Tópicos en Inteligencia Artificial
Deep Learning

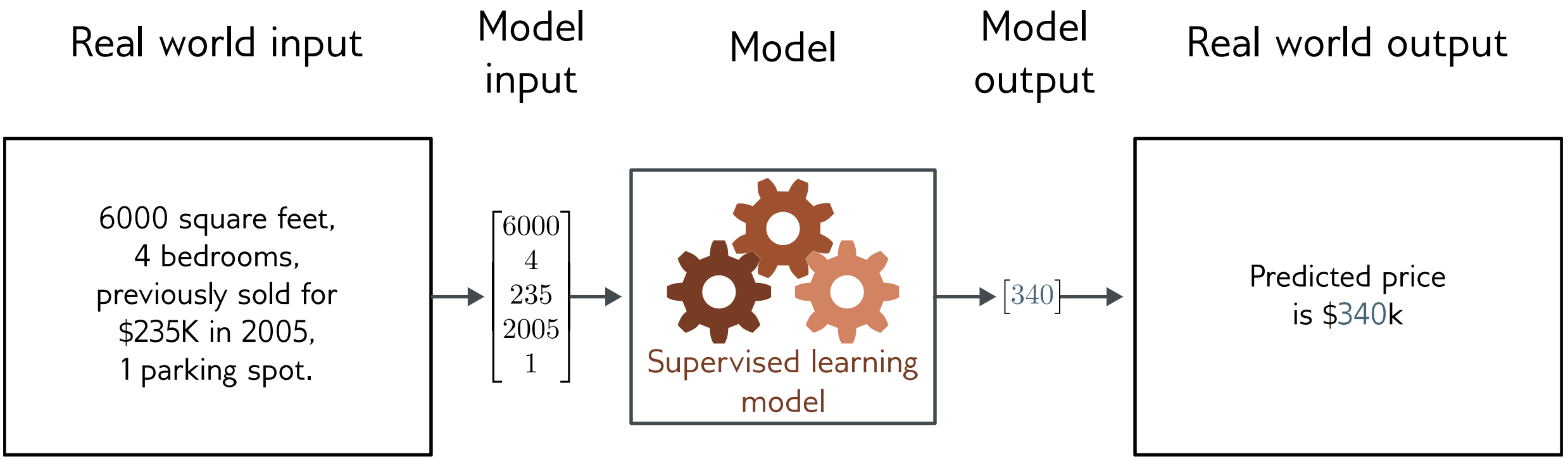
Algoritmos de optimización

Basado en el **ppt 06 – Fitting** de

Prof. Simon Prince

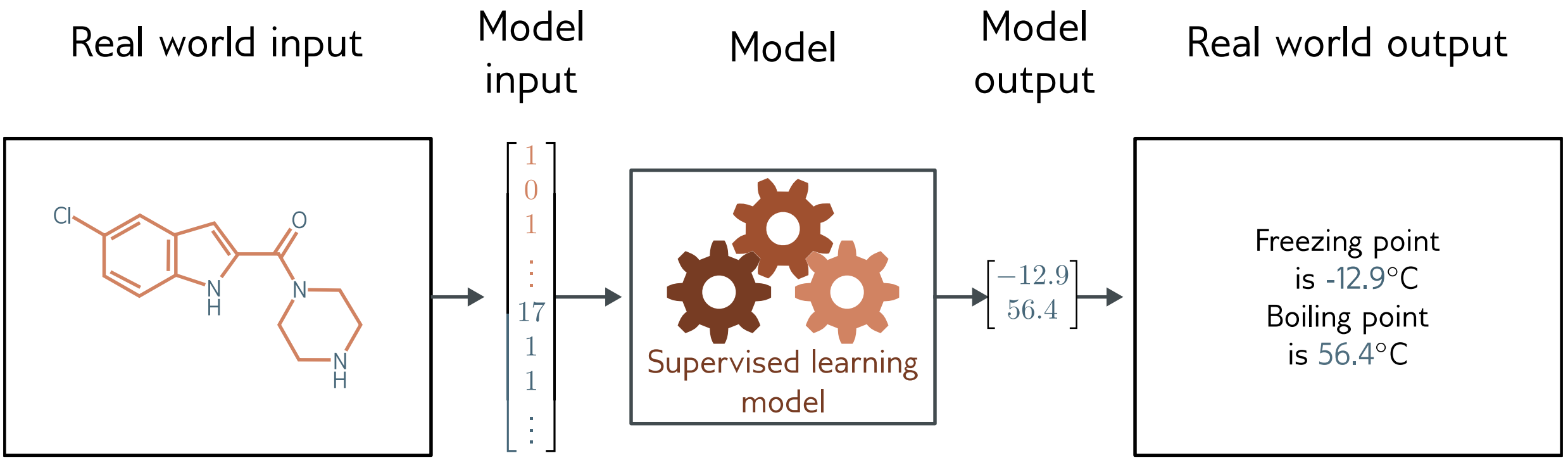
Adaptado por Prof. Fernando Crema García

Regresión



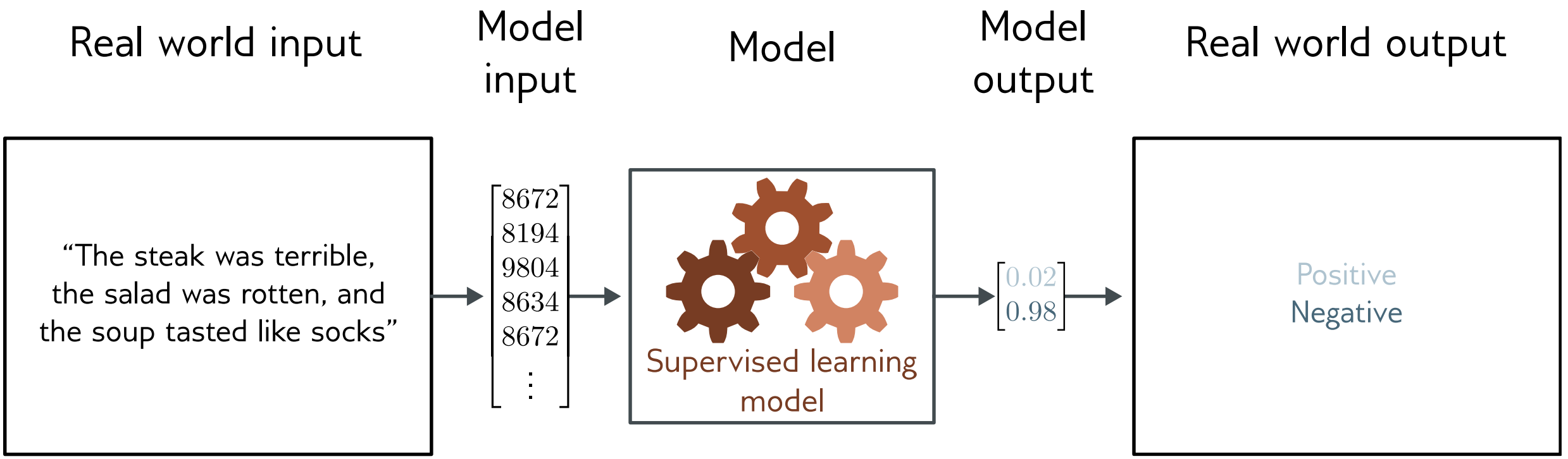
- Problema de regresión univariante (un resultado, valor real

Regresión gráfica



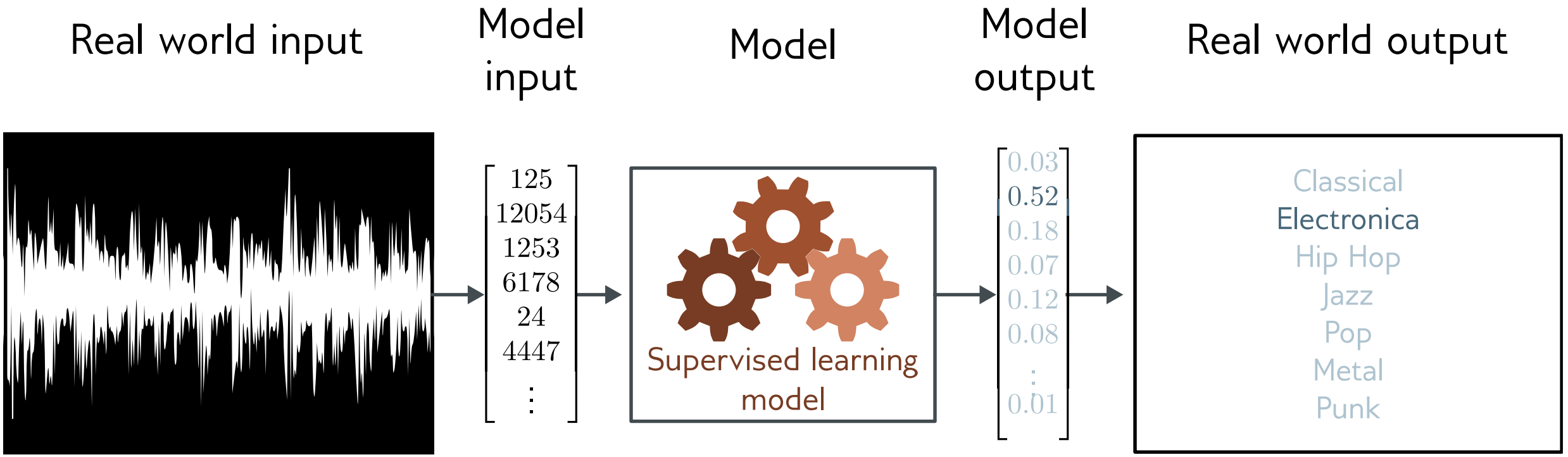
- Problema de regresión multivariante (>1 salida, valor real)

Clasificación de textos



- Problema de clasificación binaria (dos clases discretas)

Clasificación de los géneros musicales



- Problema de clasificación multiclase (clases discretas, >2 valores posibles)

Función de pérdida

- Conjunto de datos de entrenamiento de I pares de ejemplos de entrada/salida:

$$\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{I}$$

- La función de pérdida o función de coste mide lo malo que es el modelo:

$$L[\boldsymbol{\phi}, \mathrm{f}\left[\mathbf{x}_i, \boldsymbol{\phi}\right], \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{I}]$$

o para abreviar:

$$L\left[\boldsymbol{\phi}\right]$$

Devuelve un escalar que es más pequeño cuando el modelo asigna mejor las entradas a las salidas

Formación

- Función de pérdida:

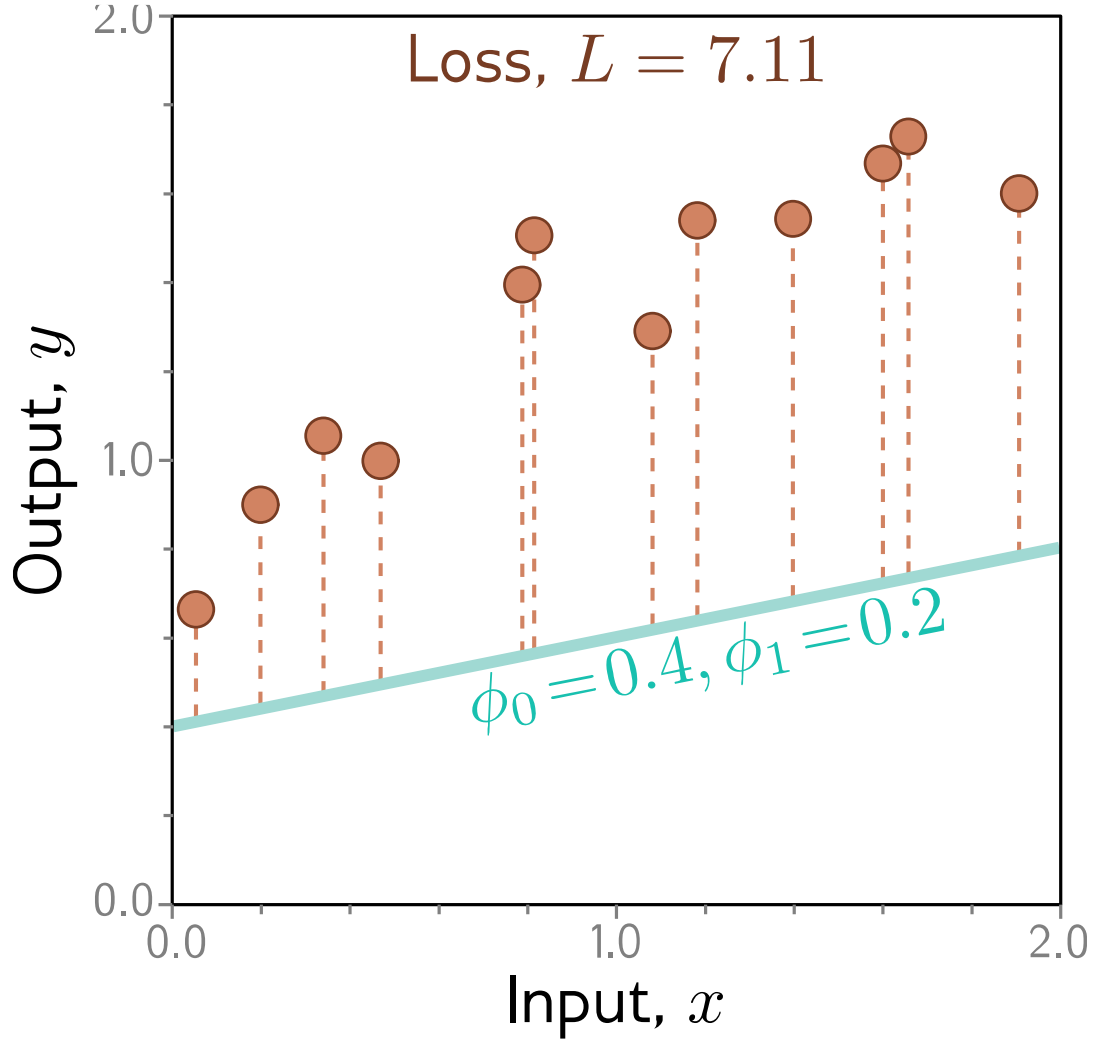
$$L\left[\boldsymbol{\phi}\right]$$

Devuelve un escalar que es más pequeño cuando el modelo asigna mejor las entradas a las salidas

- Encuentra los parámetros que minimizan la pérdida:

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\left[L[\boldsymbol{\phi}]\right]$$

Ejemplo: función de pérdida de regresión lineal 1D



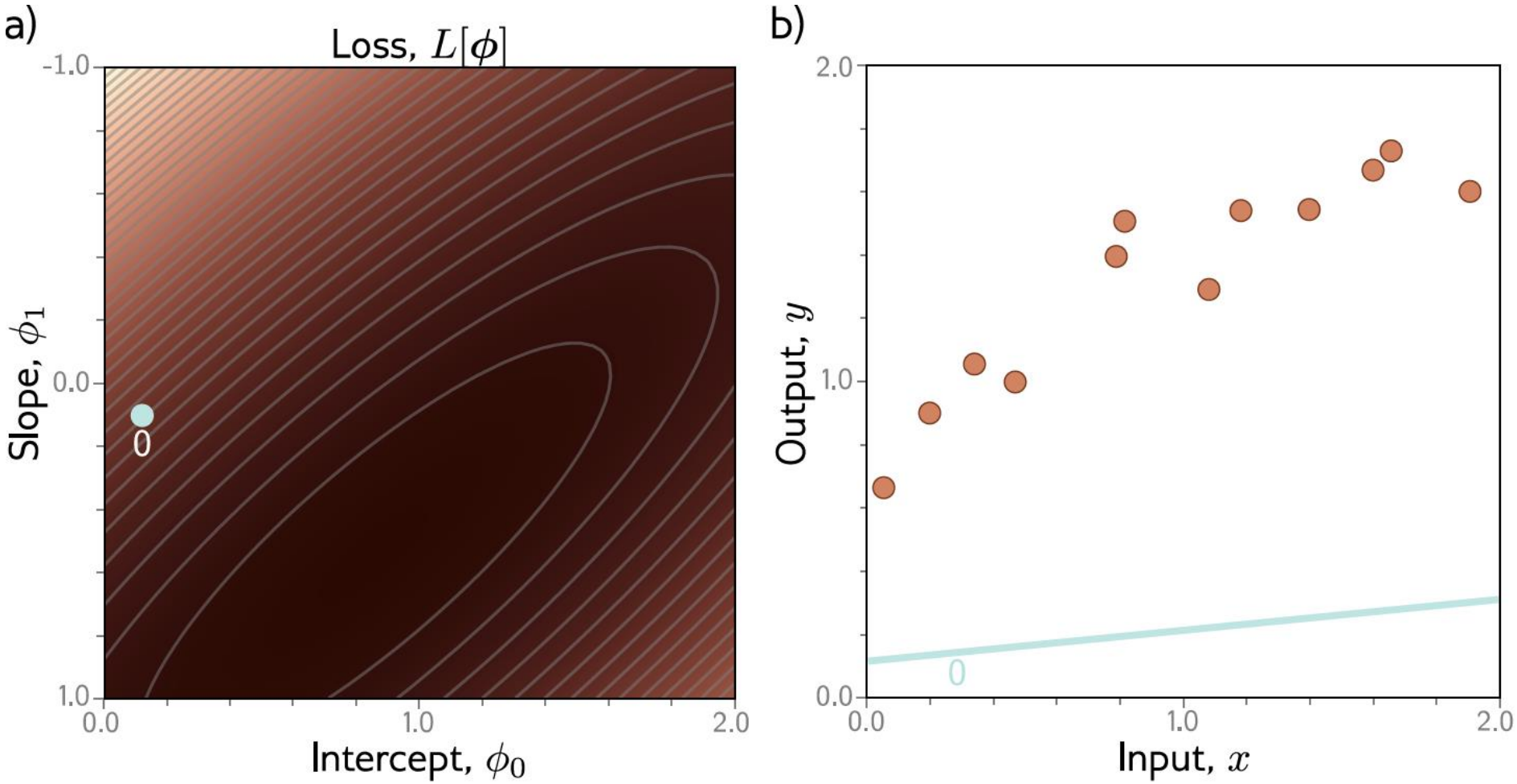
Función de pérdida:

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} (\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i)^2$$

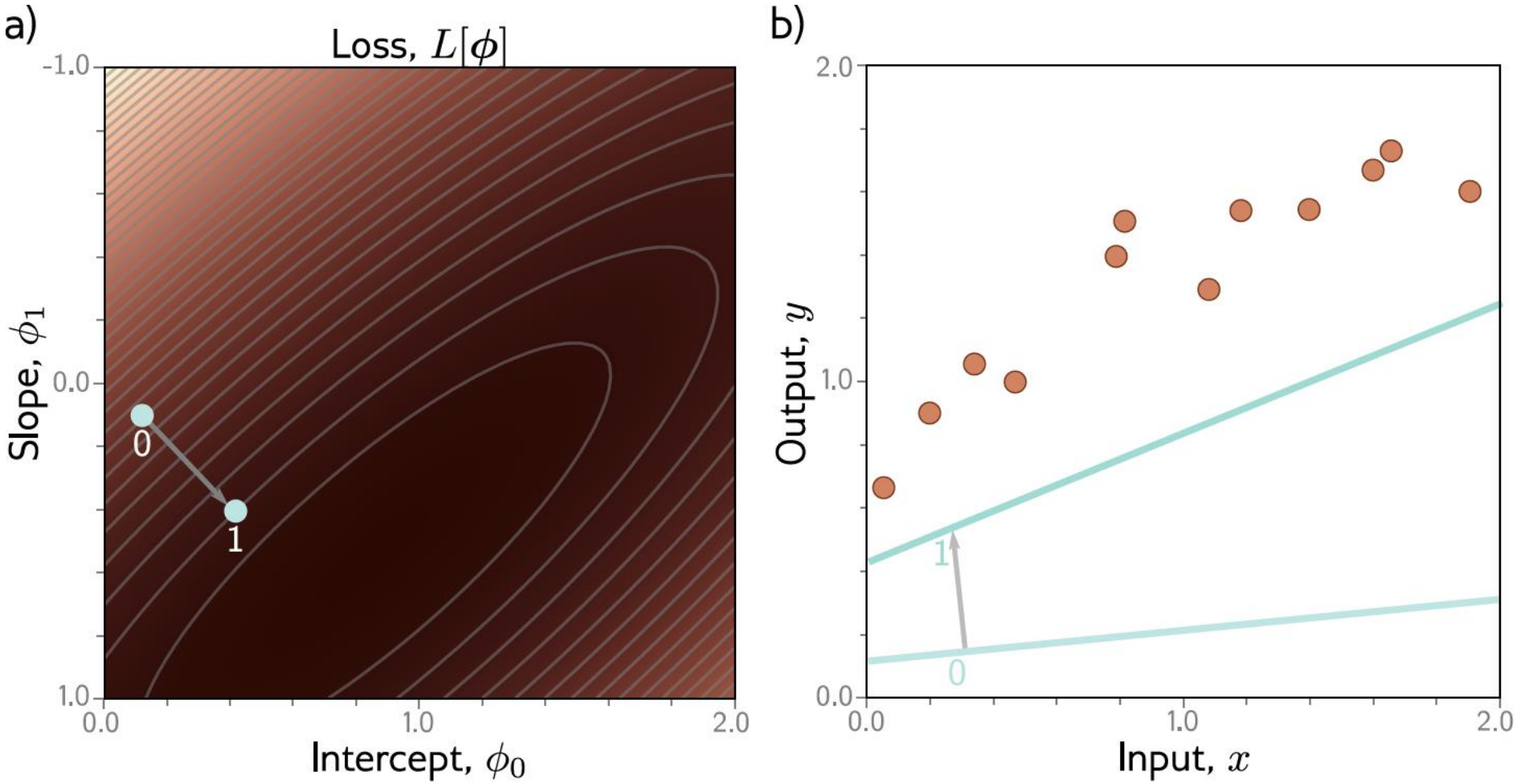
$$= \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2$$

"Función de pérdida por mínimos cuadrados"

