 + \lambda(\theta_1 + \theta_2 - 1) \tag{113}$$

- Condiciones para un punto estacionario:

$$\frac{\partial}{\partial \theta_1} L(\theta_1, \theta_2, \lambda) = 2\theta_1 + \lambda = 0 \tag{114}$$

$$\frac{\partial}{\partial \theta_2} L(\theta_1, \theta_2, \lambda) = 2\theta_2 + \lambda = 0 \tag{115}$$

$$\frac{\partial}{\partial \lambda} L(\theta_1, \theta_2, \lambda) = \theta_1 + \theta_2 - 1 = 0 \tag{116}$$

- De las dos primeras obtenemos: $\theta_1 = \theta_2$
- De la tercera: $2\theta_1 = 1$
- $\boldsymbol{\theta}^* = (0.5, 0.5)$ mínimo global (objetivo convexo y restricción afín)

8.5.2. Las condiciones KKT

- ***Caso con una única restricción y de desigualdad,*** $g(\boldsymbol{\theta}) \leq 0$***:***
  - Podemos optimizar (sin restricciones) el objetivo con un término de penalización añadido en forma de función de paso infinito:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \infty \mathbb{1}(g(\boldsymbol{\theta}) > 0) \tag{117}$$

  - Más fácilmente, añadimos una cota inferior $\mu g(\boldsymbol{\theta})$, con $\mu \geq 0$:

$$L(\boldsymbol{\theta}, \mu) = \mathcal{L}(\boldsymbol{\theta}) + \mu g(\boldsymbol{\theta}) \tag{118}$$

    lo que permite recuperar la función de paso infinito como:

$$\hat{\mathcal{L}}(\boldsymbol{\theta}) = \max_{\mu \geq 0} L(\boldsymbol{\theta}, \mu) = \begin{cases} \infty & \text{si } g(\boldsymbol{\theta}) > 0, \\ \mathcal{L}(\boldsymbol{\theta}) & \text{en otro caso} \end{cases} \tag{119}$$

  - Así, nuestro problema de optimización es:

$$\min_{\boldsymbol{\theta}} \max_{\mu \geq 0} L(\boldsymbol{\theta}, \mu) \tag{120}$$

- ***Caso general:***
  - Tenemos múltiples restricciones de desigualdad, $\boldsymbol{g}(\boldsymbol{\theta}) \leq \mathbf{0}$, y de igualdad, $\boldsymbol{h}(\boldsymbol{\theta}) = \mathbf{0}$
  - ***Lagrangiana generalizada:***

$$L(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) = \mathcal{L}(\boldsymbol{\theta}) + \sum_i \mu_i g_i(\boldsymbol{\theta}) + \sum_j \lambda_j h_j(\boldsymbol{\theta}) \tag{121}$$

  - Podemos cambiar $\lambda_j h_j$ por $-\lambda_j h_j$ ya que el signo es arbitrario
  - Nuestro problema de optimización es:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \max_{\boldsymbol{\mu} \geq \mathbf{0}, \boldsymbol{\lambda}} L(\boldsymbol{\theta}, \boldsymbol{\mu}, \boldsymbol{\lambda}) \tag{122}$$

- ***Condiciones de Karush–Kuhn–Tucker (KKT):*** si $\mathcal{L}$ y g son convexas, los puntos críticos satisfacen (bajo ciertas condiciones)
  - ***Estacionaridad:***

$$\nabla \mathcal{L}(\boldsymbol{\theta}^*) + \sum_i \mu_i \nabla g_i(\boldsymbol{\theta}^*) + \sum_j \lambda_j \nabla h_j(\boldsymbol{\theta}^*) = \mathbf{0} \tag{123}$$

  - ***Factibilidad primal:***

$$\boldsymbol{g}(\boldsymbol{\theta}^*) \leq \mathbf{0}, \qquad \boldsymbol{h}(\boldsymbol{\theta}^*) = \mathbf{0} \tag{124}$$

  - ***Factibilidad dual:***

$$\boldsymbol{\mu} \geq \mathbf{0} \tag{125}$$

  - ***Holgura complementaria:***

$$\mu_i \, g_i(\boldsymbol{\theta}^*) = 0 \quad \text{para todo } i \in \mathcal{I} \tag{126}$$

