Descenso del gradiente

Optimización determinista vs estocástica

- Métodos deterministas: gradiente descendente, Newton-Raphson.
- Métodos estocásticos: simulación de Monte Carlo, algoritmos genéticos, recocido simulado.

Determinismo: el ideal de una ciencia total

El determinismo ha sido, históricamente, uno de los pilares conceptuales de la ciencia moderna. Bajo esta concepción, todo evento en el universo es efecto necesario de causas precedentes. Si se conociera con precisión suficiente el estado actual del mundo, podría anticiparse con total exactitud su evolución futura. Esta tesis, común en la física clásica, otorga al conocimiento un poder predictivo absoluto y establece una relación directa entre explicación y predicción.

El demonio de Laplace: una inteligencia perfecta

Pierre-Simon Laplace llevó esta idea al extremo al postular una inteligencia —el célebre "demonio"— que, conociendo todas las leyes de la naturaleza y la posición de cada partícula en el universo, podría calcular el pasado y el futuro con exactitud. No se trata simplemente de una hipótesis metafísica, sino de una metáfora que encarna el optimismo epistemológico del siglo XVIII. Sin embargo, esta figura plantea interrogantes: ¿es legítimo excluir al propio sujeto cognoscente de la cadena causal que pretende conocer? ¿Puede el conocimiento ser exterior al sistema que describe?

El problema del azar: ¿ignorancia o estructura?

Frente a la certeza determinista, el pensamiento contemporáneo ha debido enfrentarse a fenómenos que escapan a esta lógica. La categoría de lo estocástico emerge para describir procesos en los cuales el resultado no está completamente determinado. Desde una perspectiva epistemológica, el azar puede interpretarse como expresión de ignorancia: no sabemos lo suficiente. Pero también existe una interpretación ontológica: algunos eventos ocurren sin causa suficiente, como parte de la estructura misma del mundo.

Física cuántica: el fin del determinismo clásico

La teoría cuántica introdujo una ruptura radical. A diferencia de la física newtoniana, los fenómenos cuánticos no admiten una descripción determinista. El principio de incertidumbre y la noción de colapso de la función de onda sugieren que ciertos eventos no están predeterminados, sino que se manifiestan probabilísticamente. ¿Significa esto que el universo es, en su fundamento, estocástico? Las interpretaciones divergen: desde la escuela de Copenhague hasta las teorías de variables ocultas y los mundos múltiples, el debate permanece abierto.

Tensiones conceptuales: más allá de la dicotomía

El contraste entre determinismo y estocasticidad no es absoluto. Existen niveles de descripción en los que ambos enfoques coexisten. Por ejemplo, un sistema determinista puede exhibir comportamientos caóticos o impredecibles debido a su sensibilidad a condiciones iniciales. Por otro lado, modelos estocásticos pueden describir regularidades agregadas sin negar la existencia de causas subyacentes. Esto plantea preguntas filosóficas relevantes: ¿qué tipo de causalidad reconocemos? ¿Qué papel otorgamos al sujeto? ¿Tiene sentido hablar de libertad bajo incertidumbre?

Consecuencias filosóficas y científicas

En la práctica científica, lo estocástico se ha naturalizado como herramienta de modelación. Las ciencias sociales, biológicas y económicas adoptan modelos probabilísticos para describir fenómenos complejos. En filosofía, el problema del libre albedrío se ve reformulado: ya no se trata solo de enfrentar un mundo de causas necesarias, sino también de interpretar el significado de la indeterminación. En la inteligencia artificial, por ejemplo, la toma de decisiones bajo incertidumbre desafía la concepción clásica de agencia racional.