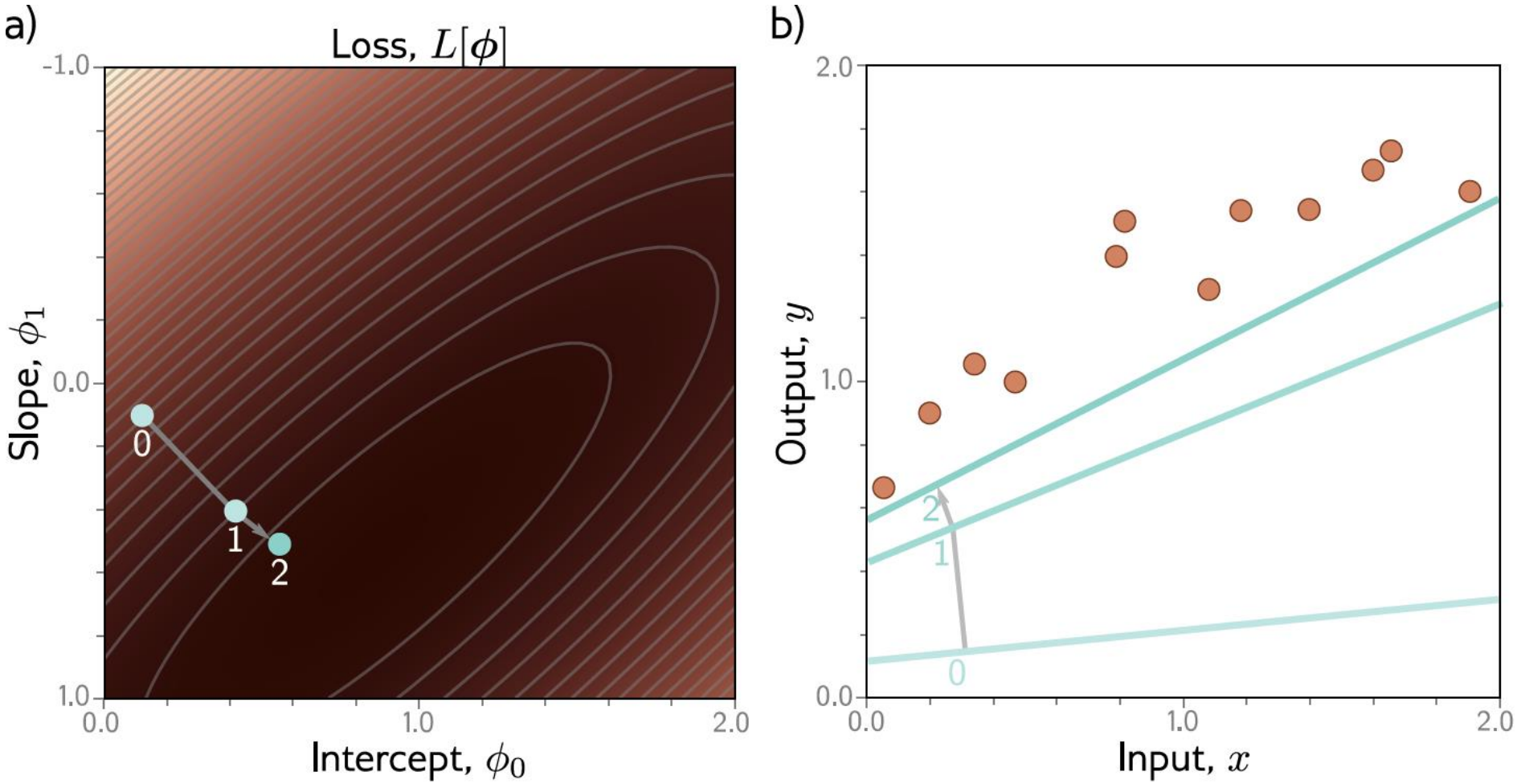
Ejemplo: entrenamiento de regresión lineal 1D



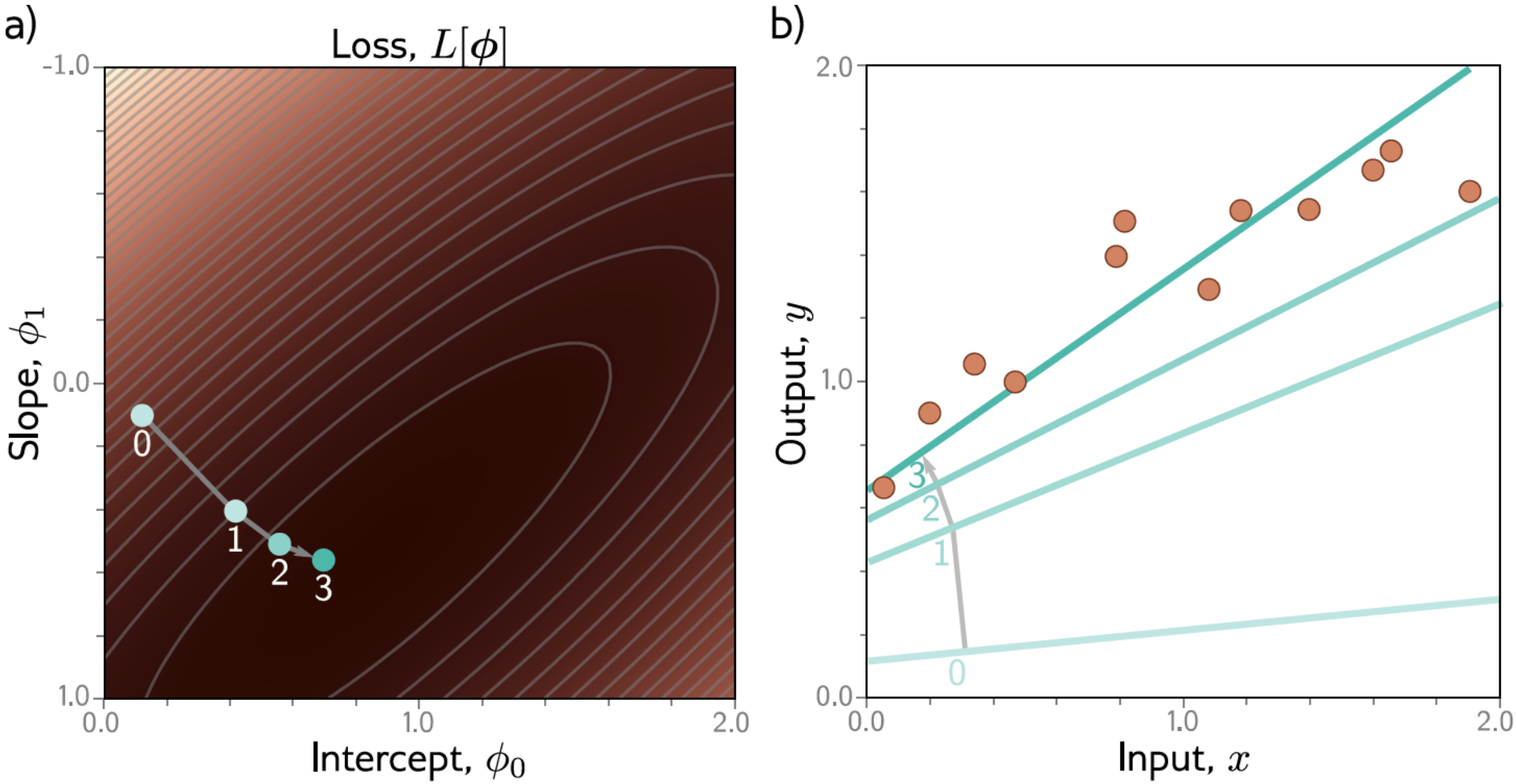
Ejemplo: entrenamiento de regresión lineal 1D



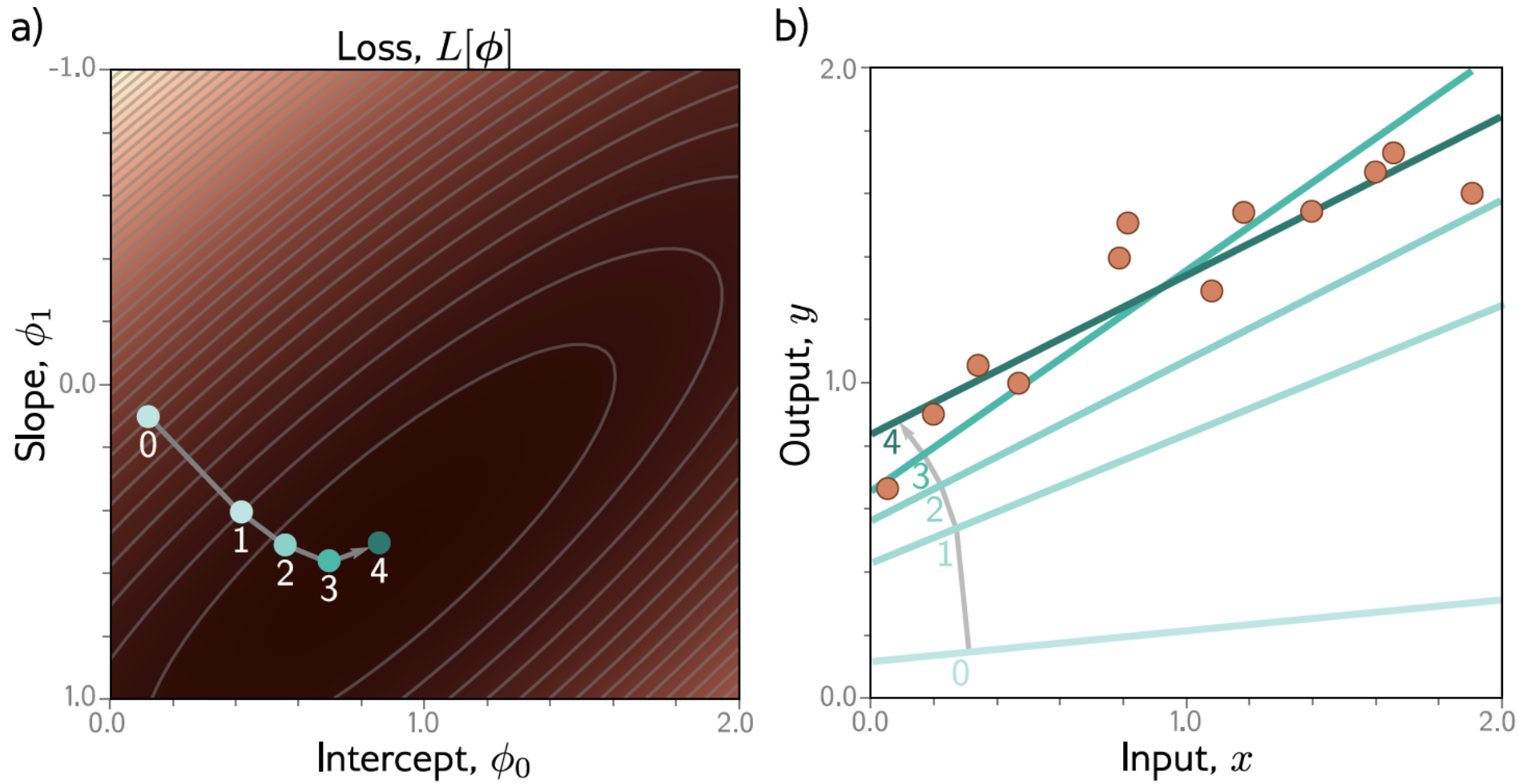
Ejemplo: entrenamiento de regresión lineal 1D



Ejemplo: entrenamiento de regresión lineal 1D



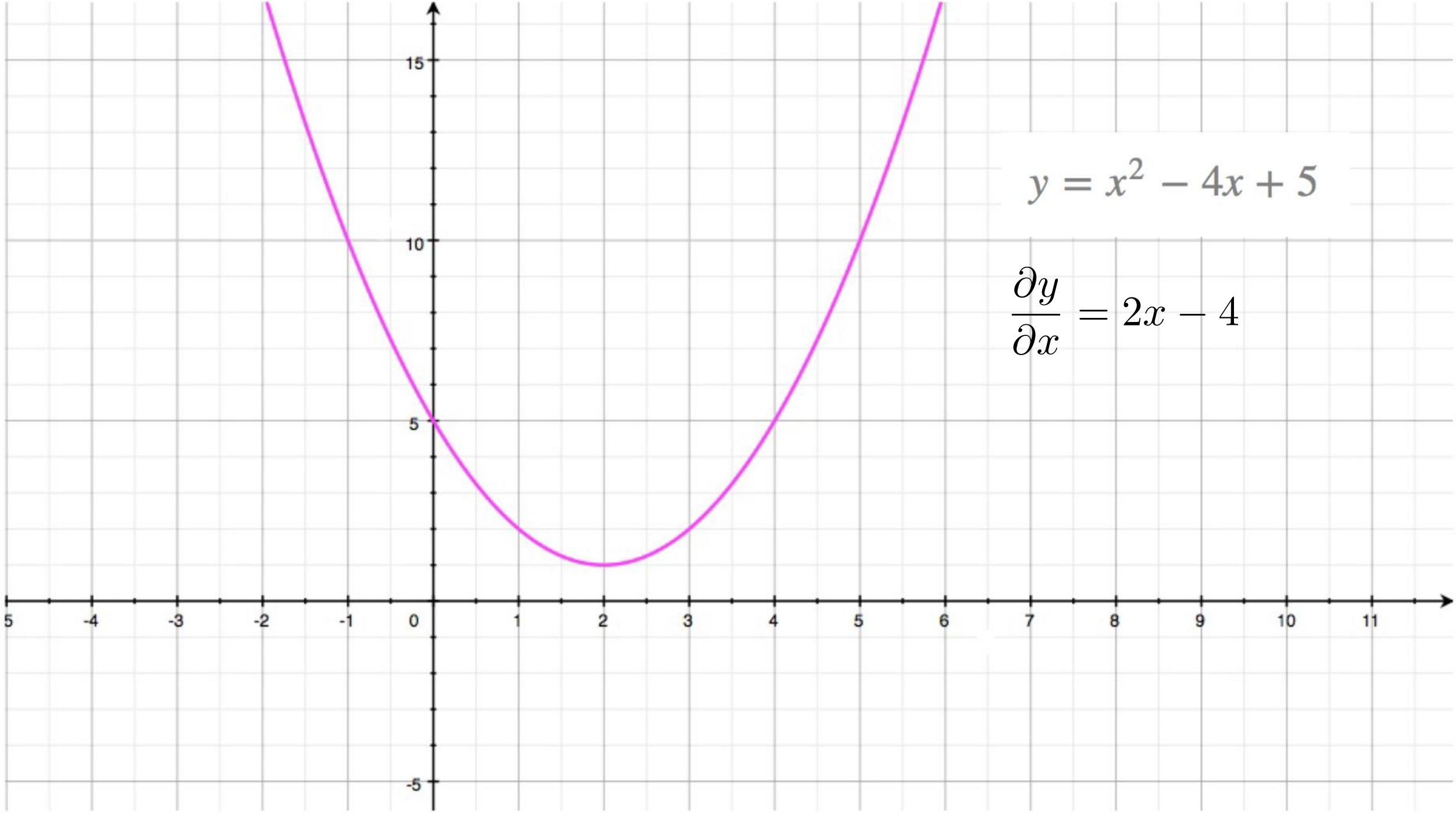
Ejemplo: entrenamiento de regresión lineal 1D



Esta técnica se conoce como descenso por gradiente

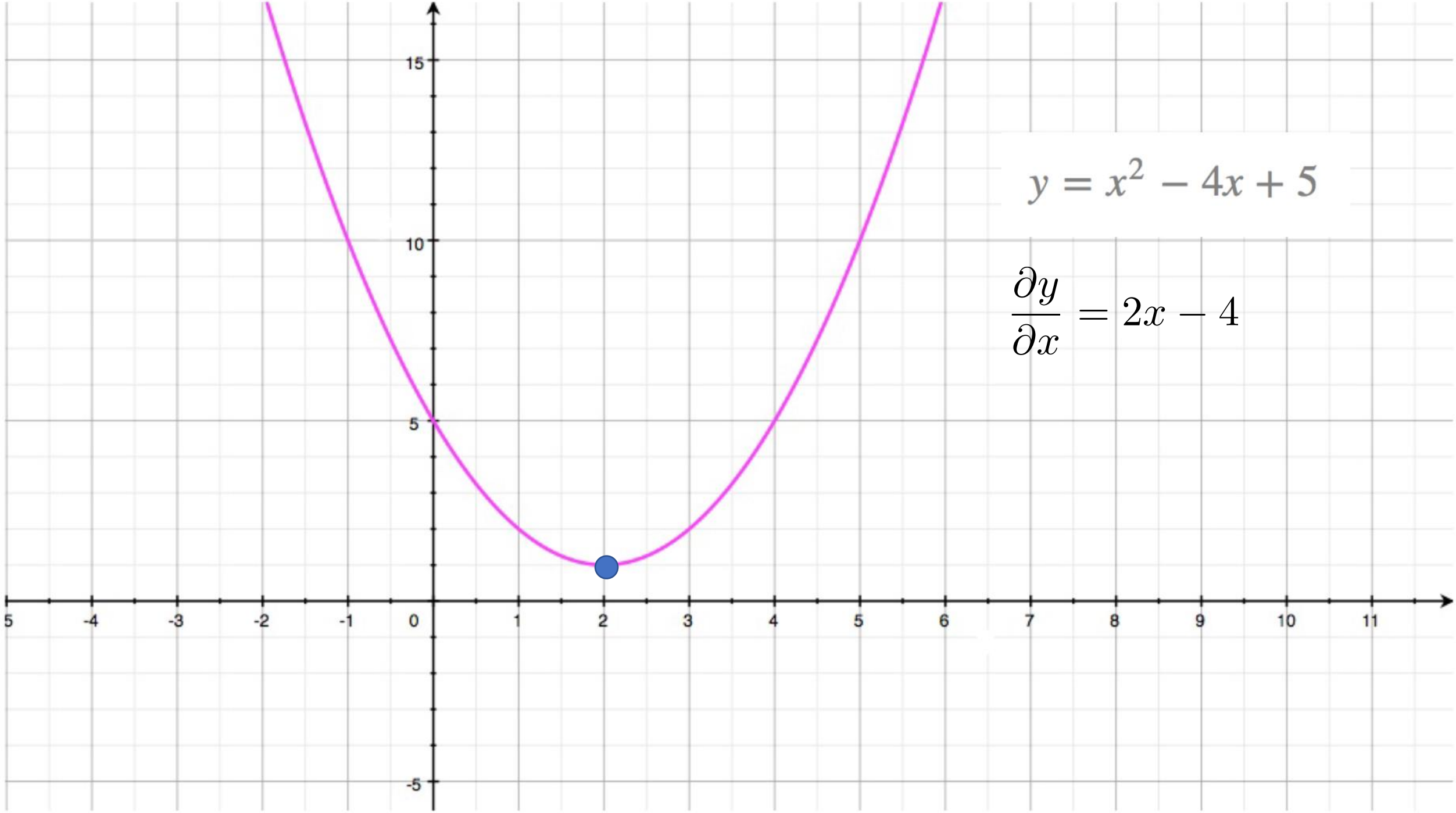
Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam