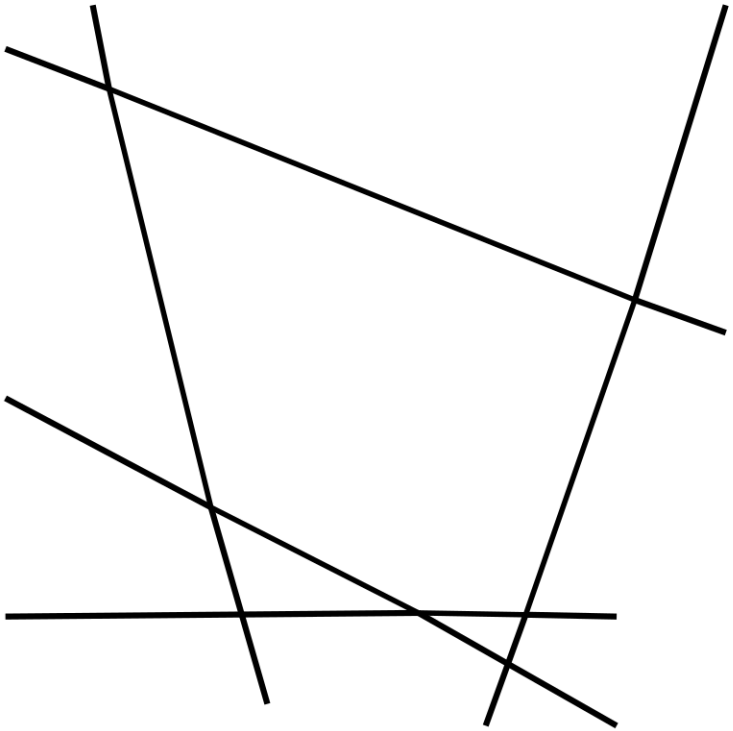
    - ↦ ***Activa:*** si $g_i(\boldsymbol{\theta}^*) = 0$, la solución se halla en la superfície de la restricción y $\nabla \mathcal{L} = \mu_i \nabla g$ para algún $\mu_i \neq 0$
    - ↦ ***Inactiva:*** si $g_i(\boldsymbol{\theta}^*) < 0$, la solución no está en la superfície de la restricción y $\nabla \mathcal{L} = \mu_i \nabla g$ con $\mu_i = 0$

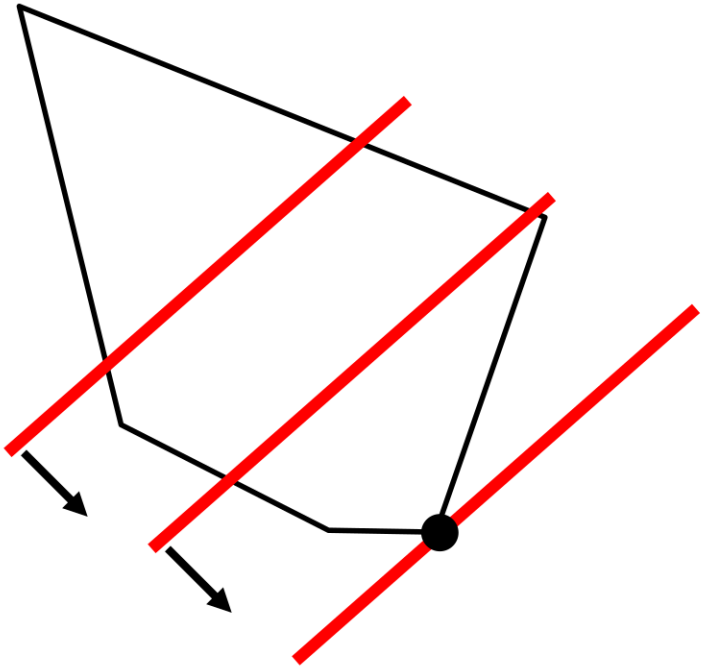
8.5.3. Programación lineal

- ***Forma estándar:*** optimización de objetivo y restricciones lineales

$$\min_{\boldsymbol{\theta}} \ \boldsymbol{c}^t\boldsymbol{\theta} \qquad \text{s.a.} \quad \mathbf{A}\boldsymbol{\theta} \leq \boldsymbol{b}, \quad \boldsymbol{\theta} \geq \mathbf{0} \tag{127}$$

- *Ejemplo:* la región factible es un poliedro convexo; el objetivo decrece hacia abajo a la derecha; óptimo en un vértice





8.5.3.1. El algoritmo simplex

- ***Algoritmo simplex:*** resuelve problemas lineales moviéndose de vértice a vértice, siguiendo la arista que mejora más el objetivo
- El simplex es muy eficiente en la práctica, si bien su complejidad temporal en el peor de los casos es exponencial con D
- Se conocen algoritmos polinómicos en el peor de los casos, pero son más lentos que simplex en la práctica

8.5.3.2. Aplicaciones

- Programación lineal se aplica en muchos campos
- En aprendizaje automático se ha aplicado en regresión lineal robusta y modelos gráficos

8.5.4. Programación cuadrática

- ***Programa cuadrático (QP):*** minimización de un objetivo cuadrático sujeto a restricciones lineales de igualdad y desigualdad

$$\min_{\boldsymbol{\theta}} \ \frac{1}{2}\boldsymbol{\theta}^t \mathbf{H} \boldsymbol{\theta} + \boldsymbol{c}^t \boldsymbol{\theta} \tag{128}$$

sujeto a

$$\mathbf{A}\boldsymbol{\theta} \leq \boldsymbol{b}, \quad \mathbf{A}_{\text{eq}}\boldsymbol{\theta} = \boldsymbol{b}_{\text{eq}} \tag{129}$$