Optimización en aprendizaje automático

- Muchos problemas de aprendizaje automático se formulan como tareas de optimización.
- El objetivo típico es
$$\min_{\mathbf{x}} f(\mathbf{x}),$$
mientras que la maximización se logra minimizando $-f(\mathbf{x})$.
- En el contexto de minimización, f recibe los nombres de *pérdida*, *costo* o *error*.
- Ejemplo: mínimos cuadrados lineales
$$f(\mathbf{x}) = \frac{1}{2} \|A\mathbf{x} - \mathbf{b}\|_2^2.$$
- El argumento que alcanza el valor óptimo se escribe
$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}).$$

Optimización basada en gradientes

- Sea una función $y = f(x)$, donde x y y son números reales.
- La derivada de la función se denota como $f'(x)$ o $\frac{dy}{dx}$, y representa la pendiente de $f(x)$ en el punto x.
- La derivada nos permite aproximar el valor de la función en un punto cercano:

$$f(x + \varepsilon) \approx f(x) + \varepsilon f'(x)$$

- Esto indica cómo un pequeño cambio ε en la entrada afecta la salida y.
- Para minimizar $f(x)$, es útil moverse en dirección opuesta al signo de la derivada:

$$f\left(x - \varepsilon \cdot \mathrm{sign}(f'(x))\right) < f(x)$$

  para ε suficientemente pequeño.
- Esta idea conduce al algoritmo conocido como **descenso de gradiente**, propuesto por Cauchy en 1847.

Problemas de optimización

- Muchos problemas prácticos implican la optimización de una función $f(\boldsymbol{x})$ respecto de sus parámetros $\boldsymbol{x}$
- La tarea puede consistir en minimizar o maximizar dicha función
- Por convención, se suele expresar como un problema de minimización. Maximizar $f(\boldsymbol{x})$ equivale a minimizar $-f(\boldsymbol{x})$
- La función $f(\boldsymbol{x})$ recibe diversos nombres según el contexto: función objetivo, criterio, función de pérdida, función de costo, entre otros
- Un ejemplo clásico es el de mínimos cuadrados:

$$f(\boldsymbol{x}) = \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2$$

- El valor óptimo se denota como:

$$\boldsymbol{x}^* = \operatorname{argmin} f(\boldsymbol{x})$$

Concepto de derivada

- Considere una función $y = f(x)$, con $x, y \in \mathbb{R}$
- La derivada $f'(x)$ (también escrita como $\frac{dy}{dx}$) representa la pendiente de la función en el punto x
- Aproximación para pequeños desplazamientos:

$$f(x + \varepsilon) \approx f(x) + \varepsilon f'(x)$$

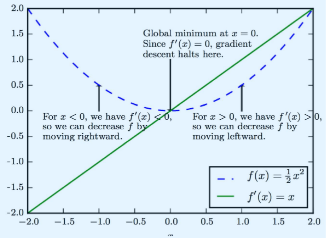
- Esto permite estimar cómo varía la salida de la función ante un pequeño cambio en su argumento
- Para reducir el valor de $f(x)$, se puede desplazar x en la dirección opuesta al signo de la derivada:

$$f\left(x - \varepsilon \, \mathrm{sign}(f'(x))\right) < f(x)$$

- Este principio es la base del método de descenso por gradiente (Cauchy, 1847)

Ejemplo ilustrativo

- Sea $f(x) = \frac{1}{2}x^2$, una función con forma de parábola que alcanza su mínimo global en $x = 0$
- Derivada: $f'(x) = x$
- Para $x > 0$: la función crece y $f'(x) > 0$
- Para $x < 0$: la función decrece y $f'(x) < 0$
- En ambos casos, moverse en dirección opuesta a $f'(x)$ reduce el valor de la función



Representación de $f(x) = \frac{1}{2}x^2$

Descenso en superficies de error

Dirección en el plano $w_0 - w_1$ que conduce al descenso más pronunciado de una función de error

- La superficie de error $E(\boldsymbol{w})$ representa la calidad de cada vector de parámetros $\boldsymbol{w}$
- Comúnmente, esta superficie presenta una forma parabólica con un único mínimo global
- Se parte de un valor inicial $\boldsymbol{w}^{(0)}$ y se actualiza iterativamente

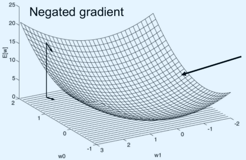


Figure: *

Superficie de error con mínimo global

Algoritmo de descenso por gradiente

Algorithm 1 *

Procedimiento iterativo para minimizar $E(\boldsymbol{w})$

Inicializar $\boldsymbol{w}^{(0)}$

Elegir tasa de aprendizaje $\eta > 0$ **for** $t = 0, 1, 2, \ldots$ *hasta convergencia* **do**