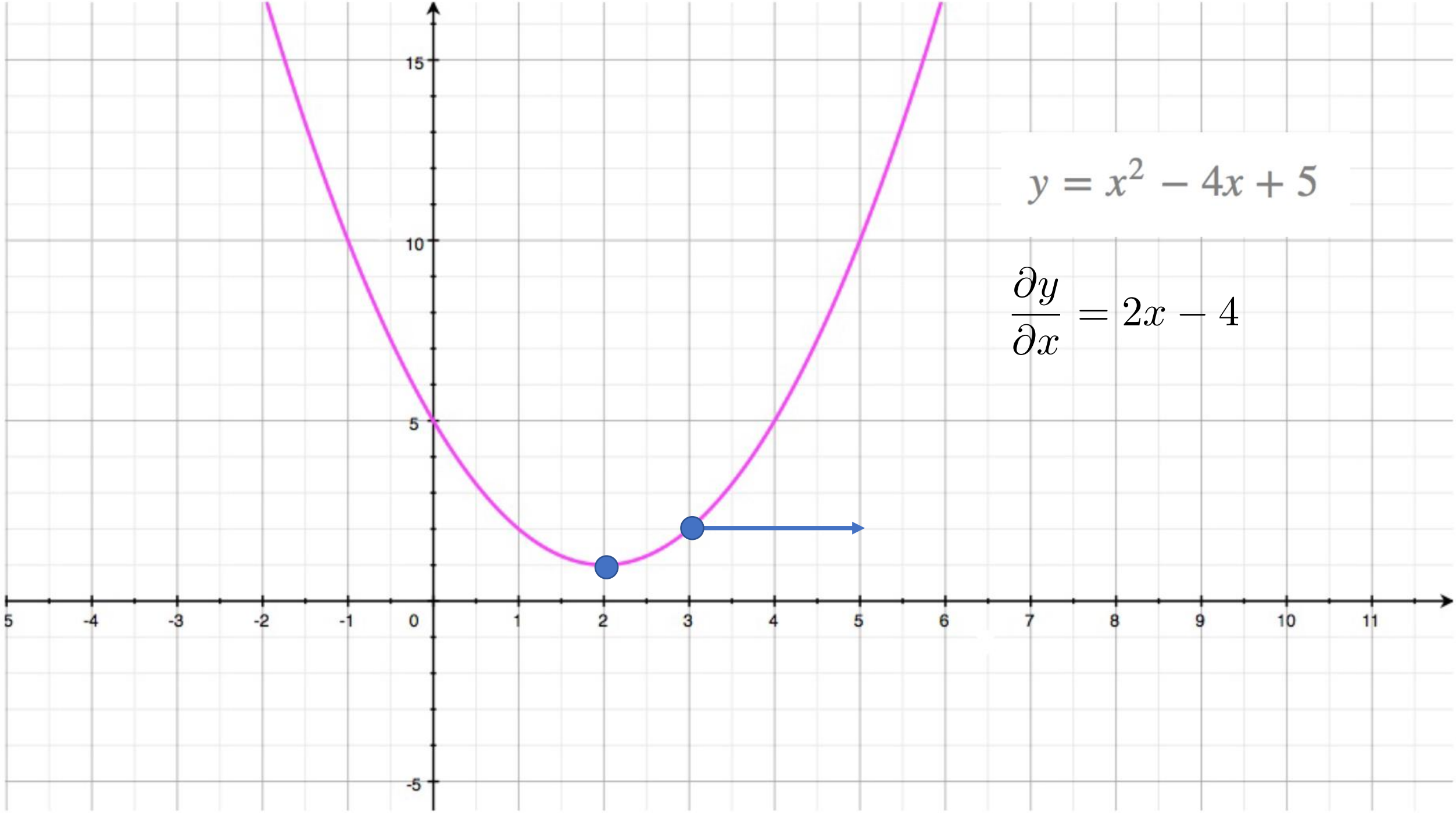
$$y = x^2 - 4x + 5$$

$$\frac{\partial y}{\partial x} = 2x - 4$$



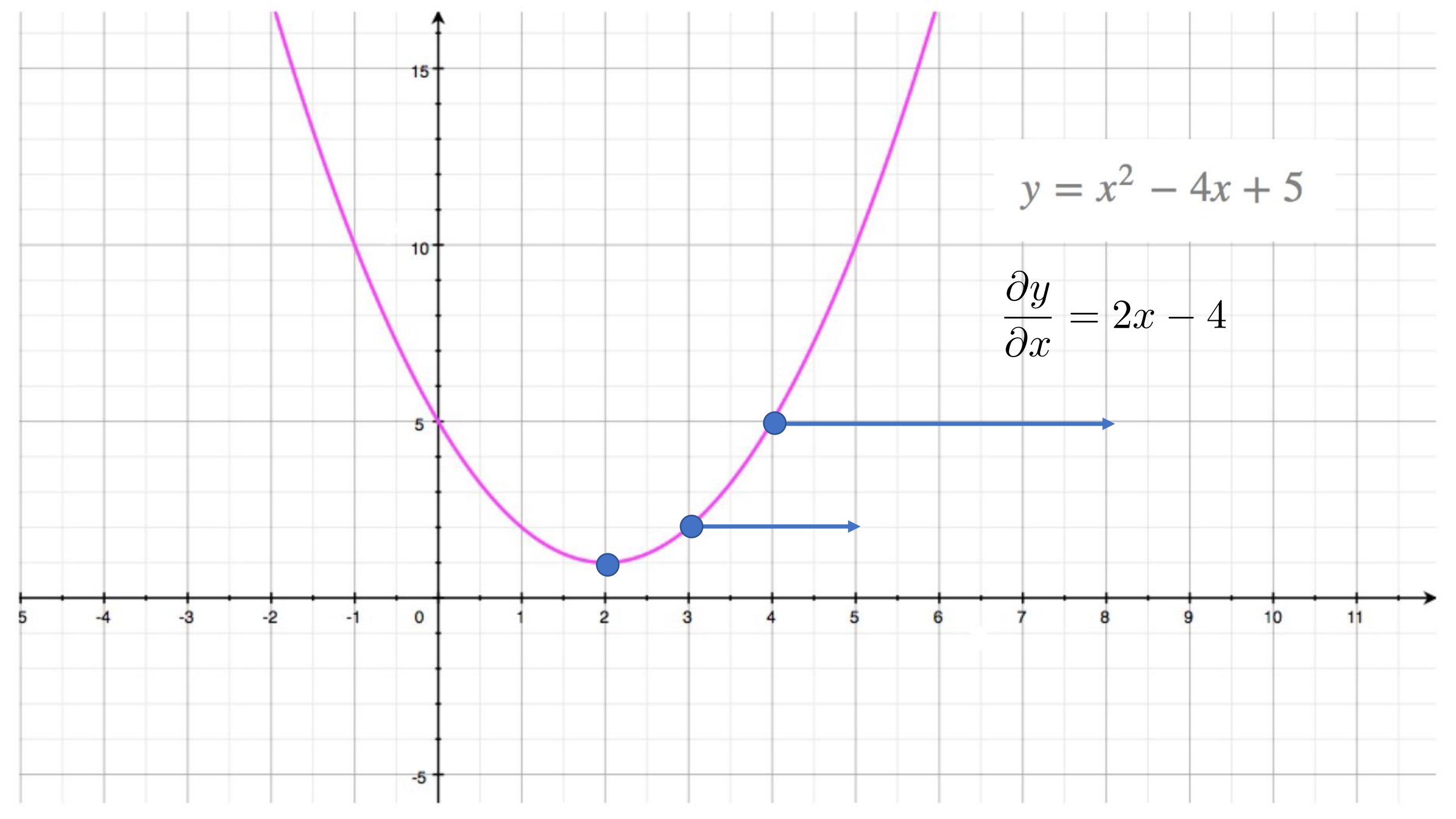
$$y = x^2 - 4x + 5$$

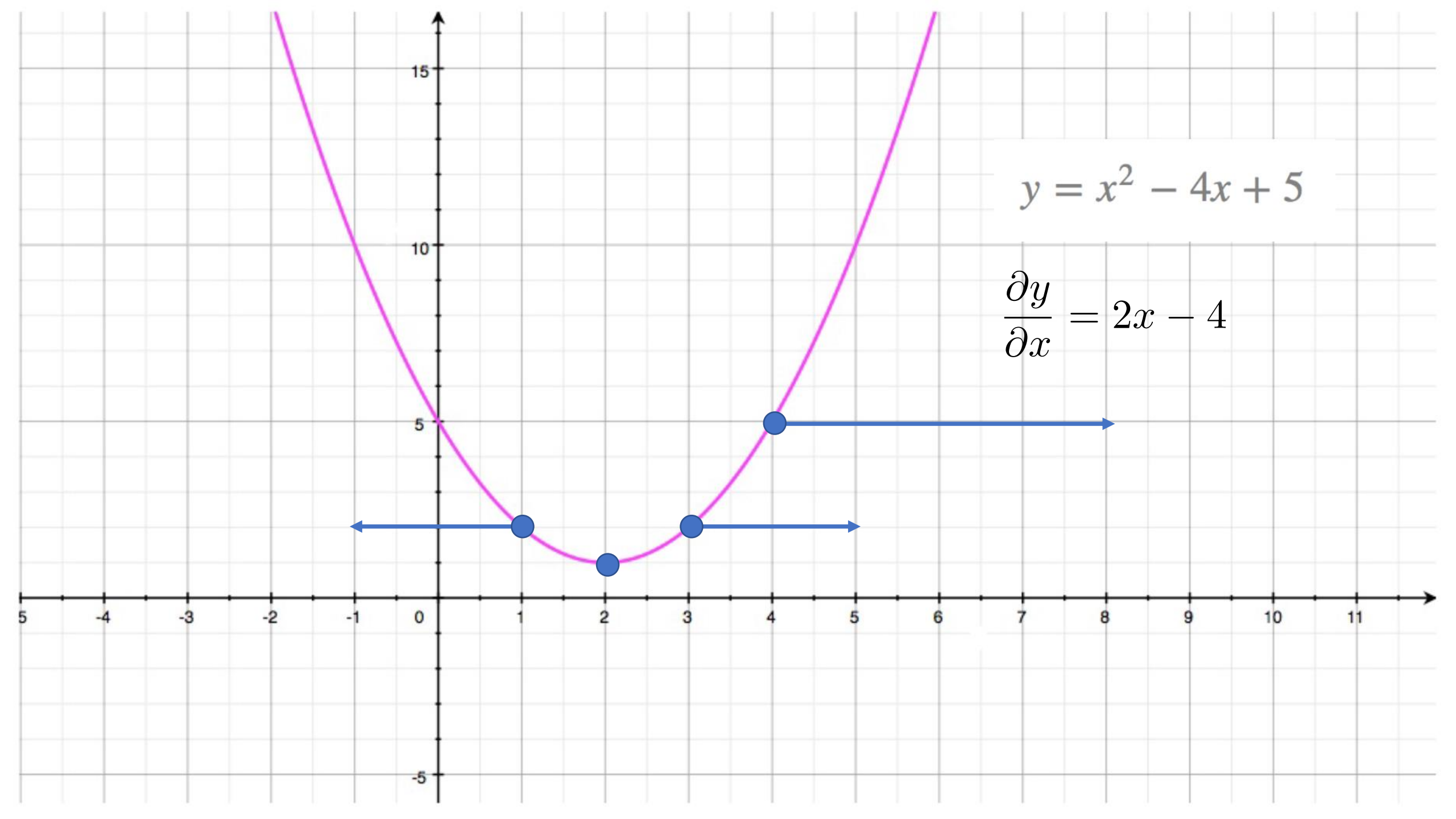
$$\frac{\partial y}{\partial x} = 2x - 4$$

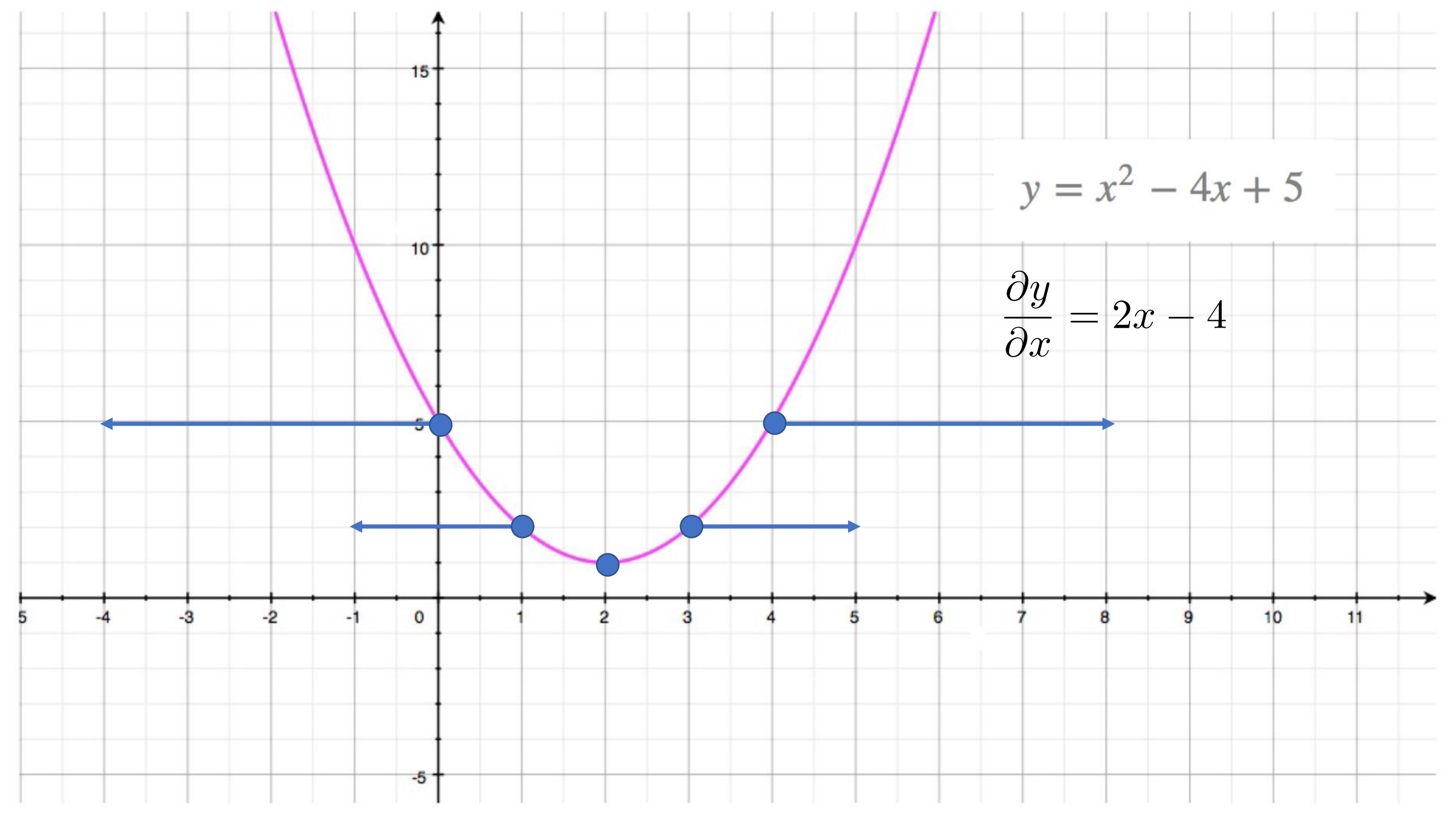


$$y = x^2 - 4x + 5$$

$$\frac{\partial y}{\partial x} = 2x - 4$$


15
10
5
-5
5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
11
$y = x^2 - 4x + 5$
$\frac{\partial y}{\partial x} = 2x - 4$


$y = x^2 - 4x + 5$
$\frac{\partial y}{\partial x} = 2x - 4$
15
10
5
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
11


$y = x^2 - 4x + 5$
$\frac{\partial y}{\partial x} = 2x - 4$
15
10
5
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
7
8
9
10
11

Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam

Algoritmo de descenso gradiente

Step 1. Compute the derivatives of the loss with respect to the parameters:

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial L}{\partial \phi_0} \\ \frac{\partial L}{\partial \phi_1} \\ \vdots \\ \frac{\partial L}{\partial \phi_N} \end{bmatrix}.$$

Step 2. Update the parameters according to the rule:

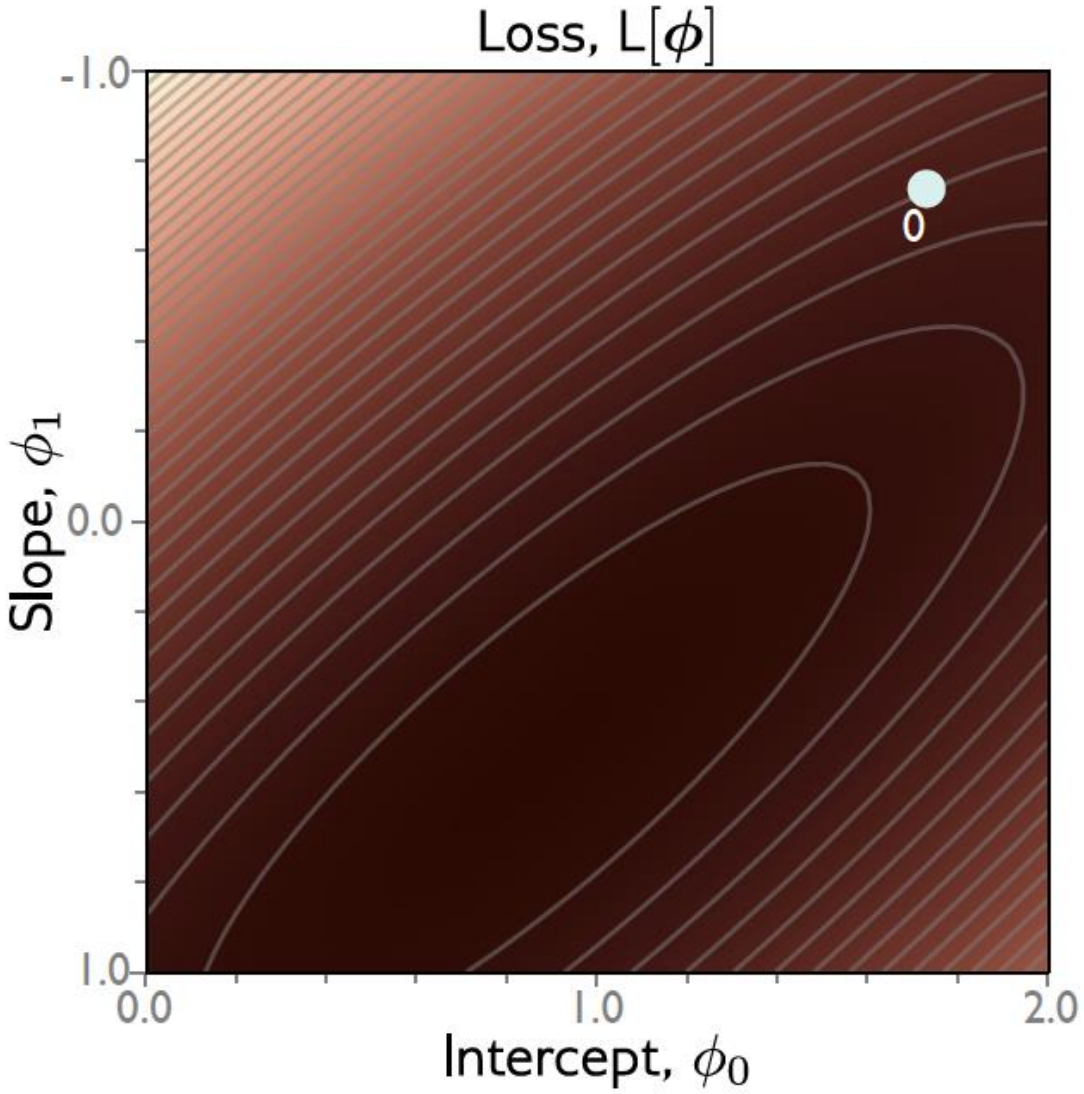
$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \frac{\partial L}{\partial \boldsymbol{\phi}},$$

where the positive scalar α determines the magnitude of the change.

Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam

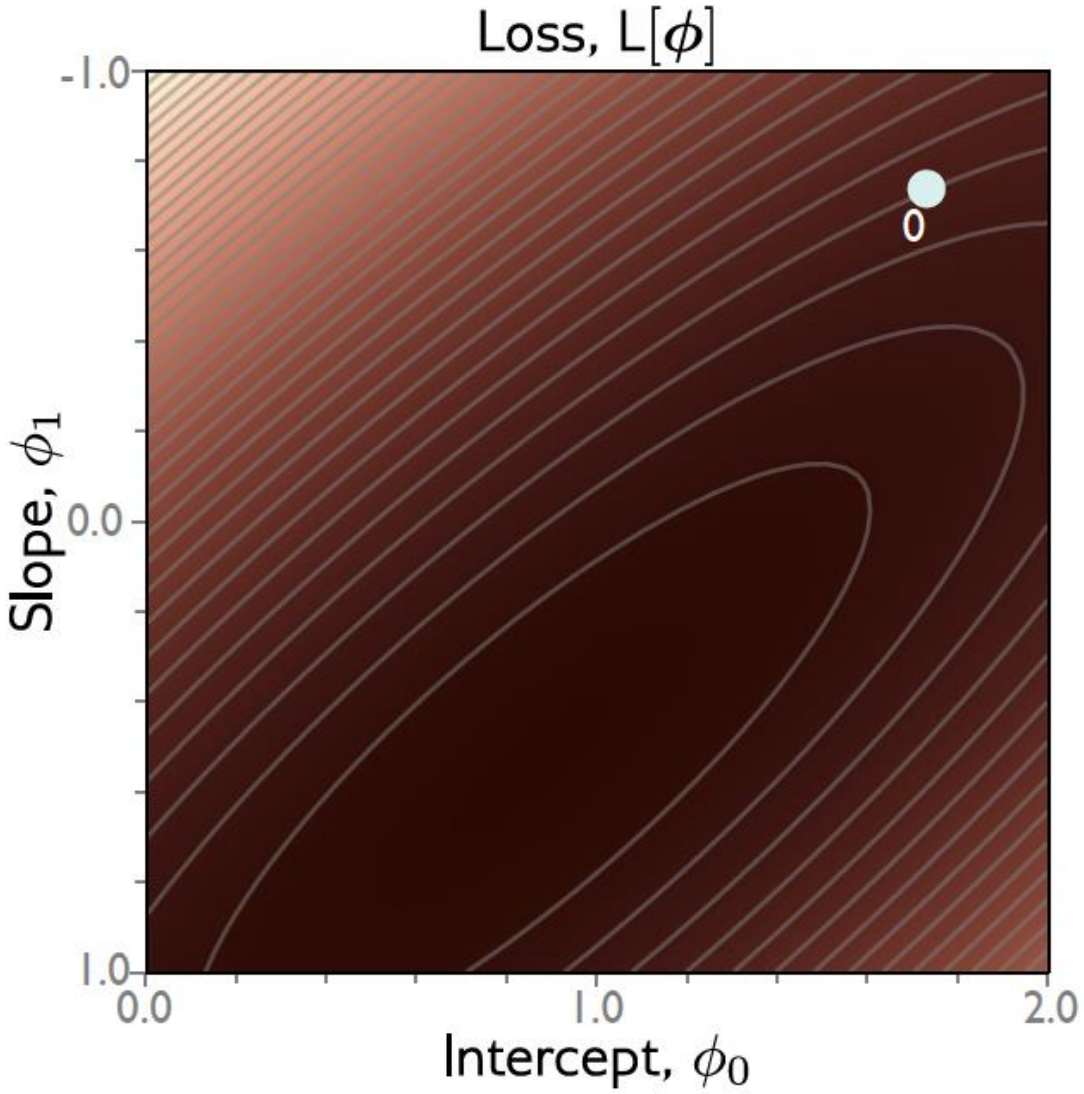
Descenso gradual



Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$
\begin{aligned}
L[\boldsymbol{\phi}] &= \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2 \\
&= \sum_{i=1}^{I} \left(\phi_0 + \phi_1 x_i - y_i\right)^2
\end{aligned}
$$

Descenso gradual

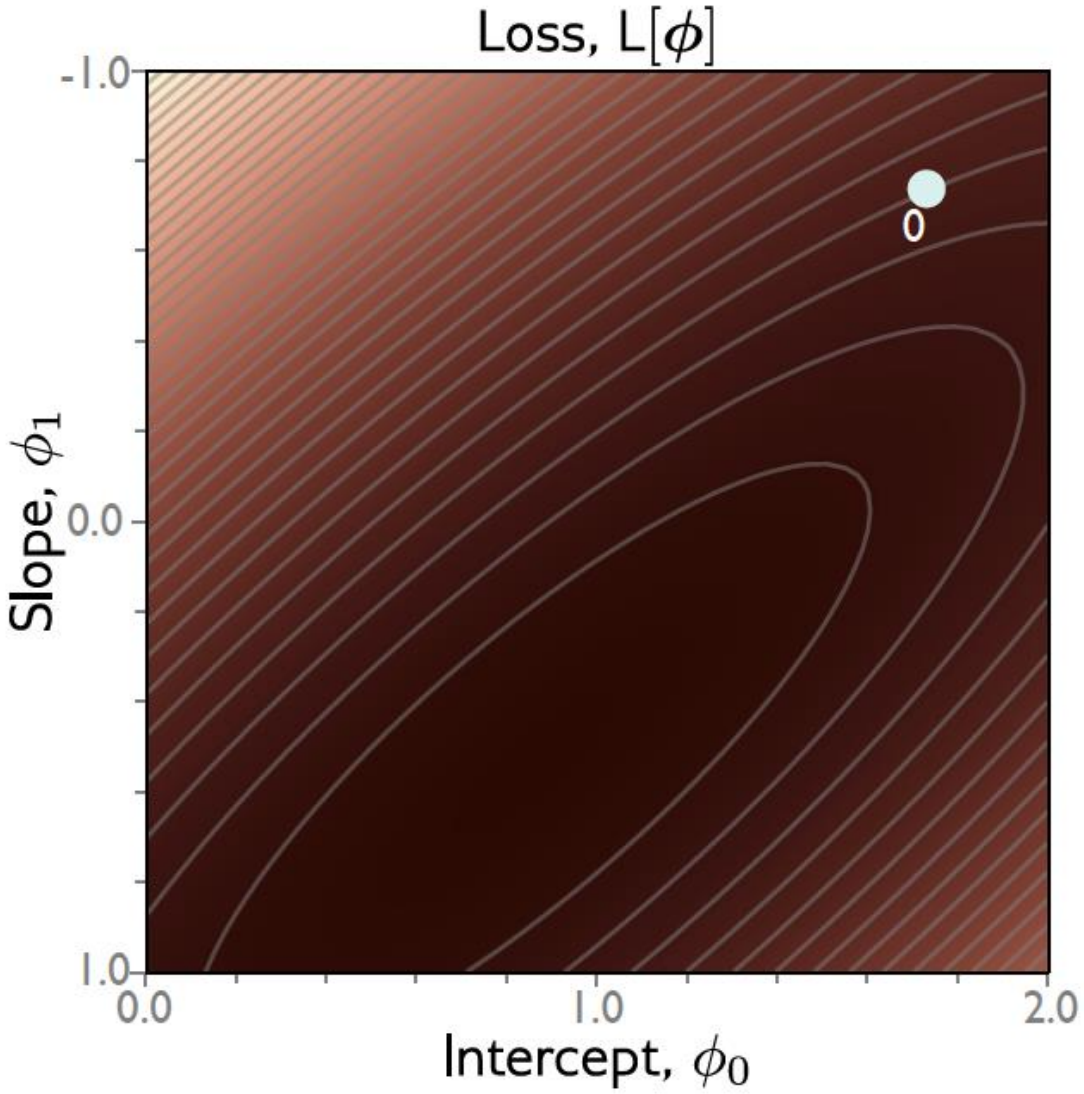


Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} (\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i)^2$$
$$= \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2$$