- Si $\mathbf{H} \succeq 0$, es un problema de optimización convexo

8.5.4.1. Ejemplo: objetivo cuadrático 2d con restricciones lineales de desigualdad

- Objetivo a minimizar:

$$\mathcal{L}(\boldsymbol{\theta}) = \left(\theta_1 - \frac{3}{2}\right)^2 + \left(\theta_2 - \frac{1}{8}\right)^2 = \frac{1}{2}\boldsymbol{\theta}^t \mathbf{H} \boldsymbol{\theta} + \boldsymbol{c}^t \boldsymbol{\theta} + \text{const} \tag{130}$$

donde $\mathbf{H} = 2\mathbf{I}$ y $\boldsymbol{c} = -(3, 1/4)$, sujeto a

$$|\theta_1| + |\theta_2| \leq 1 \tag{131}$$

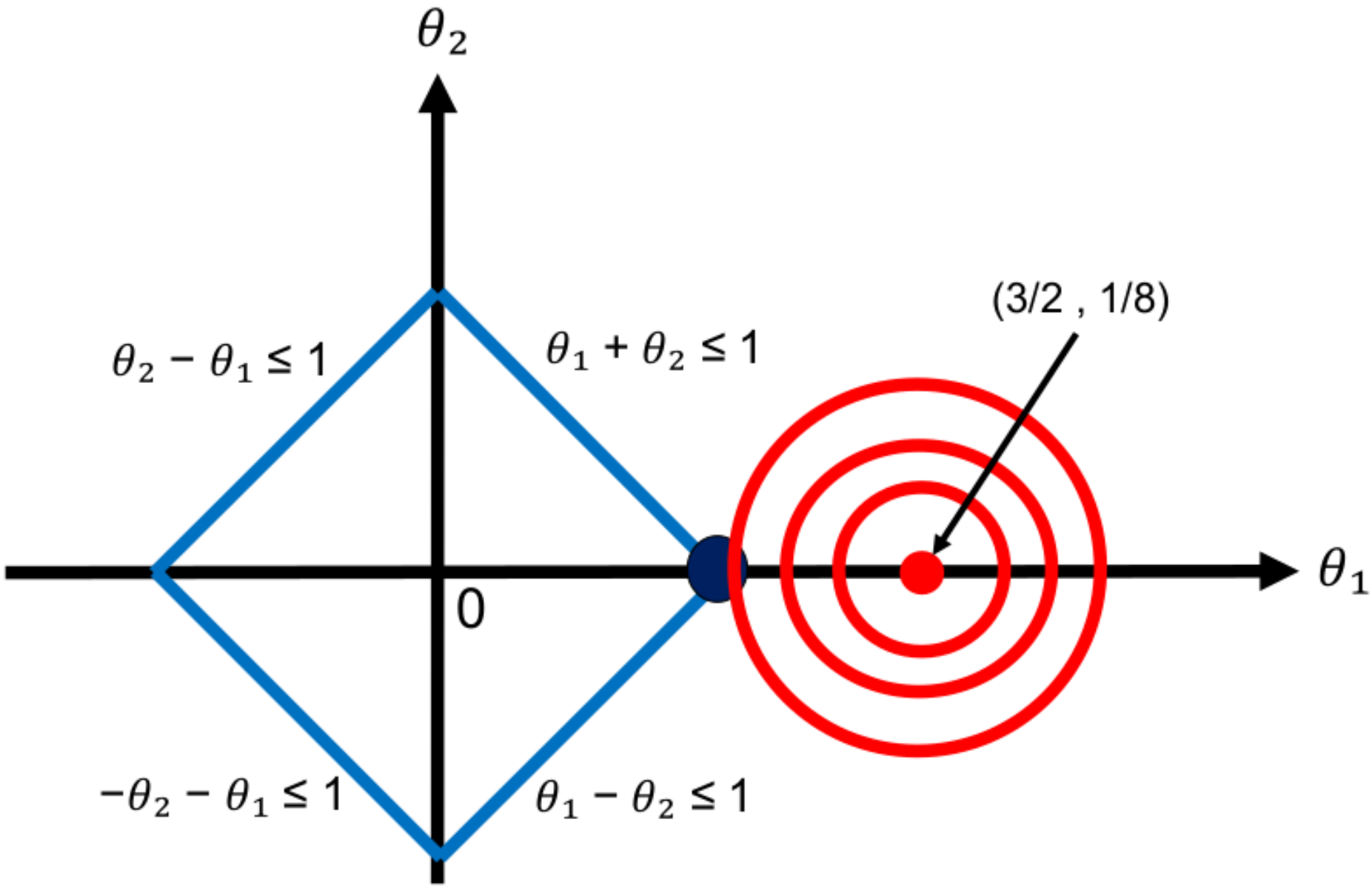
esto es,

$$g_1 : \theta_1 + \theta_2 \leq 1, \; g_2 : \theta_1 - \theta_2 \leq 1, \; g_3 : -\theta_1 + \theta_2 \leq 1, \; g_4 : -\theta_1 - \theta_2 \leq 1 \tag{132}$$

o, en forma estándar,

$$\mathbf{A}\boldsymbol{\theta} - \boldsymbol{b} \leq \mathbf{0} \quad \text{con} \quad \boldsymbol{b} = \mathbf{1} \quad \text{y} \quad \mathbf{A} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix} \tag{133}$$