Calcular el gradiente: $\nabla E(\boldsymbol{w}^{(t)})$

Actualizar parámetros:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta \nabla E(\boldsymbol{w}^{(t)})$$

Vector gradiente

El vector gradiente de una función escalar $E(\boldsymbol{w})$ respecto de un vector de parámetros $\boldsymbol{w}$ se define como:

$$\nabla E(\boldsymbol{w}) \equiv \left[\frac{\partial E}{\partial w_0}, \frac{\partial E}{\partial w_1}, \cdots, \frac{\partial E}{\partial w_n}\right]$$

- $\nabla E(\boldsymbol{w})$ es un vector que contiene las derivadas parciales de E respecto a cada componente de $\boldsymbol{w}$
- Indica la dirección de aumento más pronunciado de E
- La dirección opuesta, $-\nabla E(\boldsymbol{w})$, es aquella que produce la disminución más rápida de E

Derivada direccional

La derivada direccional en la dirección de $\mathbf{u}$ (un vector unitario) es la pendiente de la función f en la dirección de $\mathbf{u}$. Esto se evalúa como $\mathbf{u}^T \nabla_x f(x)$, donde $\nabla_x f(x)$ representa el gradiente de f respecto a x.

Derivada direccional

Para minimizar f, es necesario encontrar la dirección en la cual f disminuye más rápidamente. Esto se logra mediante el uso de la expresión $\min_{\mathbf{u},\mathbf{u}^T\mathbf{u}=1} \mathbf{u}^T \nabla_x f(x)$. Esta expresión se puede reescribir, teniendo en cuenta la restricción $\mathbf{u}^T\mathbf{u} = 1$ (que implica que $\mathbf{u}$ es un vector unitario), como $\min_{\mathbf{u}} ||\mathbf{u}|| ||\nabla_x f(x)|| \cos\theta$, donde θ es el ángulo entre $\mathbf{u}$ y el gradiente de f.

Interpretación geométrica

- La expresión
$$\min_{\mathbf{u},\, \|\mathbf{u}\|=1} \mathbf{u}^T \nabla f(x)$$
busca la dirección $\mathbf{u}$ en la que la función f disminuye más rápidamente.
- Como $\|\mathbf{u}\| = 1$, esta expresión depende solo del ángulo θ entre $\mathbf{u}$ y el gradiente $\nabla f(x)$:
$$\mathbf{u}^T \nabla f(x) = \|\nabla f(x)\| \cos \theta$$
- Para que $\mathbf{u}^T \nabla f(x)$ sea lo más pequeño posible, necesitamos que $\cos \theta = -1$, es decir, que $\theta = \pi$ (180°).
- Esto significa que la dirección de mayor descenso es:
$$\mathbf{u} = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$$
- En resumen: el gradiente apunta hacia la subida más pronunciada, y su opuesto hacia la bajada más rápida.

Derivada direccional y gradiente

Considere una dirección unitaria $\boldsymbol{u}$, es decir, $||\boldsymbol{u}||_2 = 1$. Al evaluar la tasa de cambio direccional de una función f en la dirección de $\boldsymbol{u}$, se obtiene:

- La derivada direccional es proporcional a $\cos\theta$, donde θ es el ángulo entre $\boldsymbol{u}$ y el gradiente $\nabla_x f(x)$
- Al minimizar $\cos\theta$, se alcanza el valor mínimo cuando $\boldsymbol{u}$ apunta en dirección opuesta a $\nabla_x f(x)$
- Por lo tanto, el gradiente apunta en la dirección de máximo aumento, y su opuesto $-\nabla_x f(x)$ señala la dirección de máxima disminución

Esta observación es la base de métodos de optimización como el descenso por gradiente, que actualiza la posición en la dirección de mayor disminución de f

Regla de actualización en descenso por gradiente

La regla básica del método de descenso por gradiente se expresa como:

$$\boldsymbol{w} \leftarrow \boldsymbol{w} + \Delta \boldsymbol{w}$$

donde el vector de actualización $\Delta \boldsymbol{w}$ se define como:

$$\Delta \boldsymbol{w} = -\eta \nabla E(\boldsymbol{w})$$

- $\nabla E(\boldsymbol{w})$ representa el gradiente de la función de error E respecto a $\boldsymbol{w}$
- $\eta > 0$ es la tasa de aprendizaje, que regula el tamaño del paso
- El término $-\eta \nabla E(\boldsymbol{w})$ apunta en la dirección de mayor disminución de E

Este procedimiento se repite iterativamente hasta alcanzar la convergencia hacia un mínimo local.