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

Descenso gradual



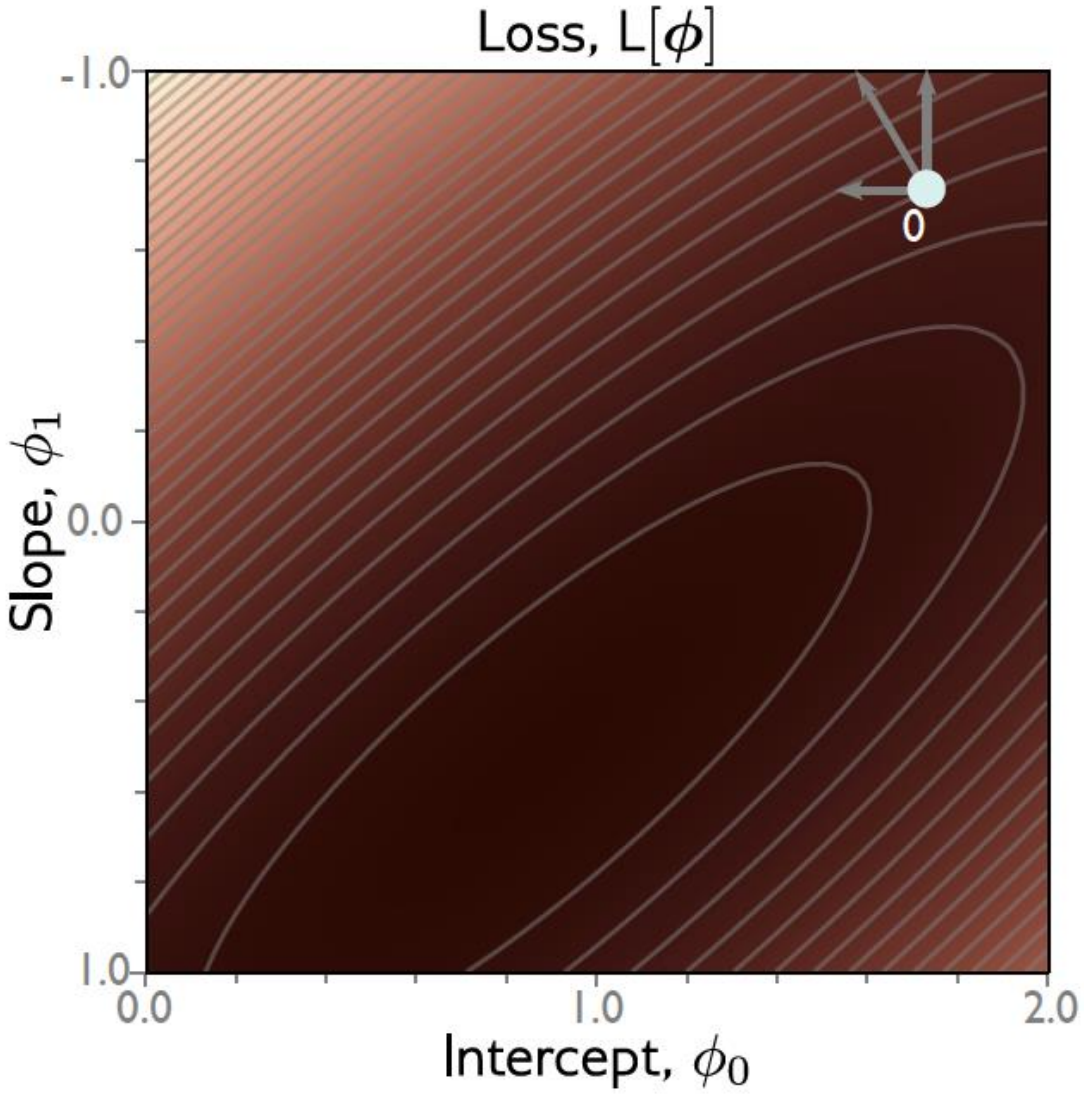
Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\begin{aligned} L[\boldsymbol{\phi}] &= \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} (\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i)^2 \\ &= \sum_{i=1}^{I} (\phi_0 + \phi_1 x_i - y_i)^2 \end{aligned}$$

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

Descenso gradual

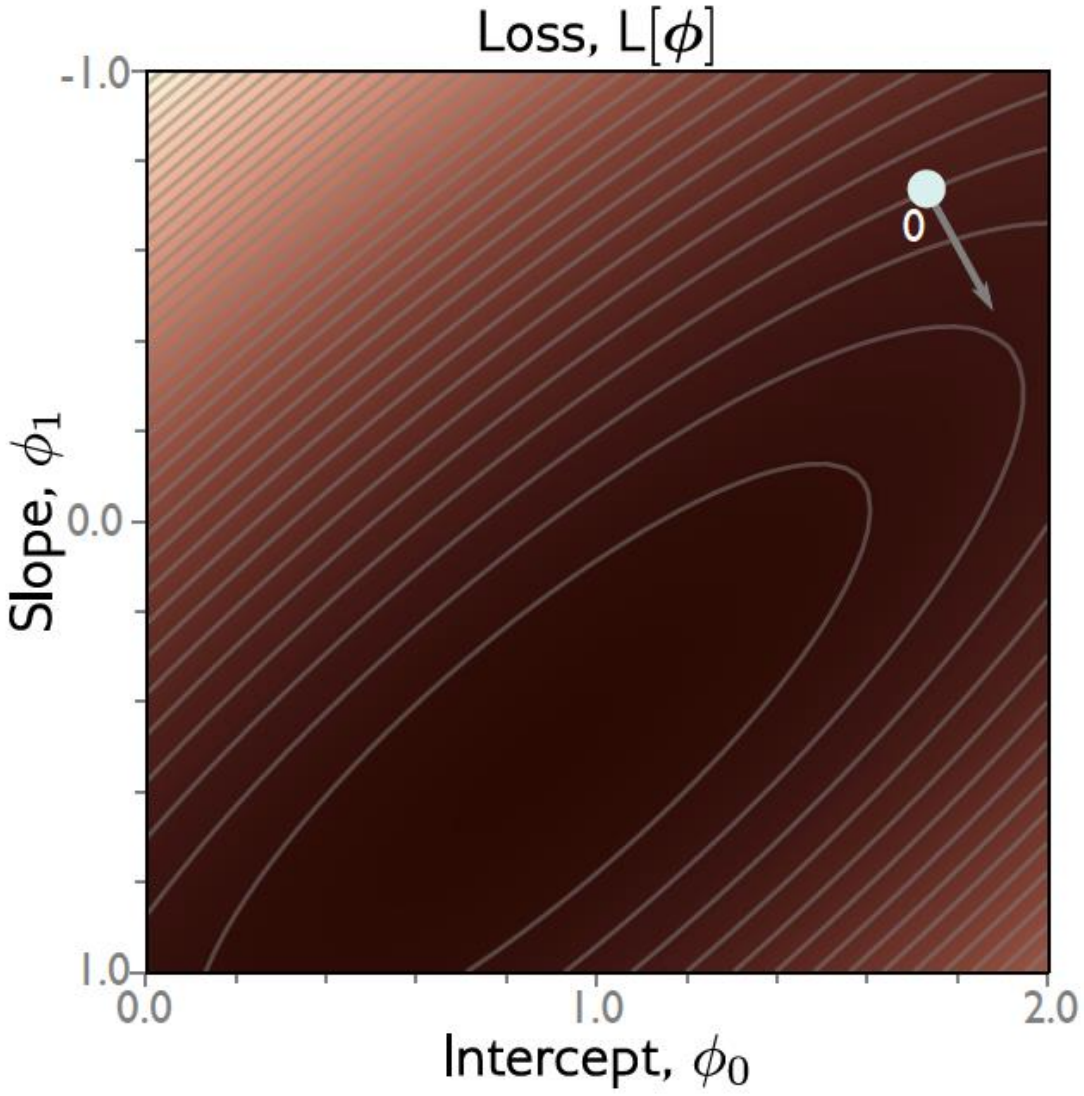


Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

Descenso gradual



Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

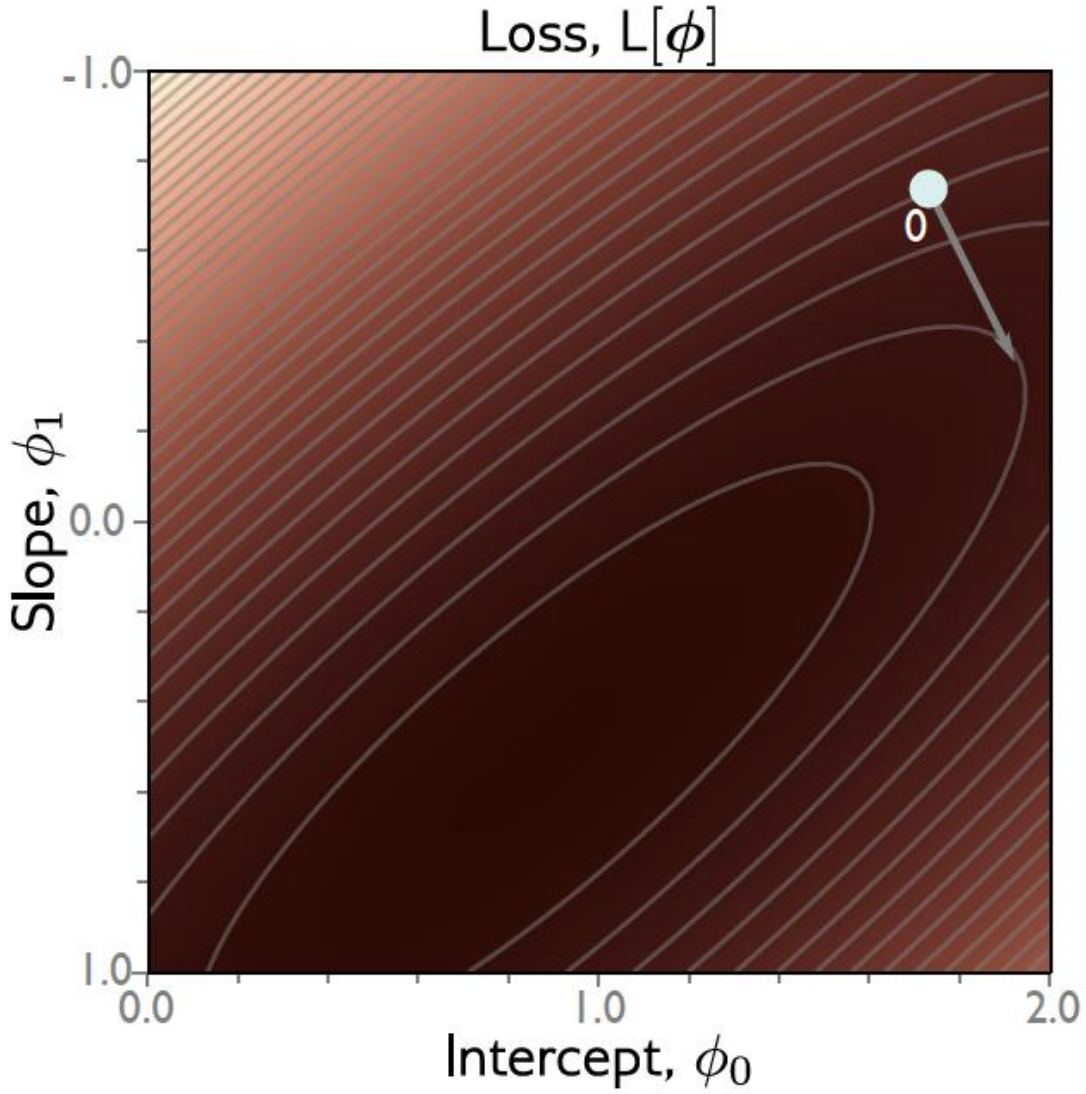
$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

Paso 2: Actualizar los parámetros según la regla

$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \frac{\partial L}{\partial \boldsymbol{\phi}}$$

= tamaño del paso o ritmo de aprendizaje si es fijo

Descenso gradual



Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

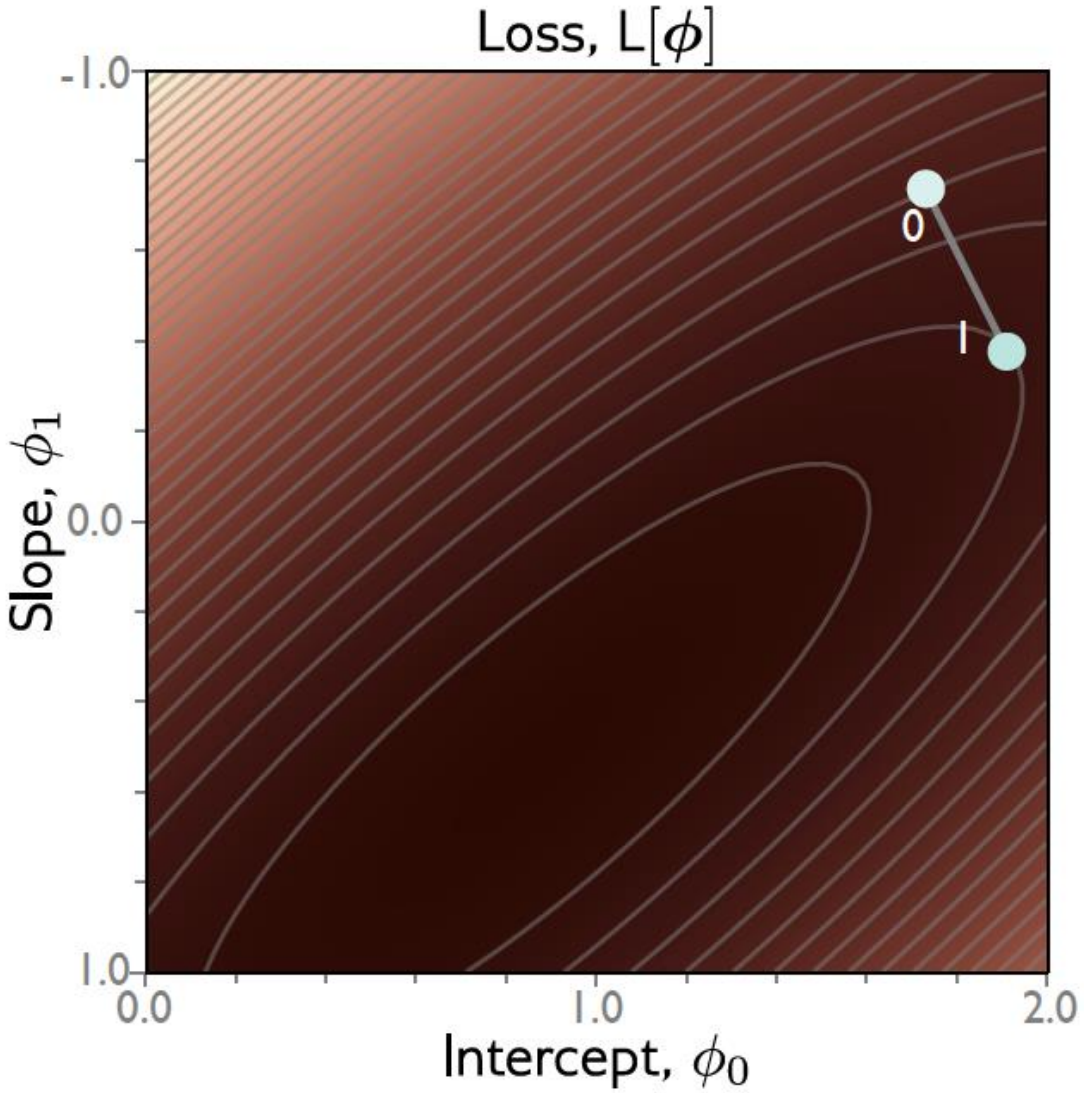
$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

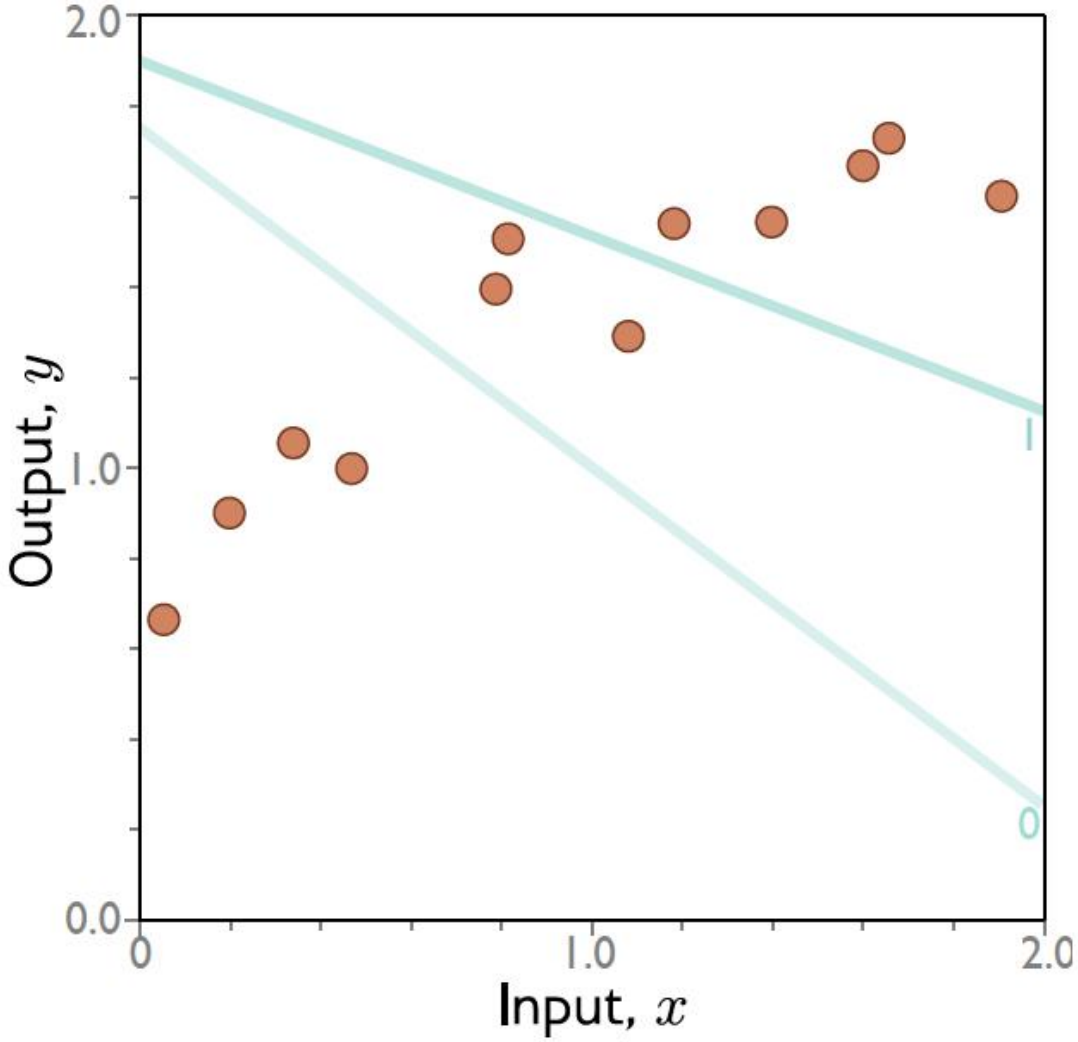
Paso 2: Actualizar los parámetros según la regla

$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \frac{\partial L}{\partial \boldsymbol{\phi}}$$