- Representación gráfica:



- Condiciones KKT:
  - g_3 y g_4 inactivas; $g_3(\boldsymbol{\theta}^*) > 0,\ g_4(\boldsymbol{\theta}^*) > 0,\ \mu_3^* = \mu_4^* = 0$
  - Estacionaridad:

$$\mathbf{H}\boldsymbol{\theta} + \boldsymbol{c} + \mathbf{A}^t\boldsymbol{\mu} = \mathbf{0} \tag{134}$$

  - Estacionaridad ignorando inactivas:

$$\begin{pmatrix} 2 & 0 & 1 & 1 \\ 0 & 2 & 1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 1/4 \\ 1 \\ 1 \end{pmatrix} \tag{135}$$

  - Solución:

$$\boldsymbol{\theta}^* = (1, 0)^t \qquad \mu^* = (0.625, 0.375, 0, 0)^t \tag{136}$$

8.5.4.2. Aplicaciones

- Regresión lineal dispersa
- SVMs

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

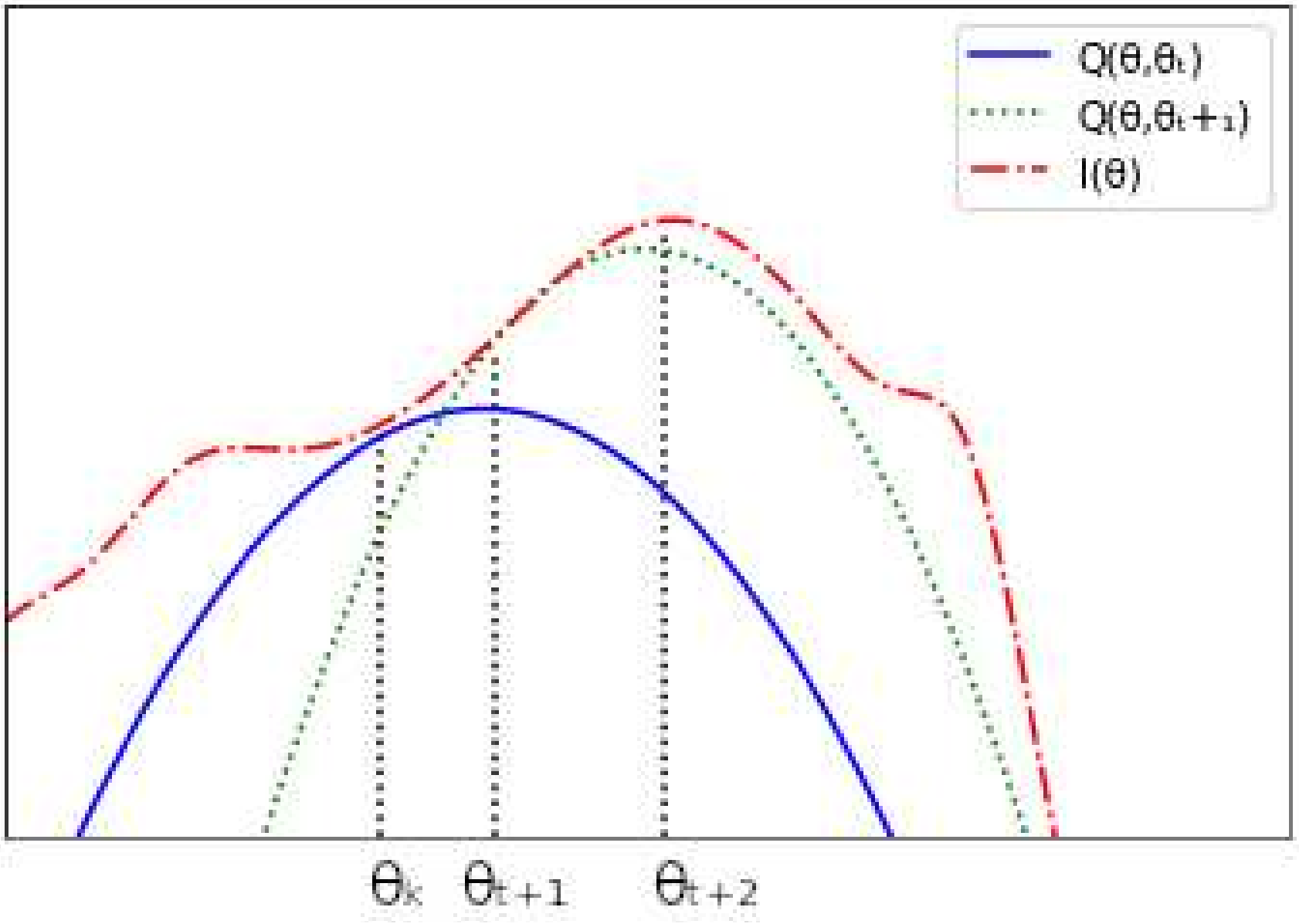
- ***Algoritmo majorize-minimize (MM):*** para $t = 0, 1, \ldots$