Método del descenso del gradiente

La tasa de aprendizaje es crucial porque un valor muy alto puede llevar a que el algoritmo sobrepase el mínimo, mientras que un valor muy bajo puede hacer que el algoritmo avance muy lentamente hacia el mínimo, aumentando el número de iteraciones necesarias para converger.

Método del descenso del gradiente

La tasa de aprendizaje η controla cuánto se actualiza el valor en cada iteración:

$$x^{(t+1)} = x^{(t)} - \eta \nabla f(x^{(t)})$$

- Si η es muy grande, el algoritmo puede *divergir*.
- Si η es muy pequeña, el algoritmo *converge muy lentamente*.

Para la función $f(x) = x^2$, su gradiente es:

$$\nabla f(x) = 2x$$

Aplicando descenso de gradiente:

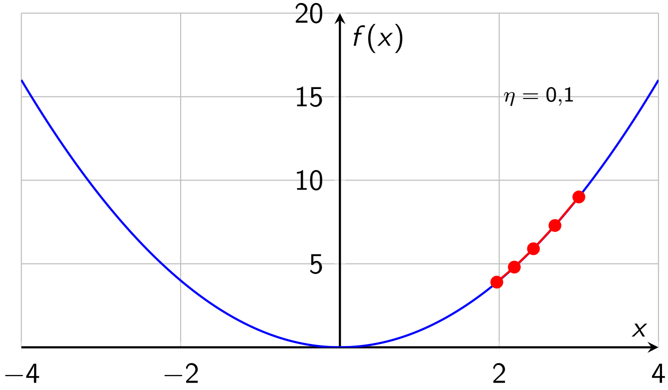
$$x^{(t+1)} = x^{(t)} - 2\eta x^{(t)} = (1 - 2\eta) x^{(t)}$$

- Si $0 < \eta < 1$, entonces $|1 - 2\eta| < 1$: hay convergencia.
- Si $\eta > 1$, entonces $|1 - 2\eta| > 1$: hay divergencia.
- Si $\eta = 0{,}001$, entonces:

  $$x^{(t)} = (1 - 0{,}002)^t x^{(0)} \approx e^{-0{,}002 t} x^{(0)}$$

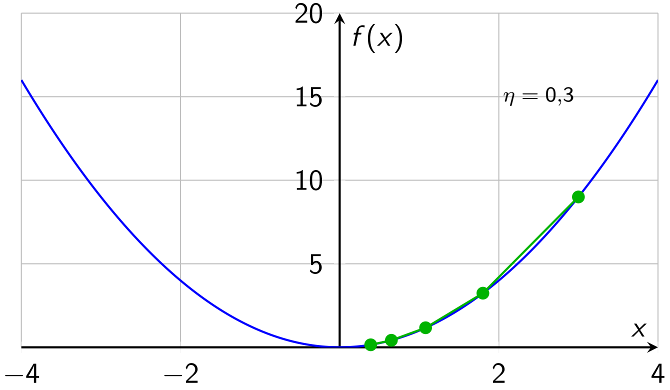
  lo cual indica una convergencia muy lenta.