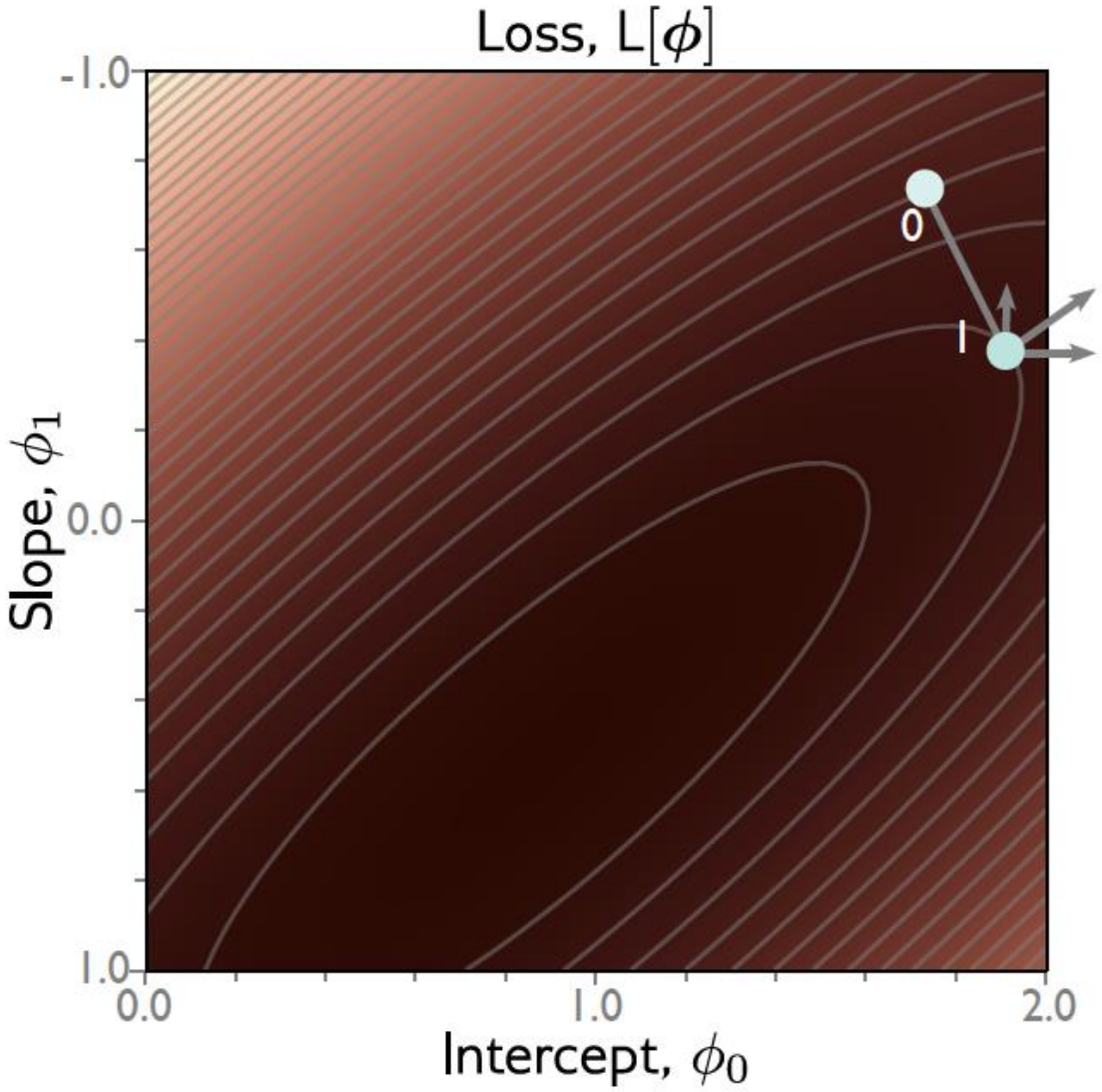
= tamaño del paso

Descenso gradual





Descenso gradual



Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

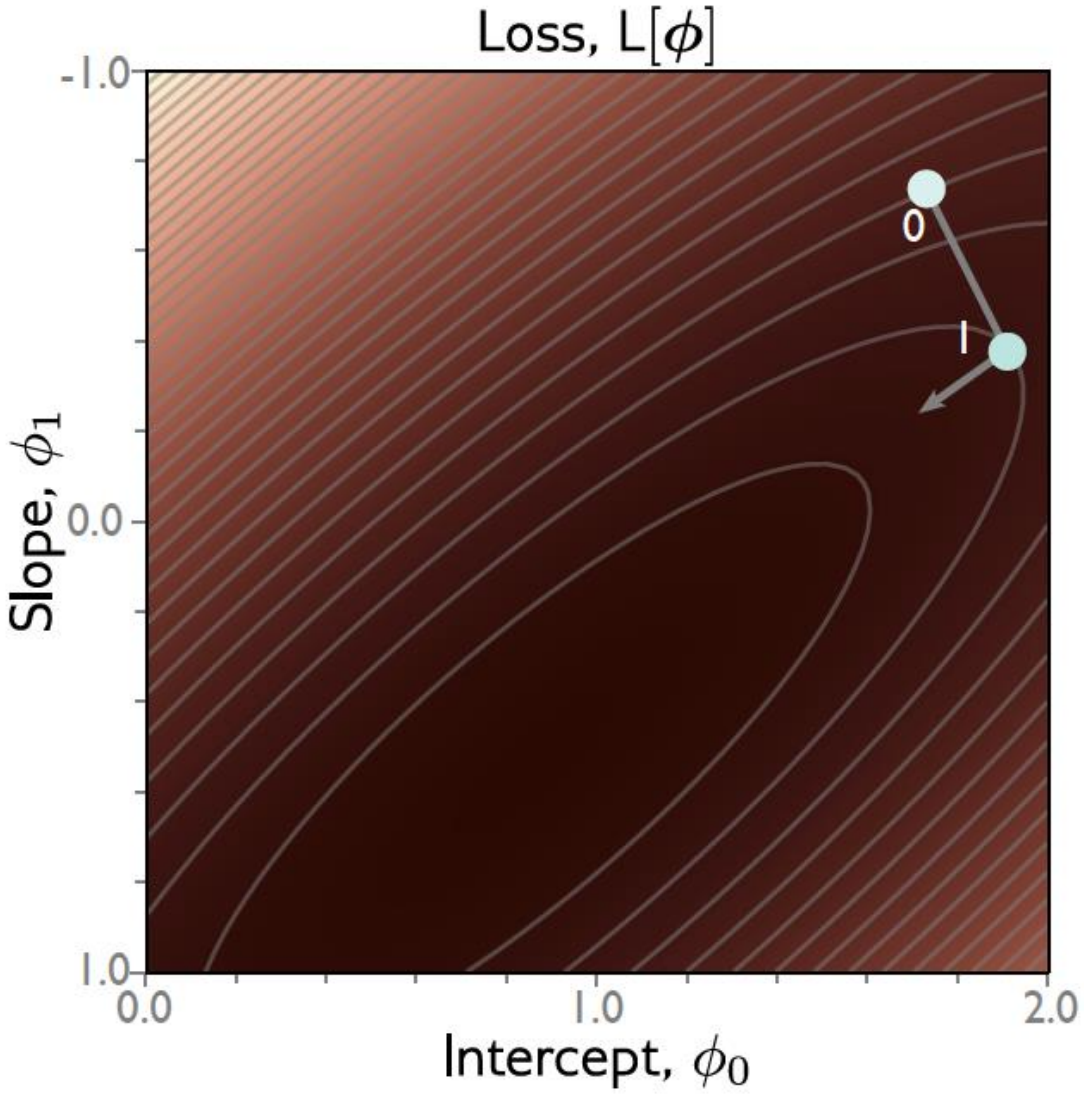
$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

Paso 2: Actualizar los parámetros según la regla

$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \frac{\partial L}{\partial \boldsymbol{\phi}}$$

= tamaño del paso

Descenso gradual



Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

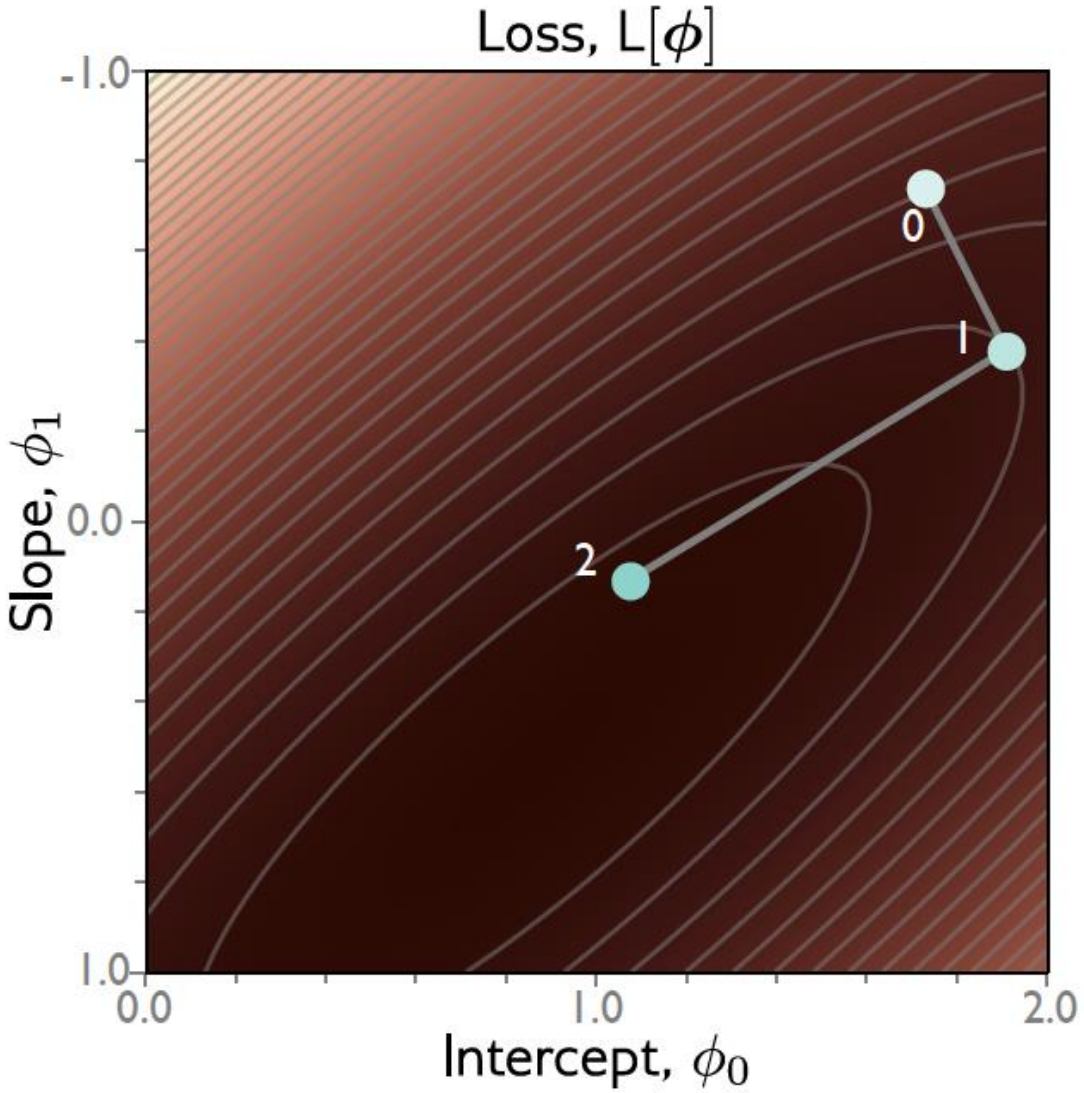
$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

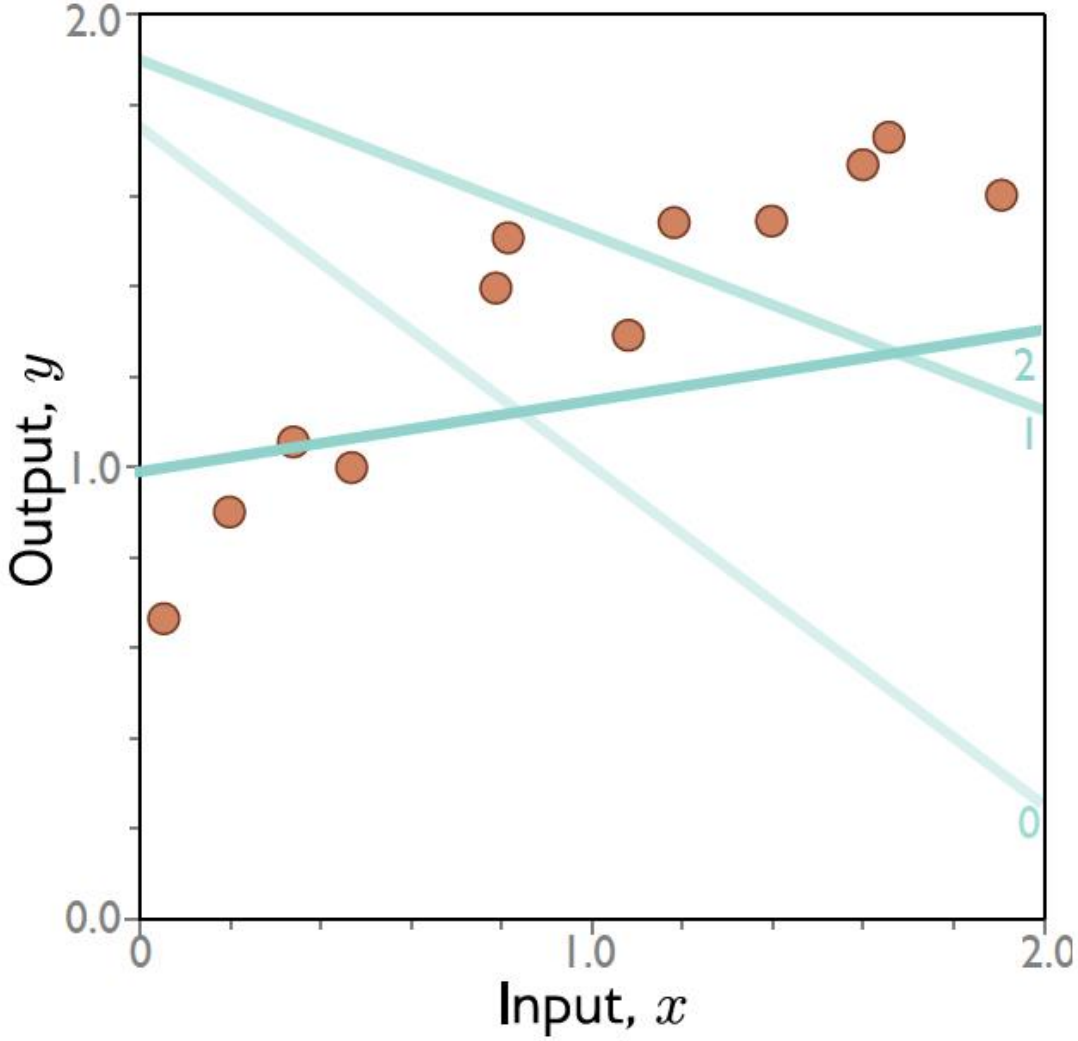
Paso 2: Actualizar los parámetros según la regla

$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \frac{\partial L}{\partial \boldsymbol{\phi}}$$