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

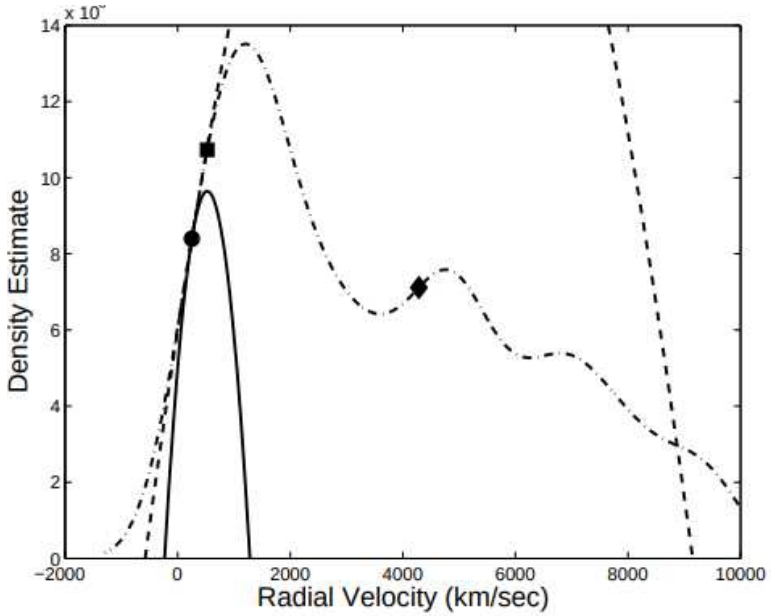
  - Si $\boldsymbol{\theta}^{t+1}$ se escoge tal que $Q(\boldsymbol{\theta}^{t+1}, \boldsymbol{\theta}^t) \geq Q(\boldsymbol{\theta}^t, \boldsymbol{\theta}^t)$:

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

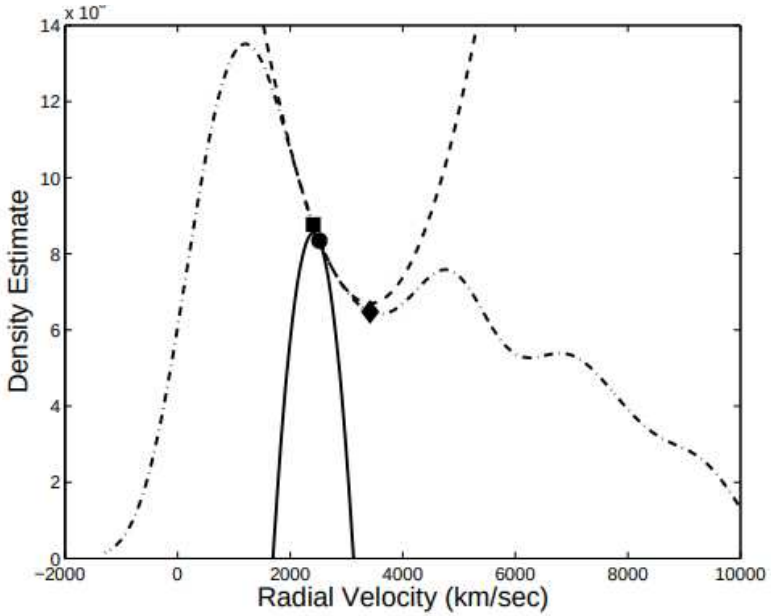
- *Ejemplo:* $Q(\theta, \theta^t)$ toca $\mathrm{LL}(\theta)$ en θ^t; su maximización da lugar a θ^{t+1} y $Q(\theta, \theta^{t+1})$ toca $\mathrm{LL}(\theta)$ en θ^{t+1}; su maximización da lugar a θ^{t+2}, etc.



- ***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 $r_k^{(t)} \triangleq \sum_n r_{nk}(t)$