Efecto de la tasa de aprendizaje: $f(x) = x^2$



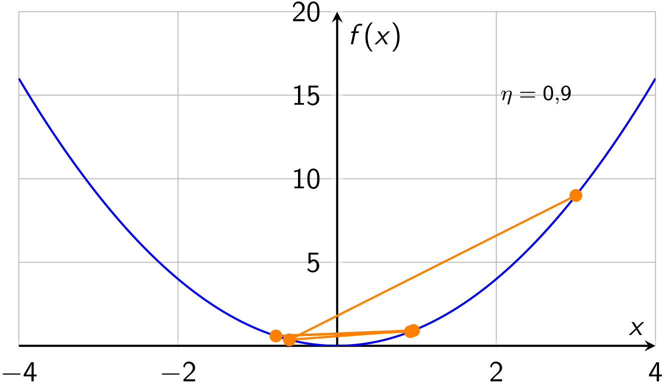
Tasa de aprendizaje pequeña: convergencia lenta

Efecto de la tasa de aprendizaje: $f(x) = x^2$



Tasa adecuada: convergencia eficiente

Efecto de la tasa de aprendizaje: $f(x) = x^2$



Tasa grande: el algoritmo oscila o diverge

Algoritmo de descenso de gradiente

Algorithm 2 Descenso de gradiente con criterio de convergencia

Input: $\boldsymbol{\theta}^1$ punto inicial, f función objetivo, δ umbral de convergencia, η tasa de aprendizaje

Output: $\boldsymbol{\theta}^t$ mínimo aproximado

```
t ← 1   repeat
    θ^{t+1} ← θ^t − η∇f(θ^t)   t ← t + 1
until ‖θ^t − θ^{t−1}‖ ≤ δ
return θ^t
```

$t \leftarrow 1$ **repeat**

$\quad \boldsymbol{\theta}^{t+1} \leftarrow \boldsymbol{\theta}^t - \eta \nabla f(\boldsymbol{\theta}^t) \quad t \leftarrow t + 1$

until $\|\boldsymbol{\theta}^t - \boldsymbol{\theta}^{t-1}\| \leq \delta$

return $\boldsymbol{\theta}^t$

Descenso de gradiente en regresión de mínimos cuadrados

Algorithm 3 Descenso de gradiente aplicado a mínimos cuadrados

Input: $A \in \mathbb{R}^{n \times p}$ matriz de diseño, $b \in \mathbb{R}^n$ vector de salida, $\boldsymbol{x}^1$ valor inicial, η tasa de aprendizaje, δ umbral de convergencia

Output: $\boldsymbol{x}^t$ mínimo aproximado de $f(\boldsymbol{x}) = \frac{1}{2}\|A\boldsymbol{x} - b\|^2$

$t \leftarrow 1$ **repeat**

  $\nabla f(\boldsymbol{x}^t) \leftarrow A^T A\boldsymbol{x}^t - A^T b$ $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t - \eta \nabla f(\boldsymbol{x}^t)$ $t \leftarrow t + 1$

until $\|\boldsymbol{x}^t - \boldsymbol{x}^{t-1}\| \leq \delta$

return $\boldsymbol{x}^t$

Fundamento del gradiente en mínimos cuadrados

Dada la función a minimizar:

$$f(\boldsymbol{x}) = \frac{1}{2}\|A\boldsymbol{x} - b\|^2$$

Su gradiente es:

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = A^T A\boldsymbol{x} - A^T b$$

- El vector gradiente apunta en la dirección de mayor aumento de f.
- Para minimizar f, el algoritmo ajusta $\boldsymbol{x}$ en la dirección opuesta:

$$\boldsymbol{x}^{t+1} = \boldsymbol{x}^t - \eta(A^T A\boldsymbol{x}^t - A^T b)$$

- Este procedimiento garantiza una disminución iterativa del error cuadrático.

Función objetivo

Consideramos la función:

$$f(x) = (x - 3)^2$$

- Tiene un mínimo en $x = 3$
- El gradiente es $\nabla f(x) = 2(x - 3)$
- Usamos descenso de gradiente para encontrar el mínimo

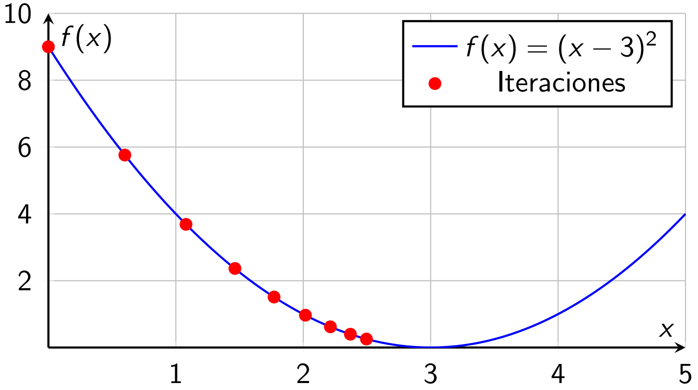
Parámetros del algoritmo

- Valor inicial: $x^{(0)} = 0$
- Tasa de aprendizaje: $\eta = 0.1$
- Criterio de parada: $|x^{(t)} - x^{(t-1)}| < 10^{-3}$