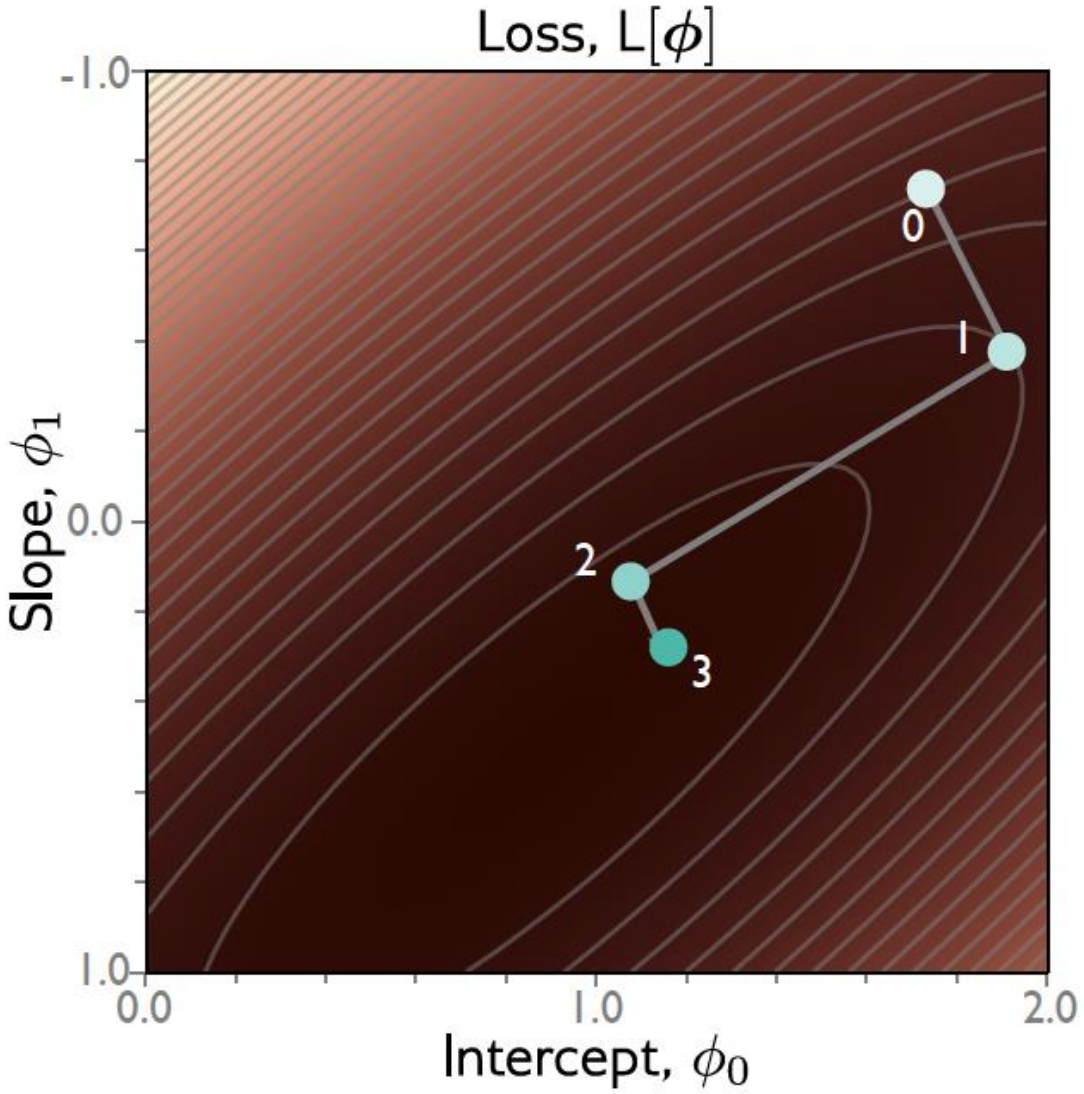
= tamaño del paso

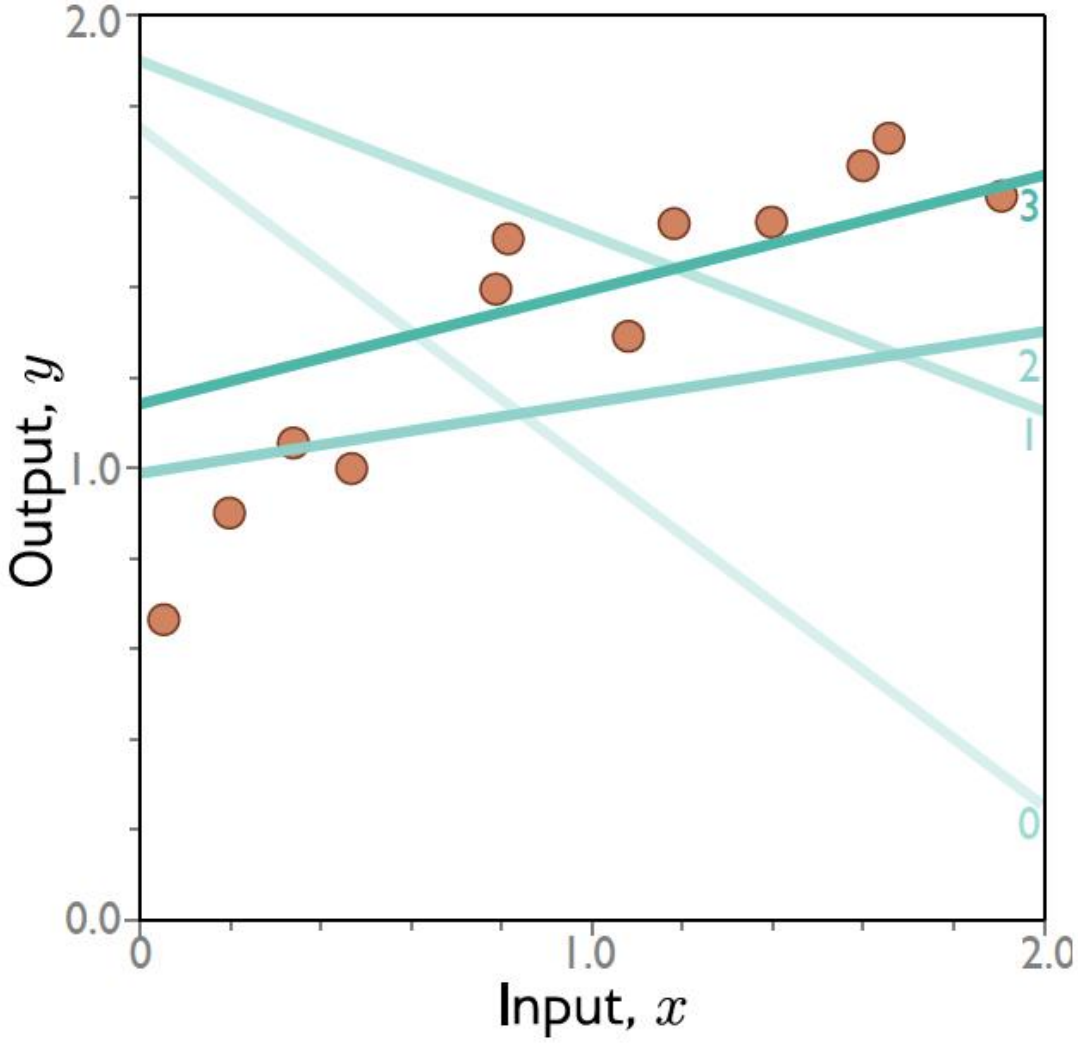
Descenso gradual



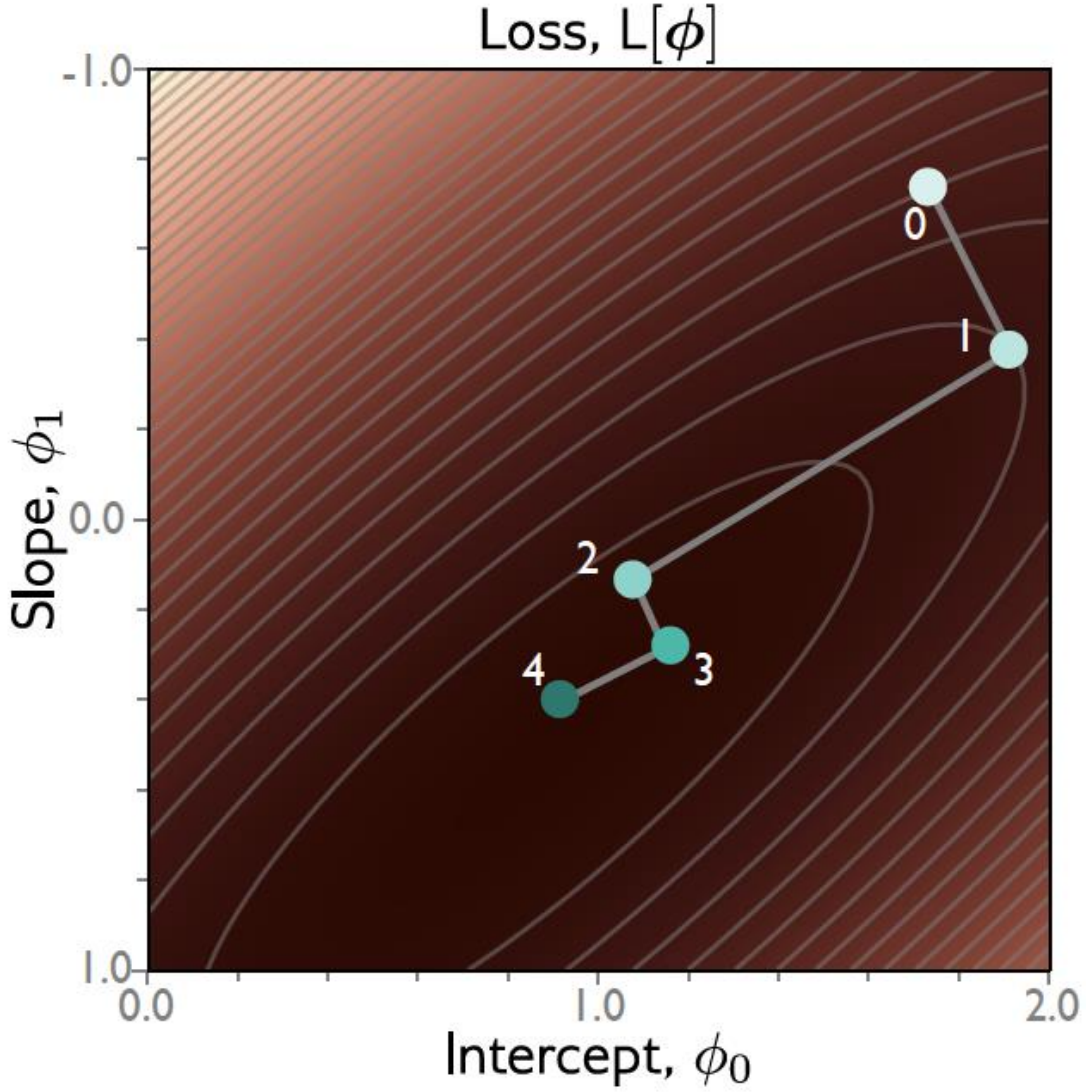


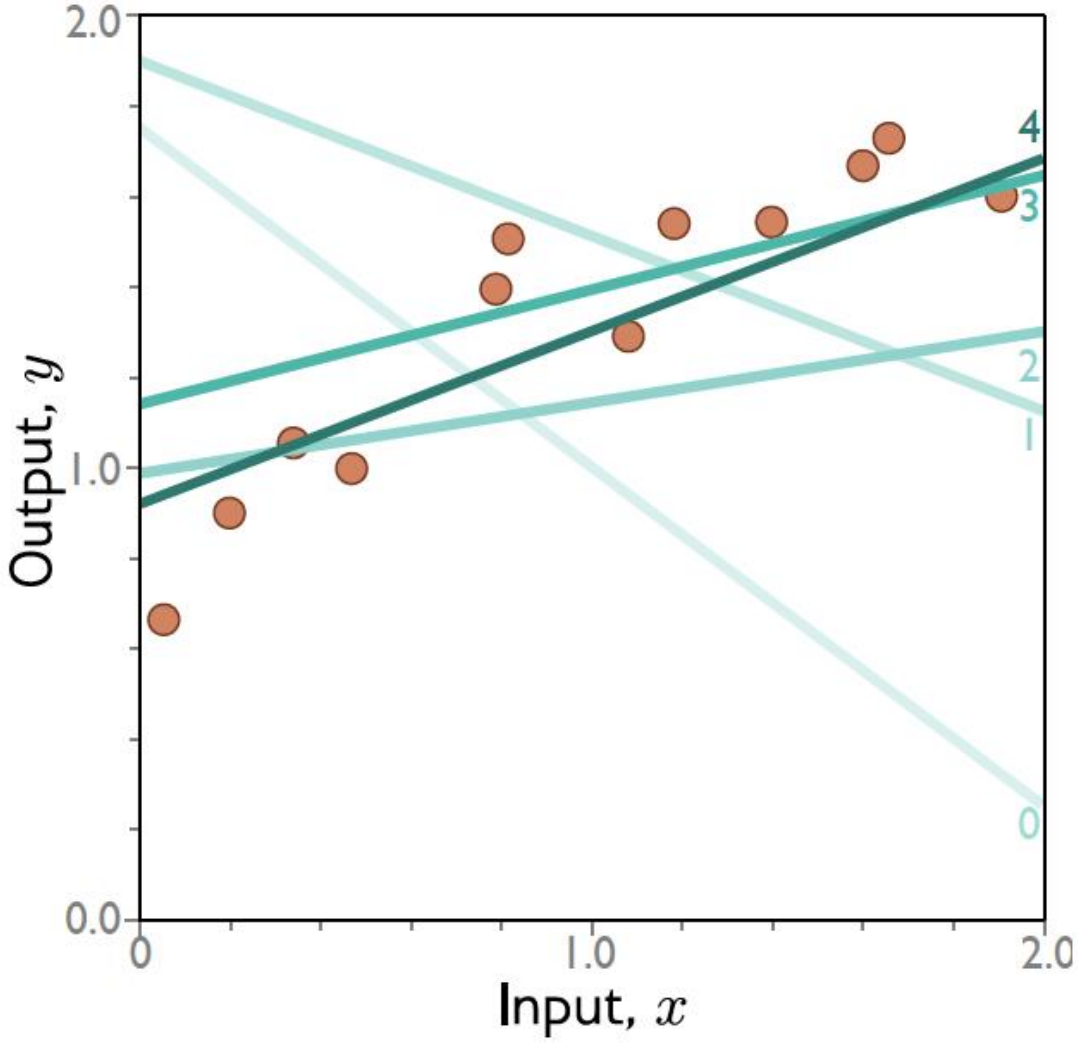
Descenso gradual



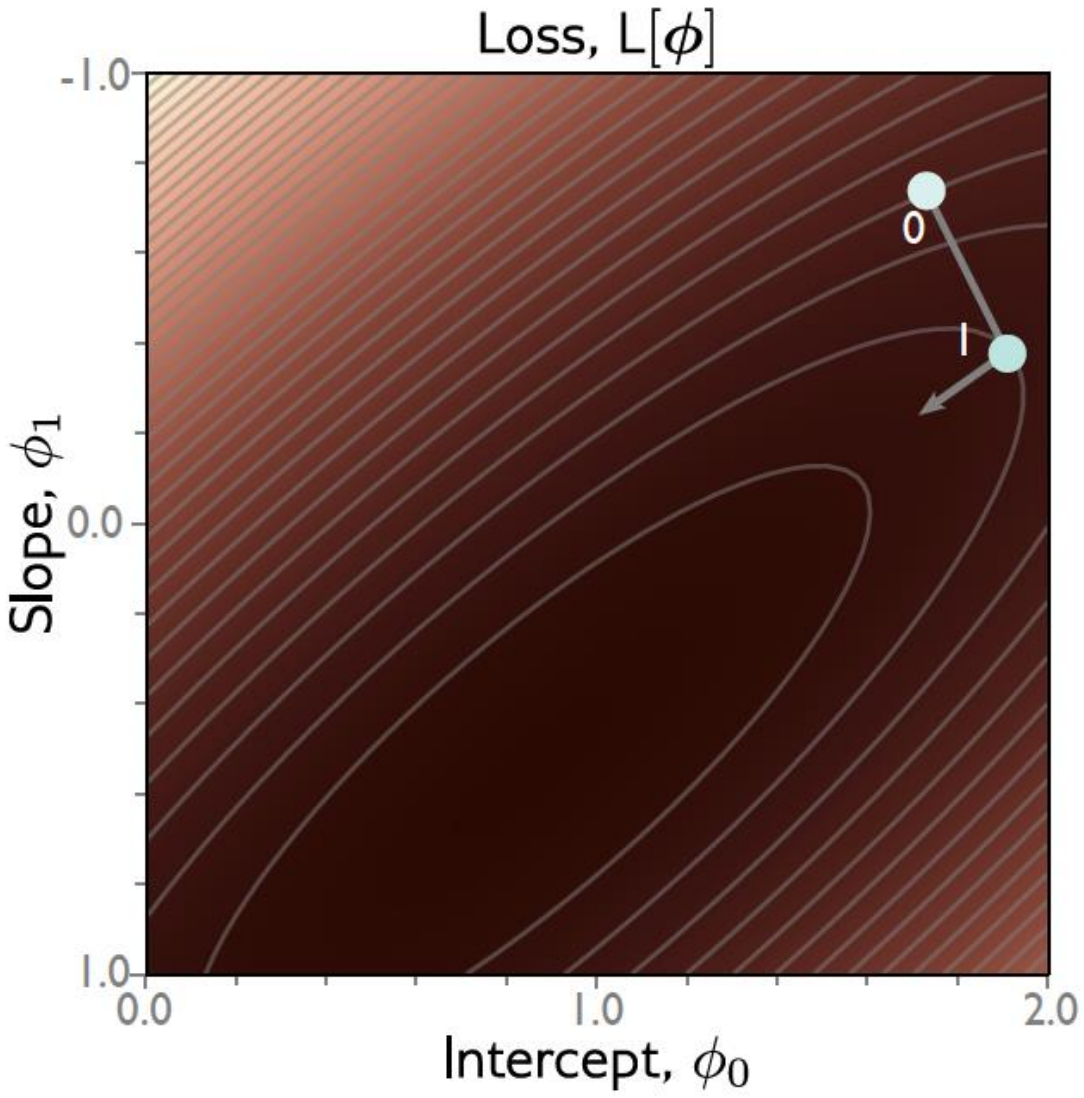


Descenso gradual





Búsqueda en línea



Paso 1: Calcular las derivadas (pendientes de la función) con Respecto a los parámetros

$$\frac{\partial L}{\partial \boldsymbol{\phi}} = \frac{\partial}{\partial \boldsymbol{\phi}} \sum_{i=1}^{I} \ell_i = \sum_{i=1}^{I} \frac{\partial \ell_i}{\partial \boldsymbol{\phi}}$$

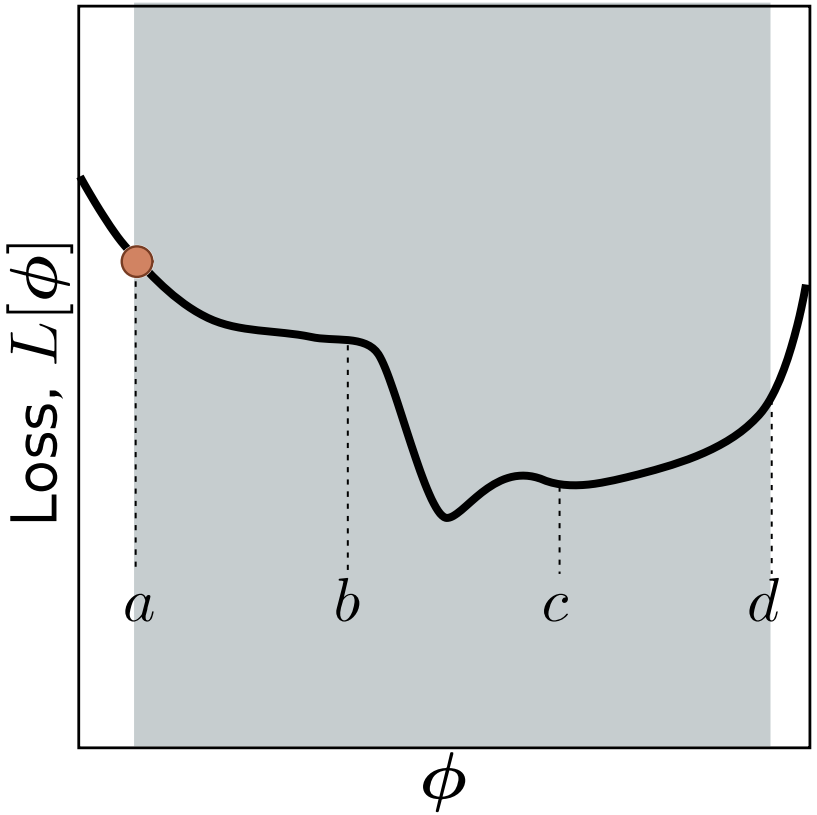
$$\frac{\partial \ell_i}{\partial \boldsymbol{\phi}} = \begin{bmatrix} \frac{\partial \ell_i}{\partial \phi_0} \\ \frac{\partial \ell_i}{\partial \phi_1} \end{bmatrix} = \begin{bmatrix} 2(\phi_0 + \phi_1 x_i - y_i) \\ 2x_i(\phi_0 + \phi_1 x_i - y_i) \end{bmatrix}$$

Paso 2: Actualizar los parámetros según la regla

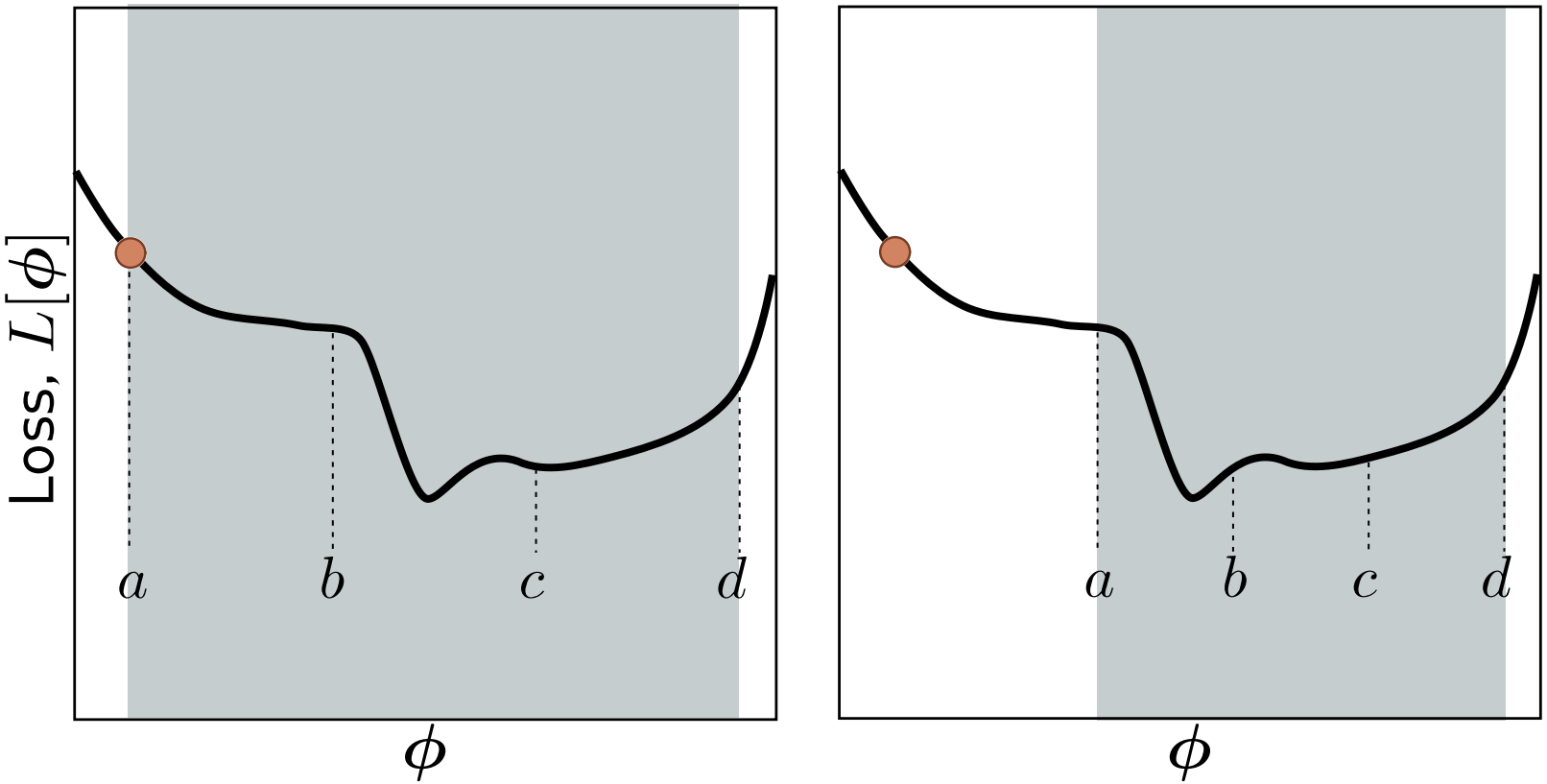
$$\boldsymbol{\phi} \longleftarrow \boldsymbol{\phi} - \alpha \frac{\partial L}{\partial \boldsymbol{\phi}}$$

= tamaño del paso

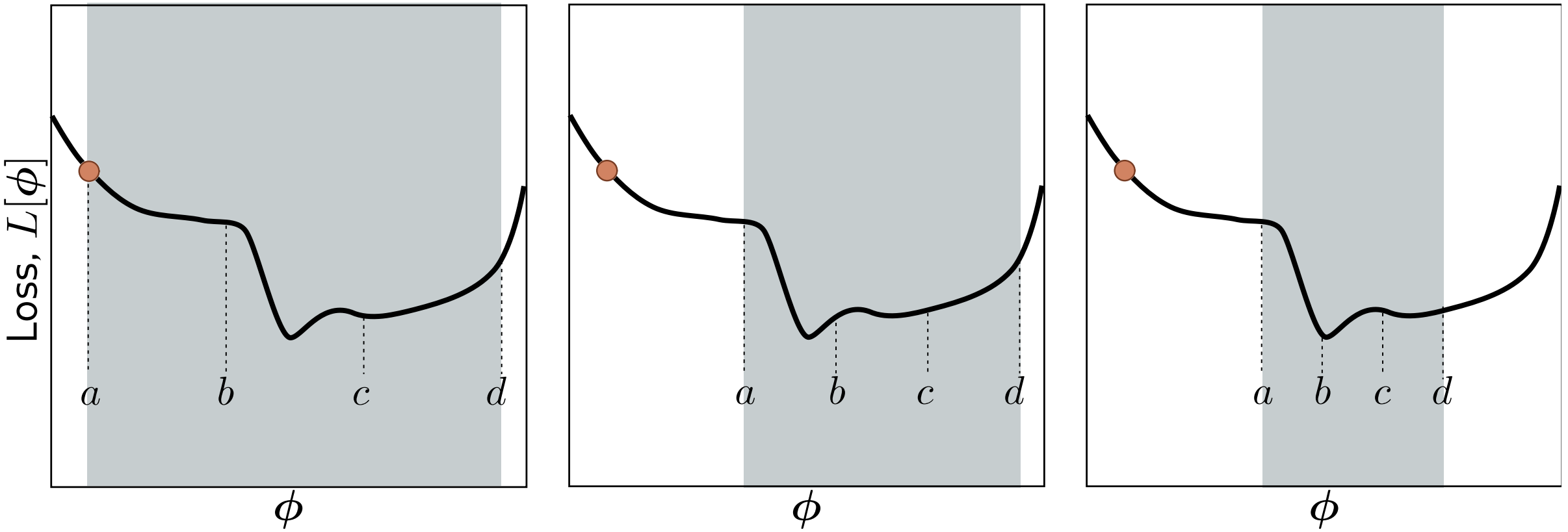
Búsqueda de líneas (entre corchetes)



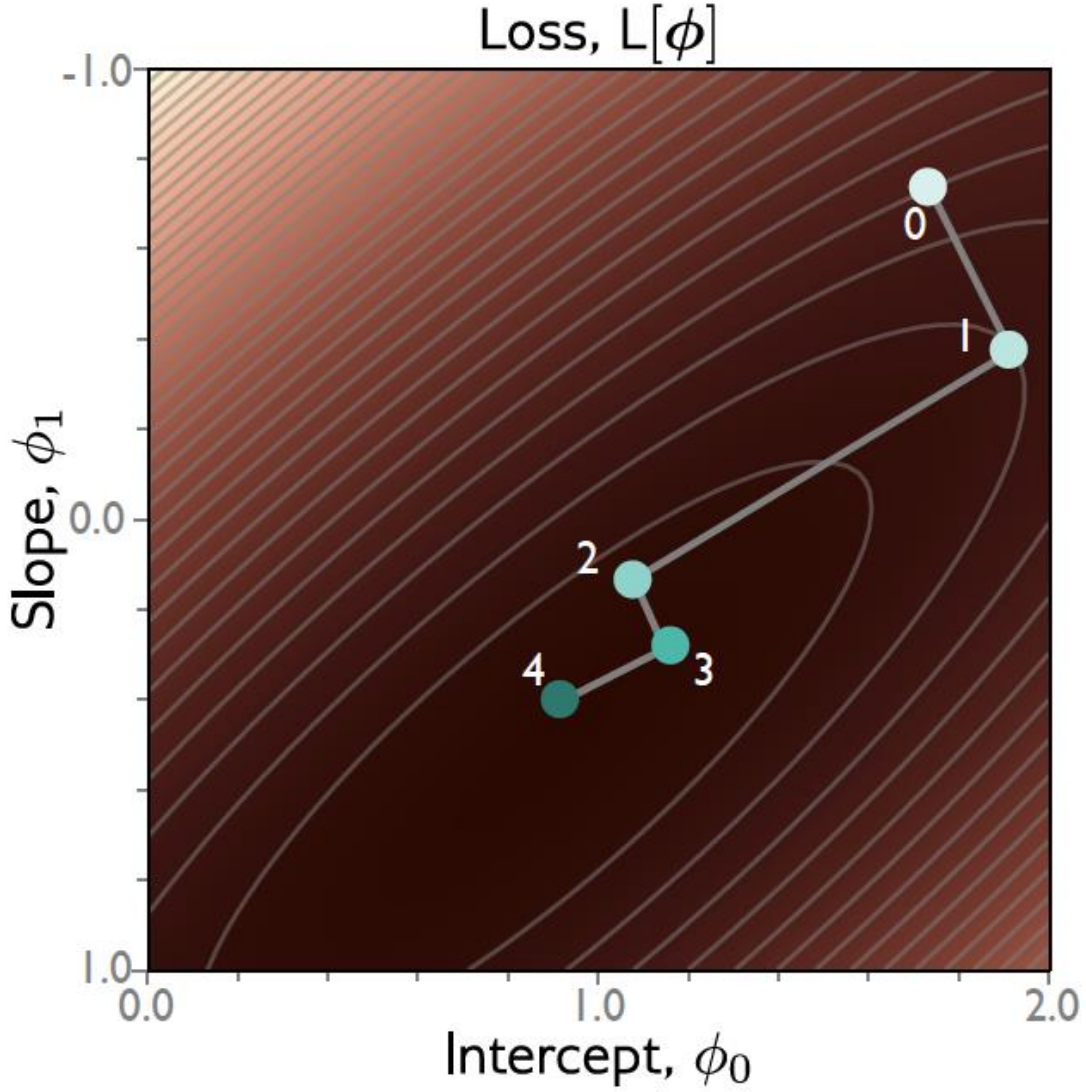
Búsqueda de líneas (entre corchetes)