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

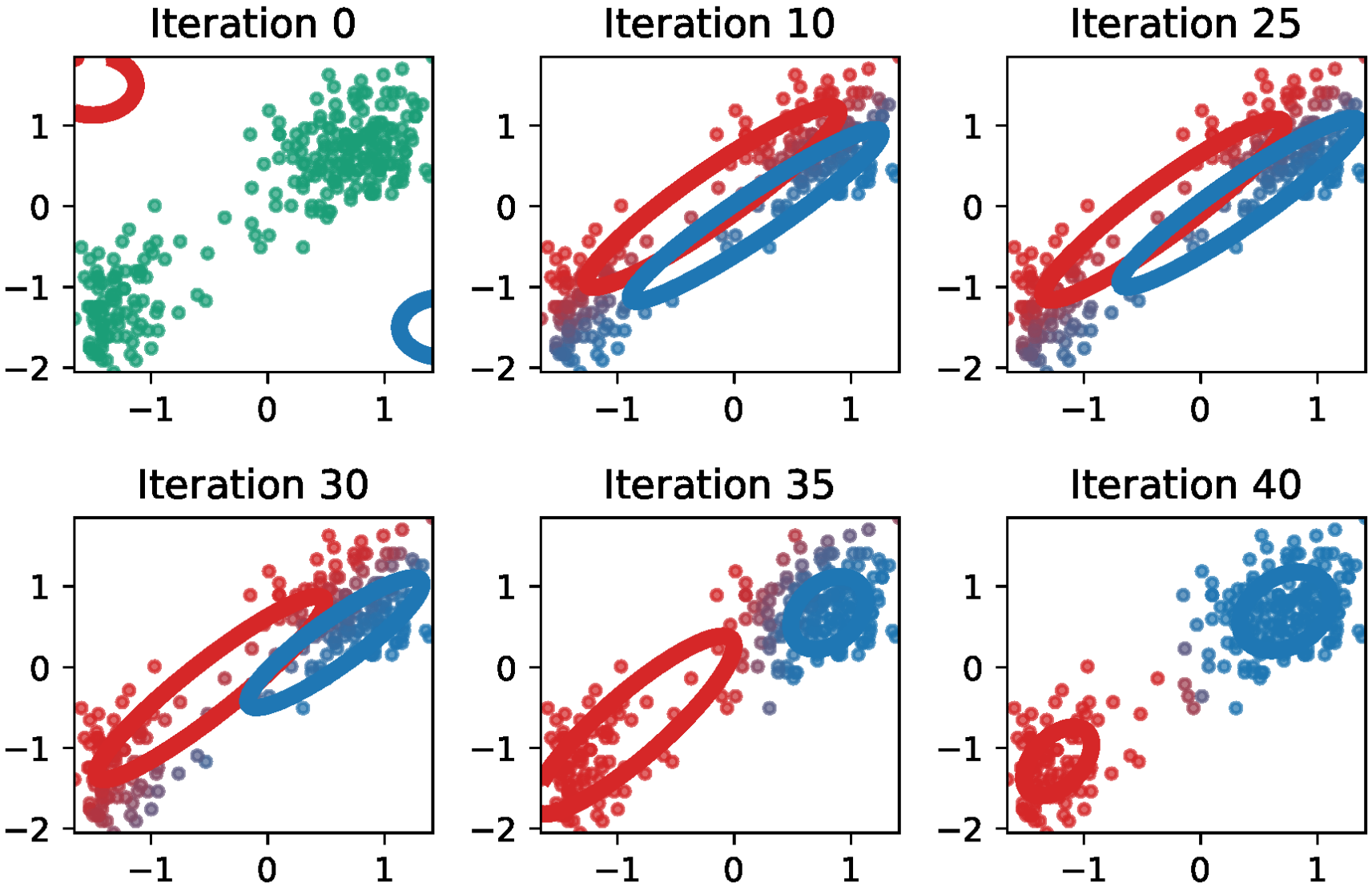
$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{r_k^{(t)}} \sum_n r_{nk}(t) (\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(t+1)})(\boldsymbol{y}_n - \boldsymbol{\mu}_k^{(t+1)})^t \tag{166}$$

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

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

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:

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

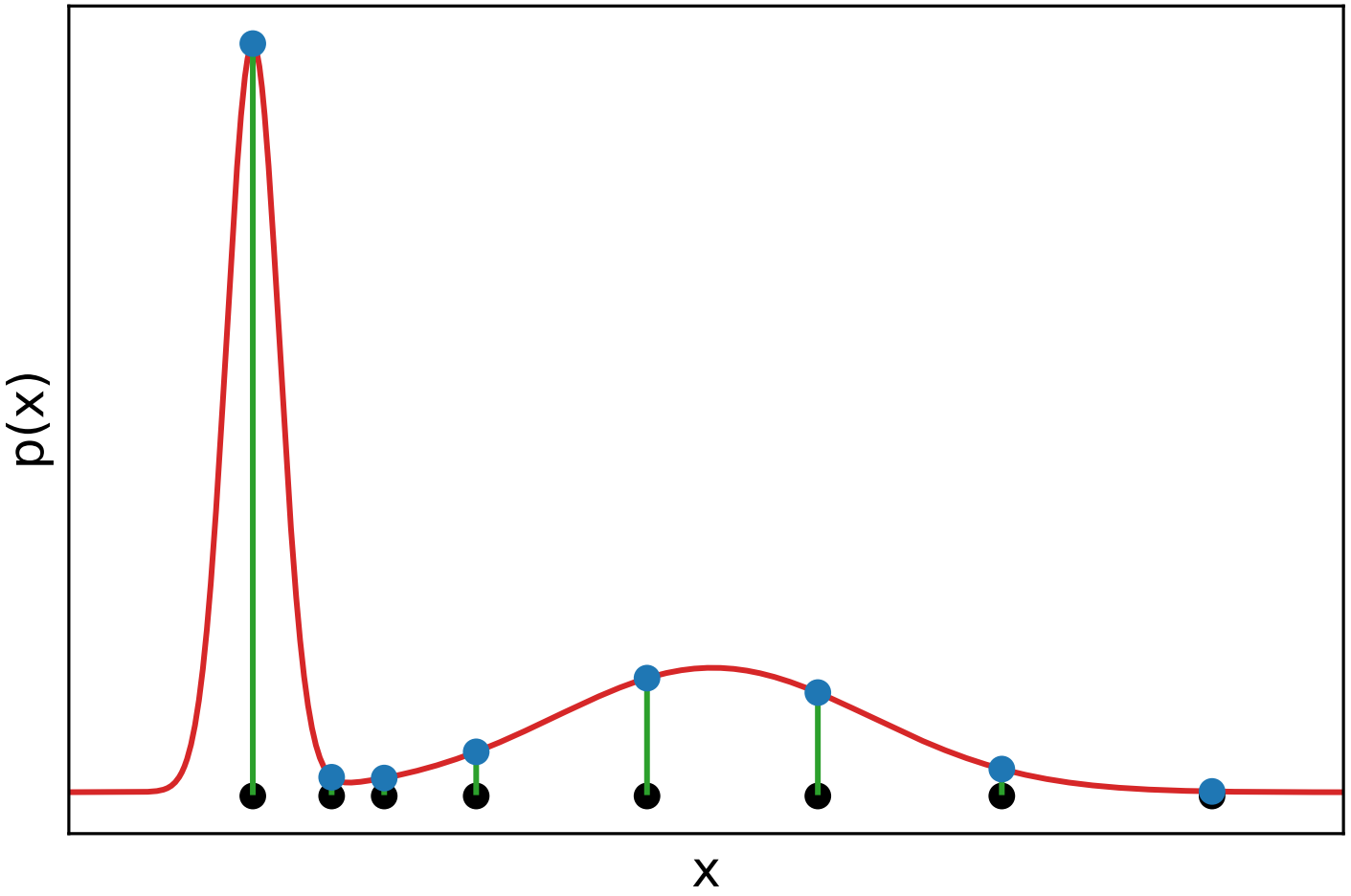
$$\tilde{\boldsymbol{\Sigma}}_k^{(t+1)} = \frac{\breve{\mathbf{S}} + \hat{\boldsymbol{\Sigma}}_k^{(t+1)}}{\breve{\nu} + r_k^{(t)} + D + 2} \tag{174}$$

  - ***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

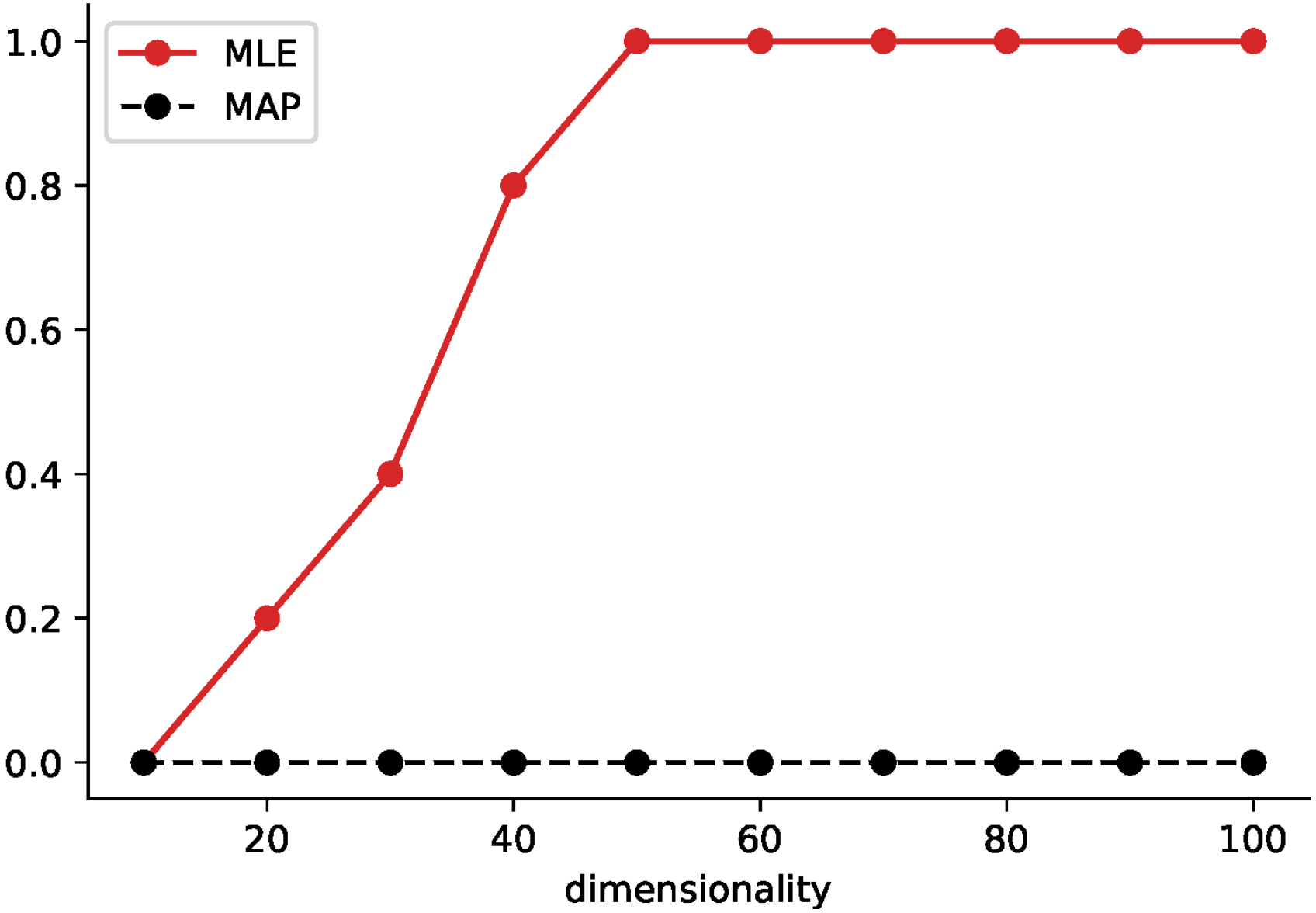
$$\breve{\mathbf{S}} = \frac{1}{K^{1/D}} \mathrm{diag}(s_1^2, \ldots, s_D^2) \tag{175}$$