Iteraciones paso a paso

t	$x^{(t)}$	$\nabla f(x^{(t)})$	$x^{(t+1)} = x^{(t)} - \eta \cdot \nabla f$
0	0.000	-6.000	0.600
1	0.600	-4.800	1.080
2	1.080	-3.840	1.464
3	1.464	-3.072	1.771
4	1.771	-2.457	2.017
5	2.017	-1.966	2.213
6	2.213	-1.573	2.370
7	2.370	-1.260	2.496
8	2.496	-1.008	2.597

Trayectoria del descenso de gradiente



Problema de optimización

Considere la función:

$$f(x) = x^4 - 4x^2 + 4$$

- Tiene varios puntos críticos.
- Buscamos encontrar un mínimo local aplicando descenso de gradiente.
- Se utilizará un punto inicial $x^{(0)} = -2$.

Gradiente de la función

La derivada de la función es:

$$f'(x) = 4x^3 - 8x$$

La regla de actualización en el descenso de gradiente es:

$$x^{(t+1)} = x^{(t)} - \eta \cdot f'(x^{(t)})$$

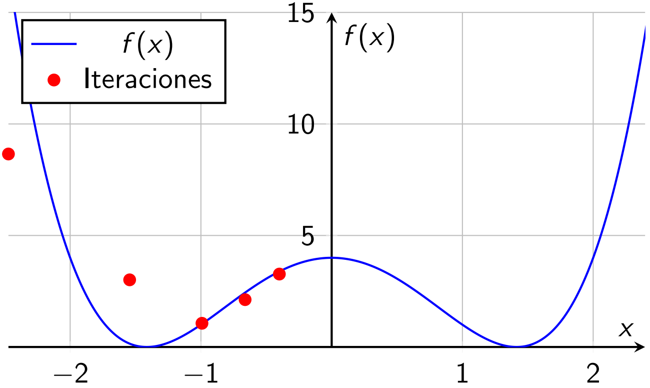
Usamos una tasa de aprendizaje fija $\eta = 0.1$

Iteraciones paso a paso

t	$x^{(t)}$	$f'(x^{(t)})$	$x^{(t+1)}$
0	-2.000	-16	-0.400
1	-0.400	2.624	-0.662
2	-0.662	3.296	-0.992
3	-0.992	5.527	-1.545
4	-1.545	9.270	-2.472
5	-2.472	29.254	-5.397

- El algoritmo oscila debido a la curvatura pronunciada.
- Puede divergir si η no se ajusta correctamente.

Visualización de la trayectoria



Problema en dos dimensiones

Queremos minimizar:

$$f(x,y) = x^2 + 2y^2 + xy$$

- Tiene un mínimo único en $(x^*, y^*) = (0,0)$
- El gradiente es:
$$\nabla f(x,y) = \begin{bmatrix} 2x + y \\ 4y + x \end{bmatrix}$$
- Utilizaremos descenso de gradiente para encontrar el mínimo.

Regla de actualización

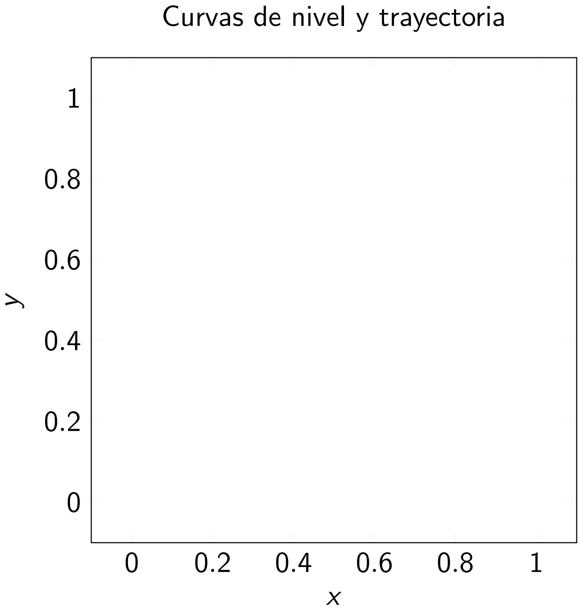
A partir del punto actual $(x^{(t)}, y^{(t)})$, se actualiza con:

$$\begin{bmatrix} x^{(t+1)} \\ y^{(t+1)} \end{bmatrix} = \begin{bmatrix} x^{(t)} \\ y^{(t)} \end{bmatrix} - \eta \cdot \begin{bmatrix} 2x^{(t)} + y^{(t)} \\ 4y^{(t)} + x^{(t)} \end{bmatrix}$$

Usamos:

- Punto inicial: $(x^{(0)}, y^{(0)}) = (2, 2)$
- Tasa de aprendizaje: $\eta = 0.1$

Trayectoria del descenso (curvas de nivel)



Gradiente y matriz Hessiana

- El gradiente $\nabla f(\boldsymbol{x})$ indica la dirección de mayor aumento local de f
- La matriz Hessiana $H(\boldsymbol{x})$ contiene las segundas derivadas parciales:

$$H(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

- La Hessiana describe la curvatura local de f
- Permite determinar si un punto crítico es mínimo, máximo o silla

Función cuadrática

Considere la función:

$$f(x, y) = 3x^2 + 2xy + y^2$$

- Gradiente:

$$\nabla f(x, y) = \begin{bmatrix} 6x + 2y \\ 2x + 2y \end{bmatrix}$$

- Hessiana:

$$H(x, y) = \begin{bmatrix} 6 & 2 \\ 2 & 2 \end{bmatrix}$$

- La matriz Hessiana es constante y simétrica

Función cuadrática

- Como H es simétrica, puede analizarse mediante sus autovalores
- Autovalores de H:

$$\lambda_1 = 1.17, \quad \lambda_2 = 6.83 \quad \text{(aproximadamente)}$$

- Ambos autovalores son positivos $\Rightarrow$ la matriz es definida positiva
- Entonces, $f(x, y)$ tiene un **mínimo estricto** en el punto donde el gradiente se anula
- Este punto es $(x, y) = (0, 0)$

Implicancia en optimización

- La dirección de descenso más pronunciado se obtiene con $-\nabla f(\boldsymbol{x})$
- El comportamiento local cerca del mínimo depende de la curvatura, dada por H
- En métodos de segundo orden (como Newton), se usa la Hessiana para ajustar la dirección y el tamaño del paso:

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - H^{-1}(\boldsymbol{x}_t)\nabla f(\boldsymbol{x}_t)$$

- Cuando la matriz Hessiana no es positiva definida o está mal condicionada, el uso de métodos que dependen de su inversión puede producir pasos numéricamente inestables o incorrectos. En tales casos, métodos de primer orden como el descenso por gradiente simple son más seguros, aunque más lentos.

Función con punto silla

Considere la función:

$$f(x, y) = x^2 - y^2$$

- Gradiente:

$$\nabla f(x, y) = \begin{bmatrix} 2x \\ -2y \end{bmatrix}$$

- Hessiana:

$$H(x, y) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

- La matriz Hessiana tiene autovalores 2 y $-2 \Rightarrow$ indefinida

Función con punto silla

- El punto crítico $(0,0)$ de la función $f(x,y) = x^2 - y^2$ cumple que el gradiente se anula, es decir, $\nabla f = 0$, pero no corresponde a un mínimo. En realidad, se trata de un punto silla: la función alcanza un mínimo en la dirección x y un máximo en la dirección y.

.

$$\begin{bmatrix} x \\ y \end{bmatrix} \leftarrow \begin{bmatrix} x \\ y \end{bmatrix} - \eta \begin{bmatrix} 2x \\ -2y \end{bmatrix} = \begin{bmatrix} x(1-2\eta) \\ y(1+2\eta) \end{bmatrix}$$