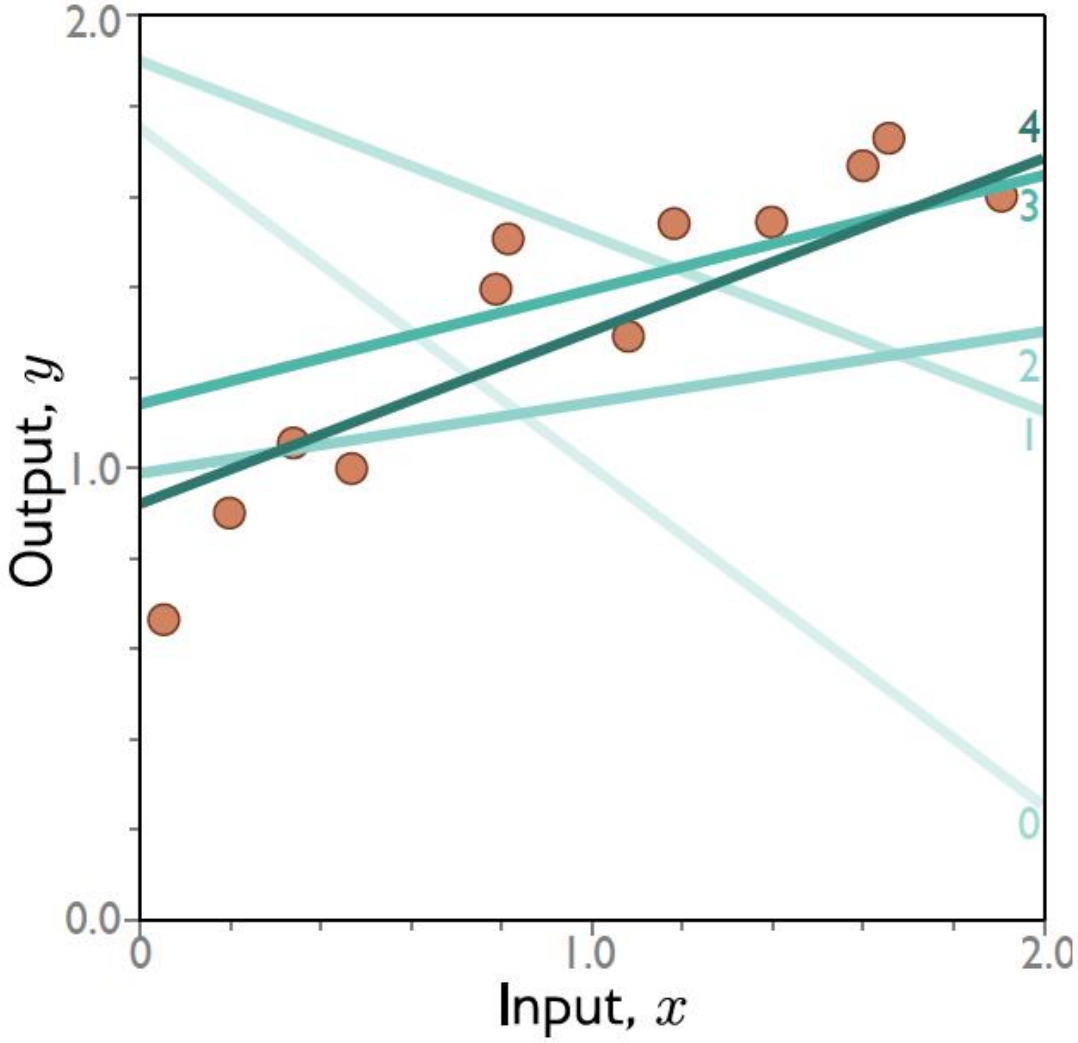


Búsqueda de líneas (entre corchetes)

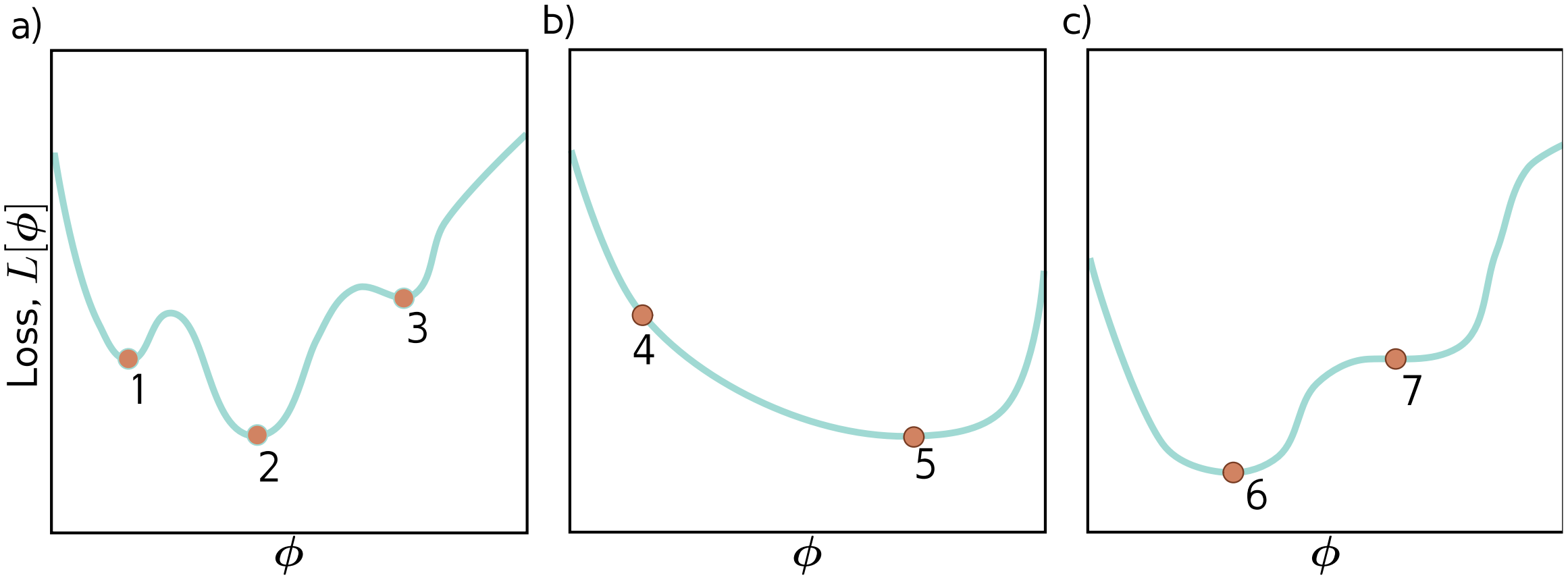


Descenso gradual





Problemas convexos

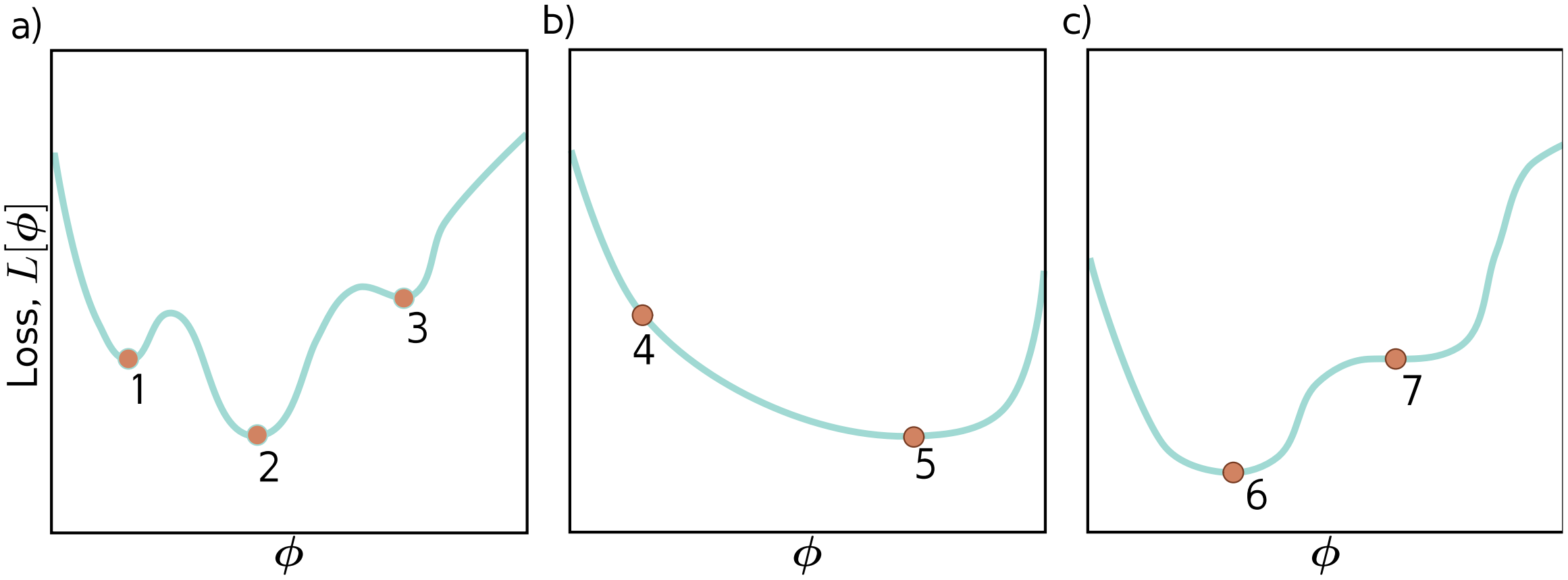


No convexo

Convexo

No convexo

Problemas convexos



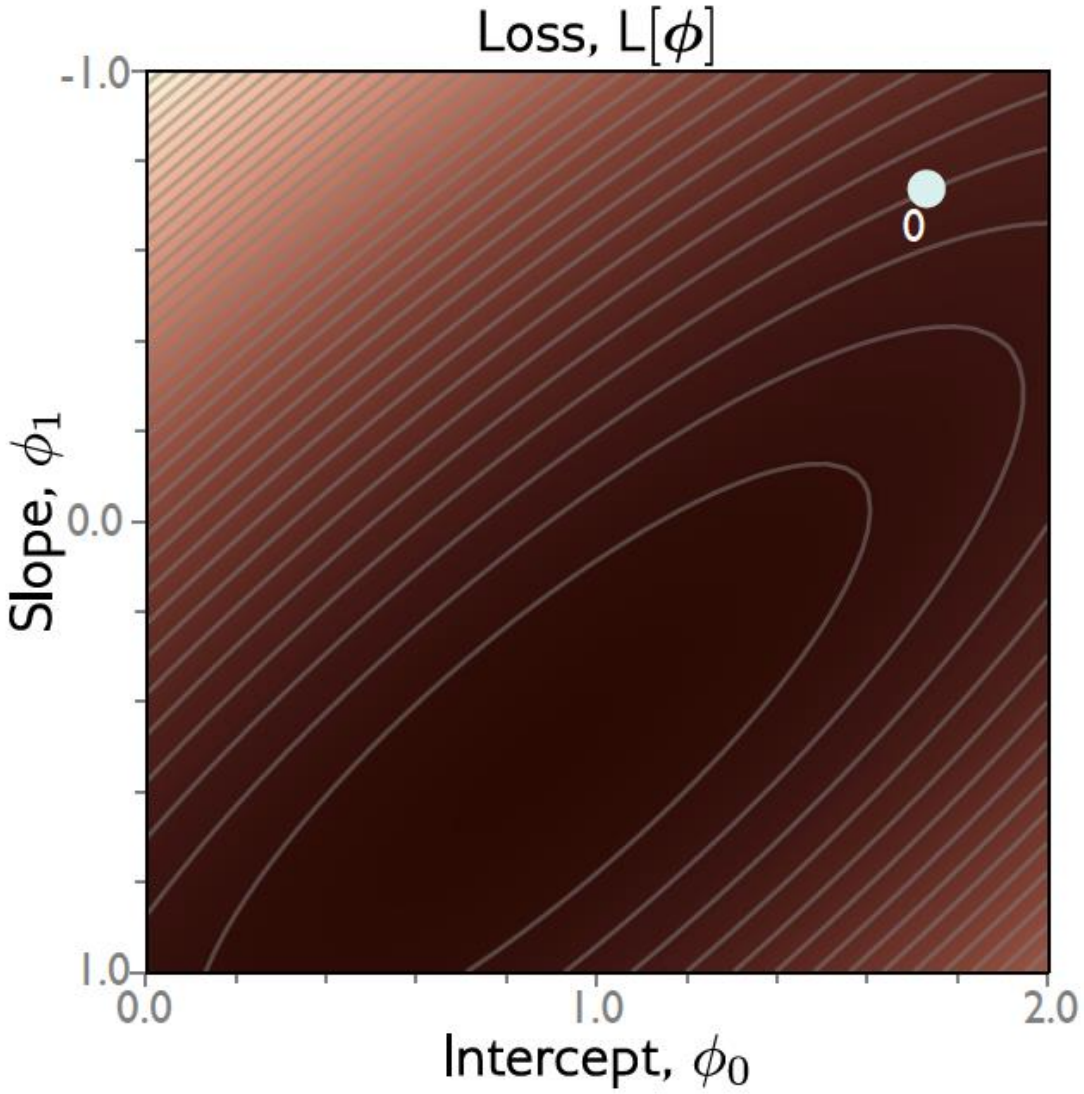
No convexo

Convexo

No convexo

La prueba de convexidad es que 2nd la derivada es positiva en todas partes

Convexidad en dimensiones superiores



La prueba de convexidad es que el determinante del hessiano (2da matriz derivada) es positivo en todas partes.

$$\mathbf{H}[\boldsymbol{\phi}] = \begin{bmatrix} \frac{\partial^2 L}{\partial \phi_0^2} & \frac{\partial^2 L}{\partial \phi_0 \partial \phi_1} \\ \frac{\partial^2 L}{\partial \phi_1 \partial \phi_0} & \frac{\partial^2 L}{\partial \phi_1^2} \end{bmatrix}$$

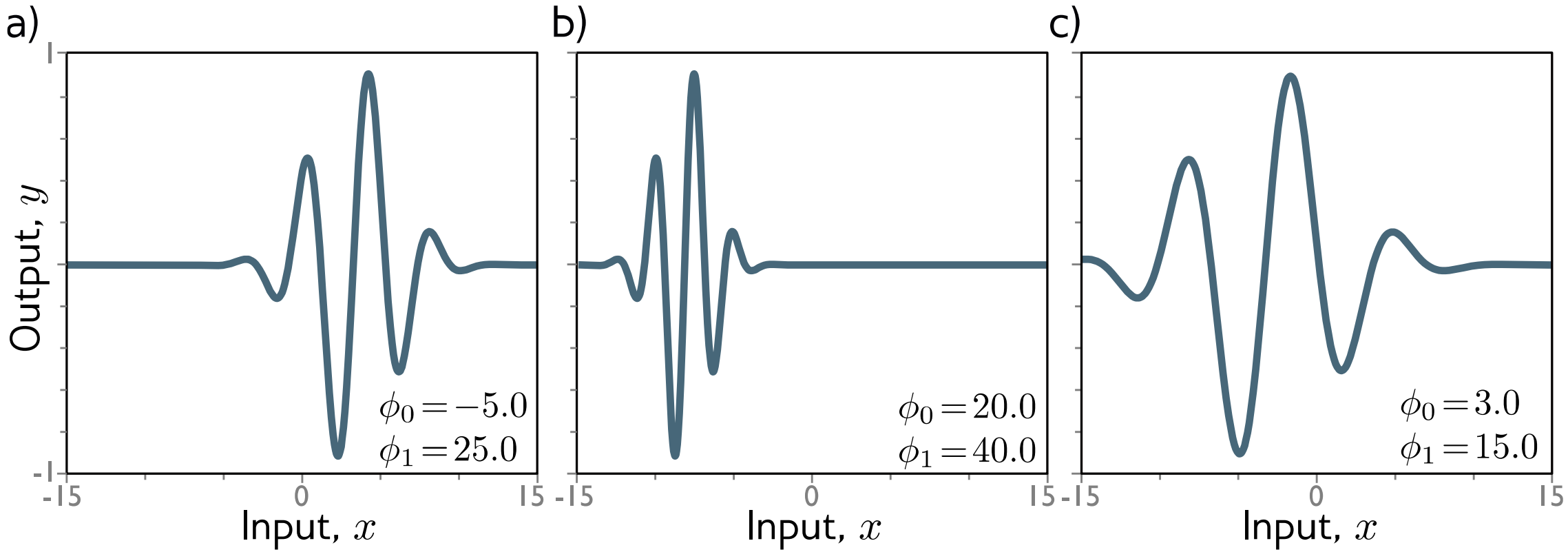
$$\mathbf{H}[\boldsymbol{\phi}] = \frac{\partial^2 L}{\partial \phi_0^2}\frac{\partial^2 L}{\partial \phi_1^2} - \frac{\partial^2 L}{\partial \phi_0 \partial \phi_1}\frac{\partial^2 L}{\partial \phi_1 \partial \phi_0}$$

Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam

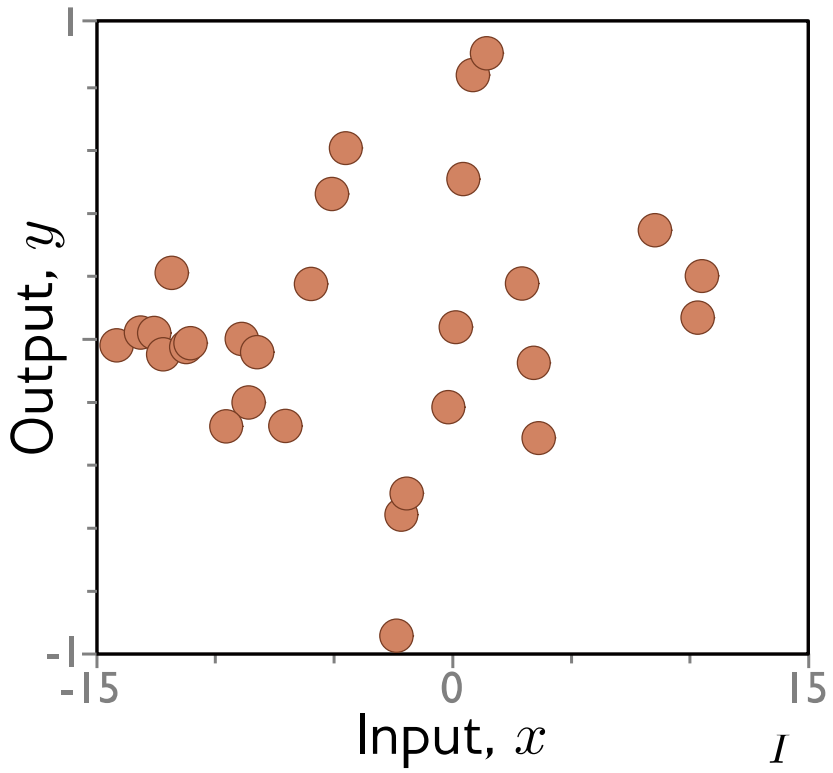
Modelo Gabor

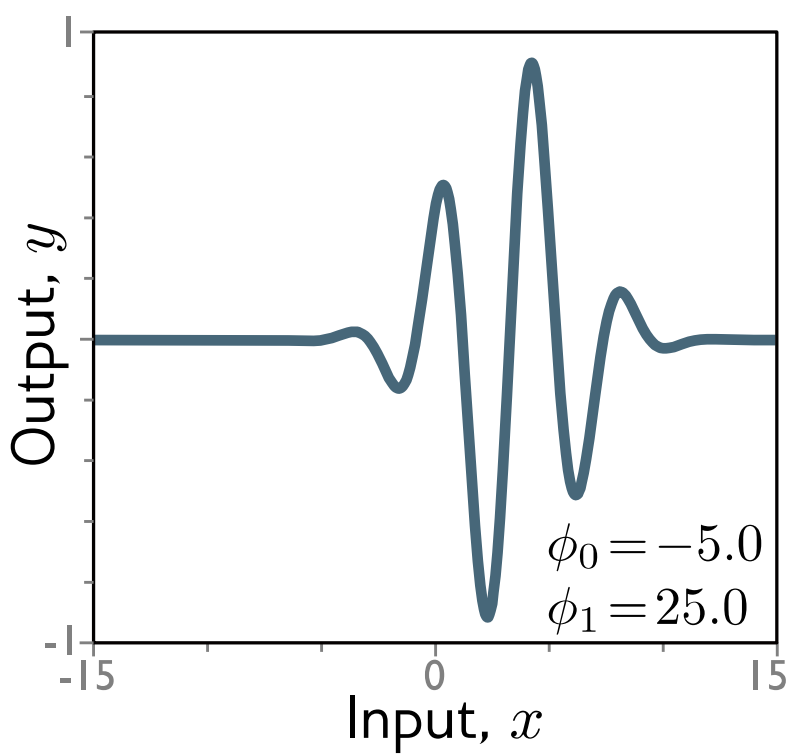
$$\mathrm{f}[x, \boldsymbol{\phi}] = \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{8.0}\right)$$