  - 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

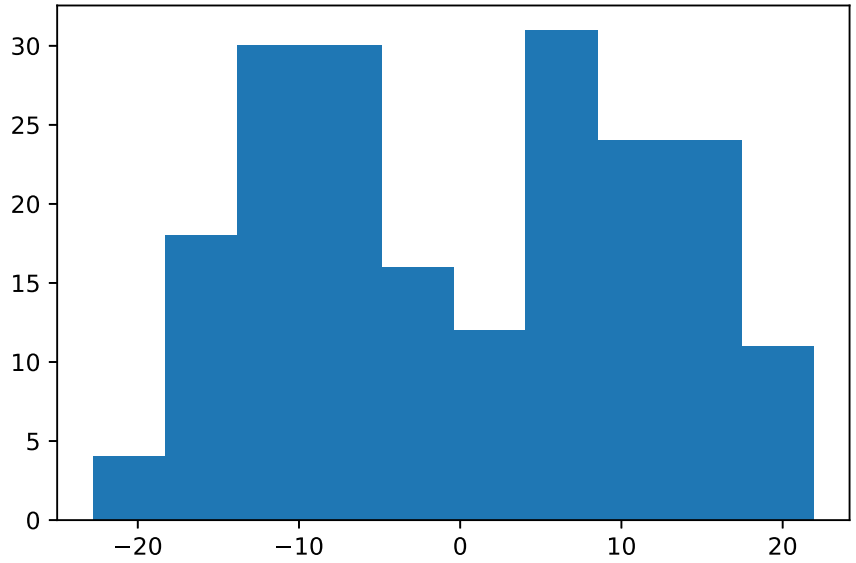


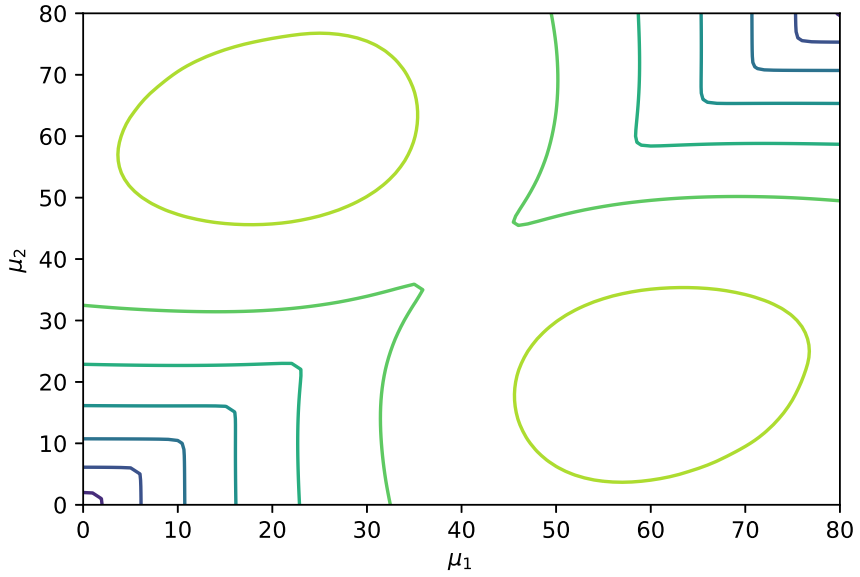
8.7.3.5. Noconvexidad de la NLL

- ***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

8.8. Optimización sin derivadas y caja-negra

- ***Optimización sin derivadas y caja-negra:*** optimización de problemas en los que la forma funcional del objetivo es desconocida

- Aplicación en ***selección de modelos:*** tipo y complejidad dependiente de un vector de hiper-parámetros $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$ y objetivo $\mathcal{L}(\boldsymbol{\lambda})$ usualmente definido como pérdida en un conjunto de validación

- ***Búsqueda en rejilla:*** se define una rejilla discreta de posibles hiper-parámetros y se enumeran todos