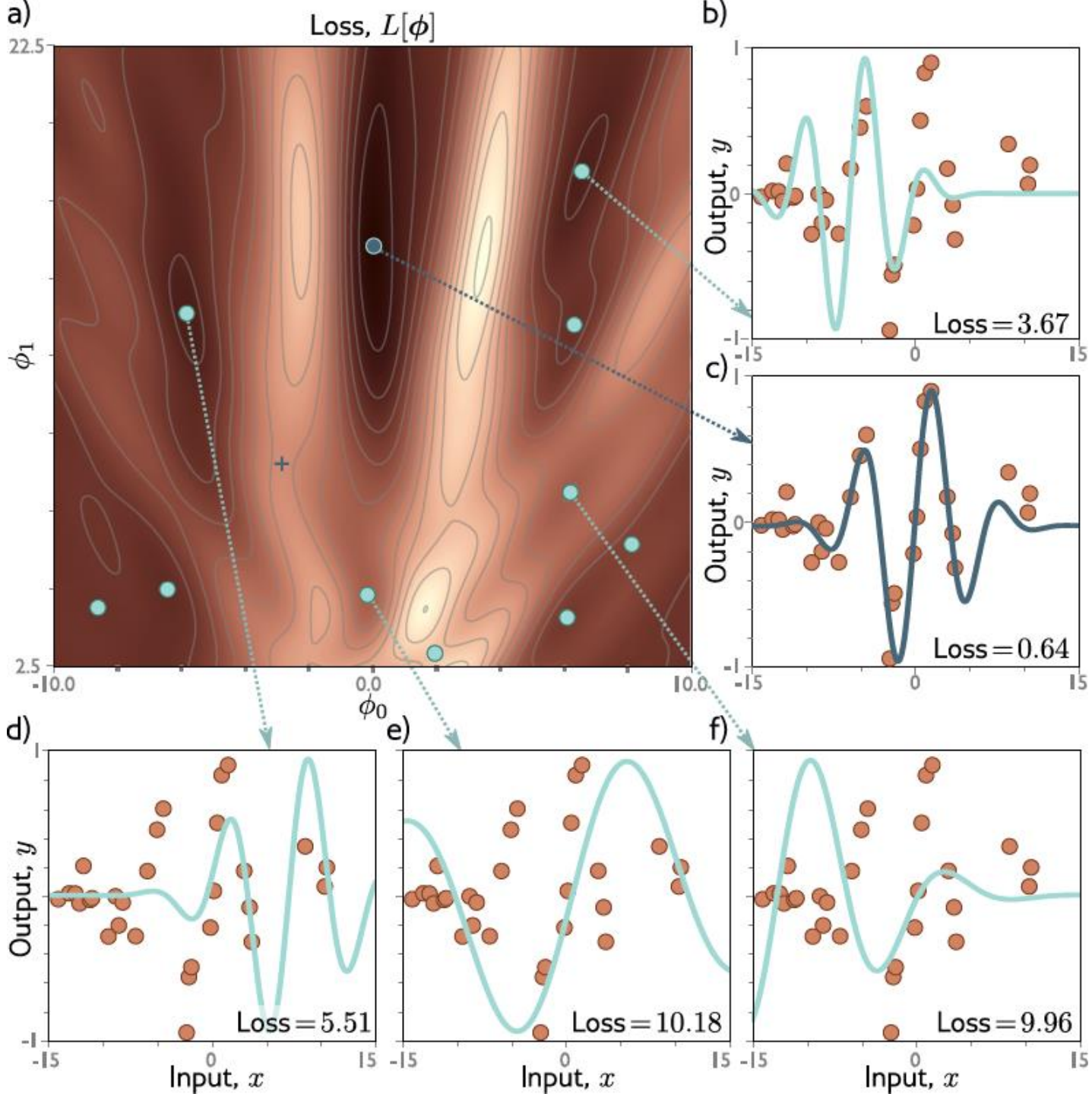
Modelo Gabor

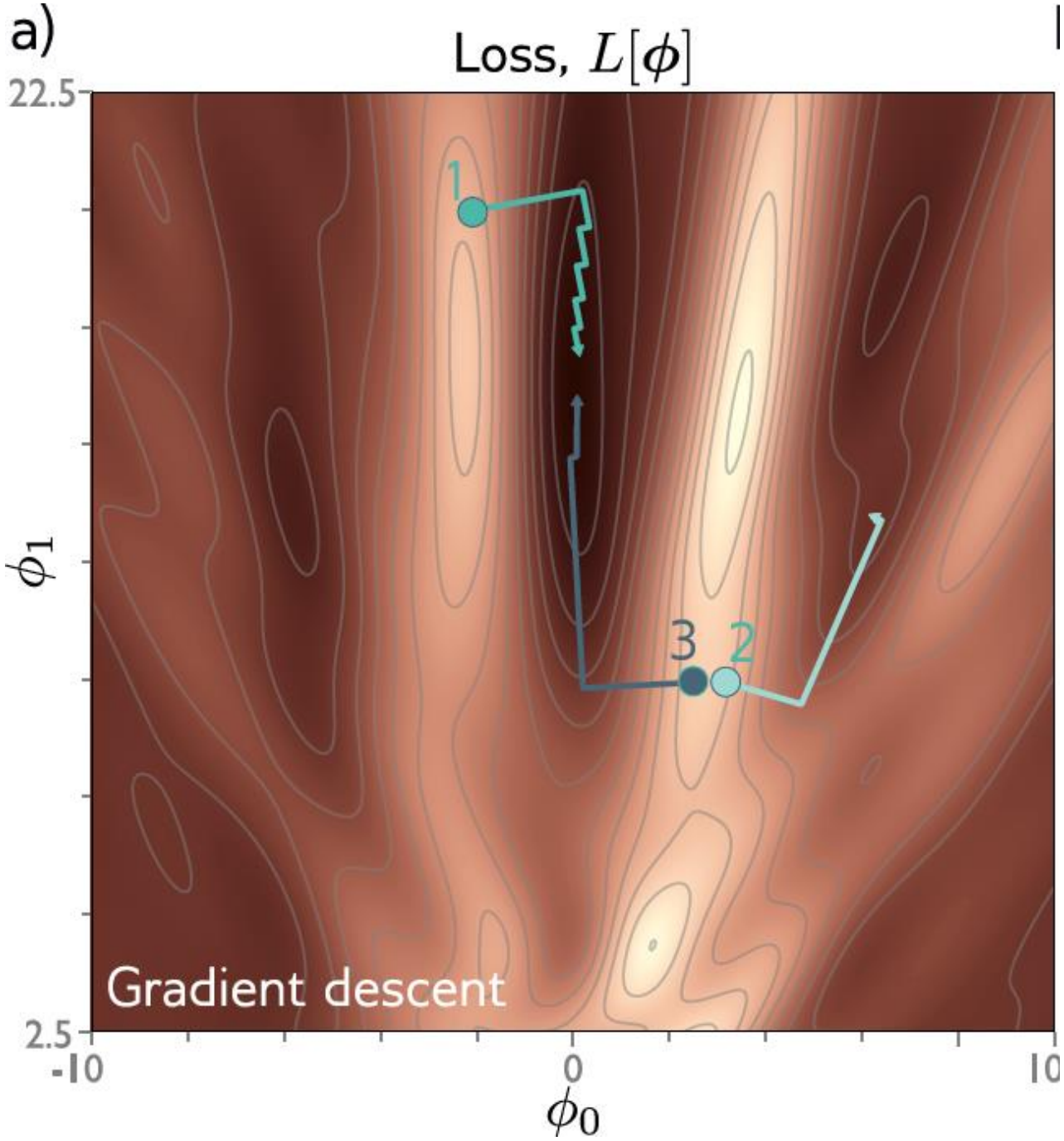
$$\mathrm{f}[x, \boldsymbol{\phi}] = \sin[\phi_0 + 0.06 \cdot \phi_1 x] \cdot \exp\left(-\frac{(\phi_0 + 0.06 \cdot \phi_1 x)^2}{8.0}\right)$$





$$L[\boldsymbol{\phi}] = \sum_{i=1}^{I} \left(\mathrm{f}[x_i, \boldsymbol{\phi}] - y_i\right)^2$$

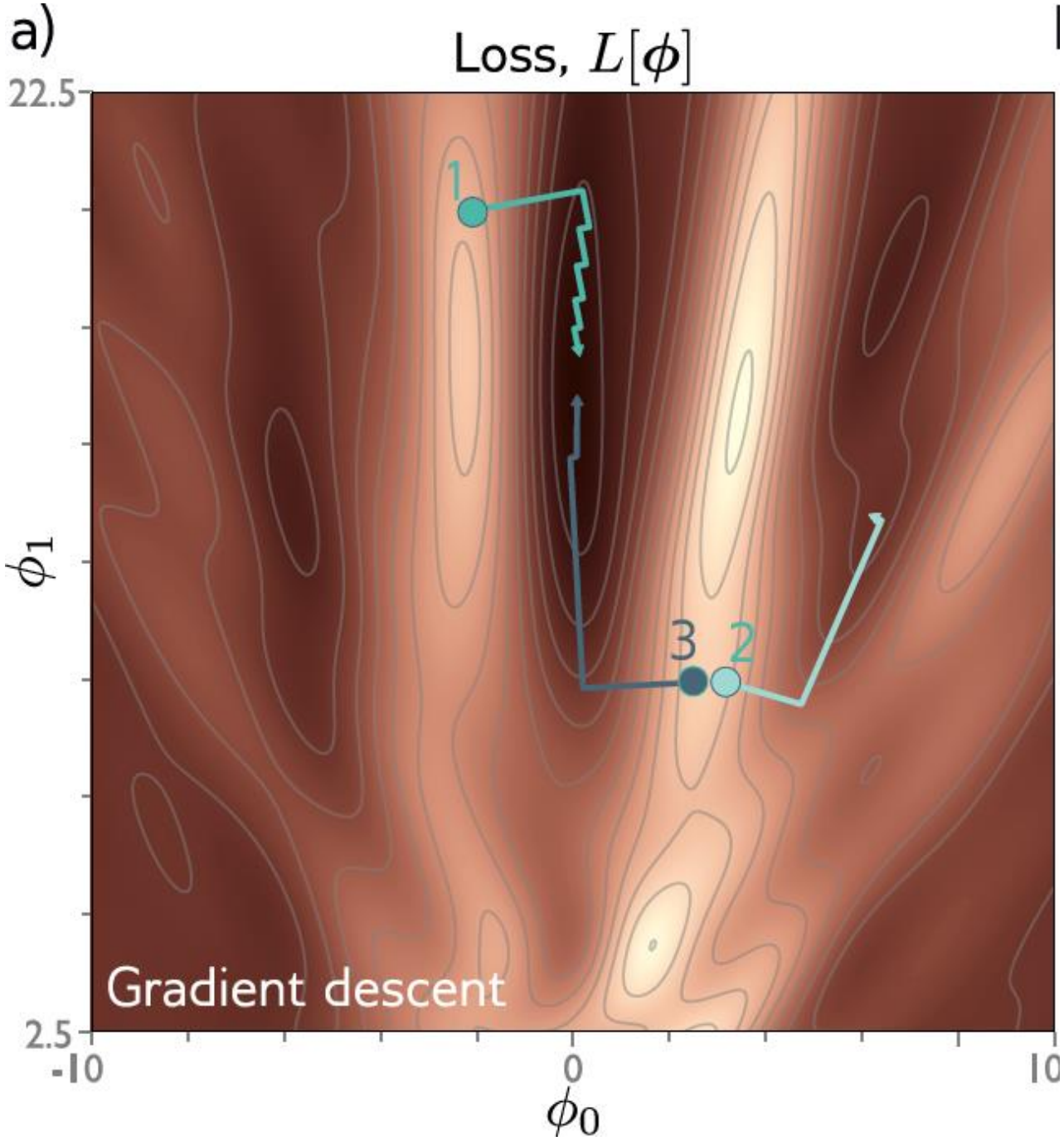

a)
Loss, $L[\phi]$
22.5
2.5
ϕ_1
-10.0
0.0
10.0
ϕ_0
b)
Output, y
Loss = 3.67
c)
Loss = 0.64
d)
Loss = 5.51
e)
Loss = 10.18
f)
Loss = 9.96
Input, x
-15
0
15
1
-1



- El descenso gradiente llega al mínimo global si empezamos en el "valle" correcto
- En caso contrario, descenso a un mínimo local
- O quedarse atascado cerca de un punto de silla *(saddle point)* de montar

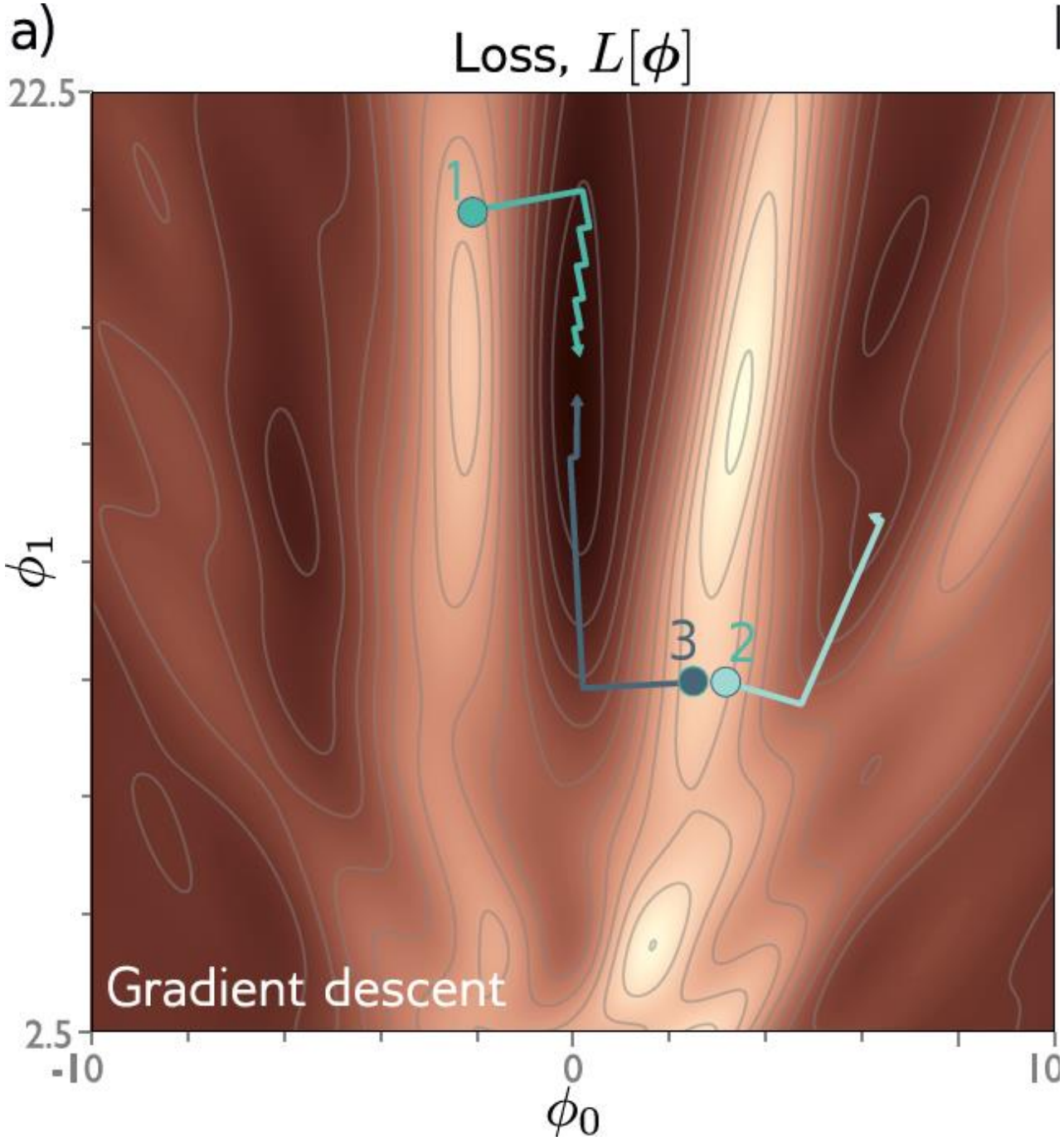
Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam



IDEA: añadir ruido

- Descenso de gradiente estocástico
- Calcular el gradiente basándose sólo en un subconjunto de puntos: un minilote *(minibatch)*
- Trabajar a través del muestreo de conjuntos de datos sin reemplazo
- Una pasada por los datos se denomina época *(epoch)*



Descenso de gradiente estocástico

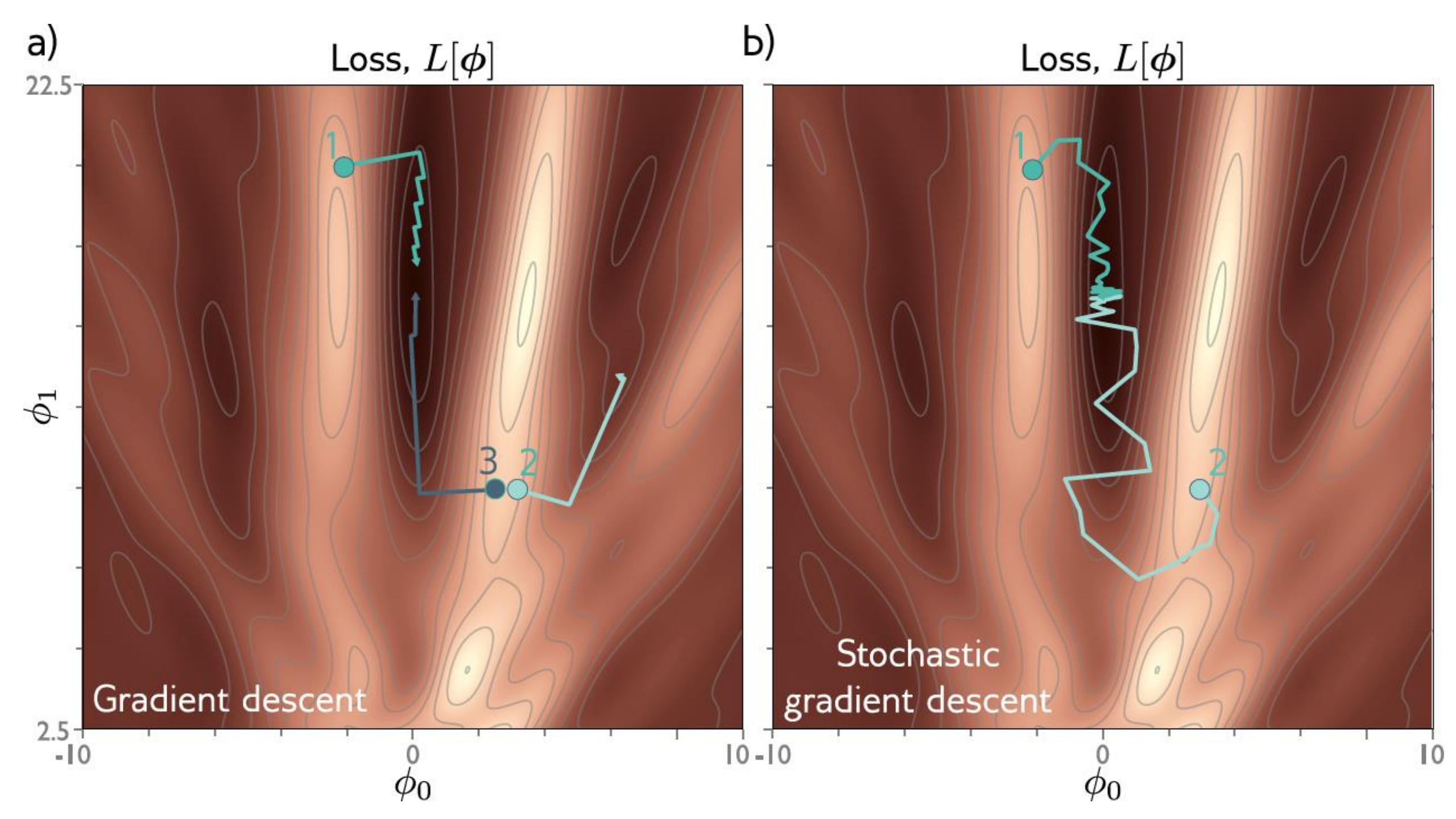
Antes (descenso por lotes completos)

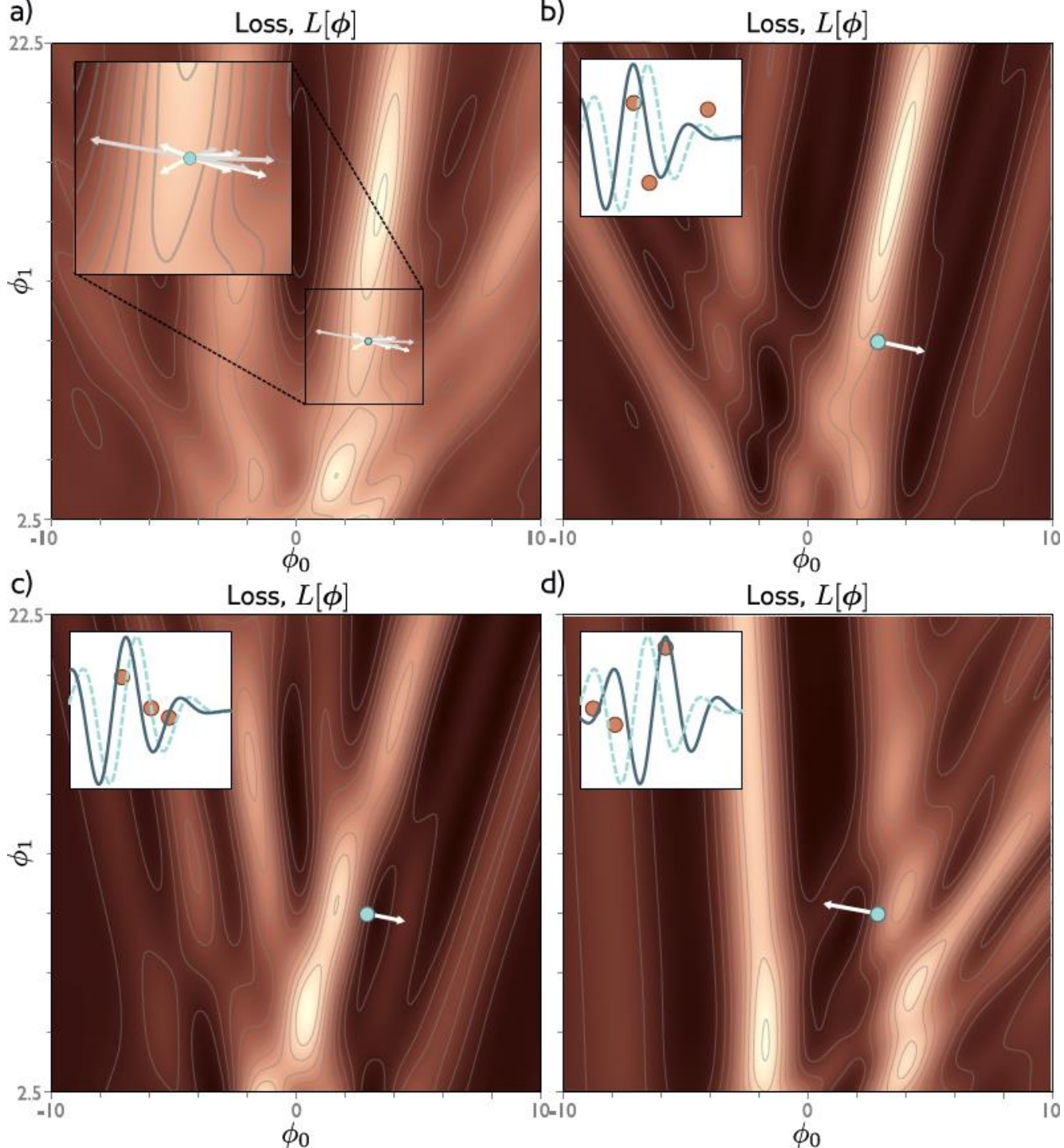
$$\boldsymbol{\phi}_{t+1} \longleftarrow \boldsymbol{\phi}_t - \alpha \sum_{i=1}^{I} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}},$$

Después (SGD)

$$\boldsymbol{\phi}_{t+1} \longleftarrow \boldsymbol{\phi}_t - \alpha \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}},$$

Tasa de aprendizaje fija


a)
Loss, $L[\phi]$
22.5
2.5
ϕ_1
1
2
3
Gradient descent
-10
0
10
ϕ_0
b)
Loss, $L[\phi]$
1
2
Stochastic
gradient descent
-10
0
10
ϕ_0


a)
Loss, $L[\phi]$
b)
Loss, $L[\phi]$
c)
Loss, $L[\phi]$
d)
Loss, $L[\phi]$
22.5
2.5
ϕ_1
-10
0
10
ϕ_0

Propiedades de SGD

- Puede escapar de mínimos locales
- Añade ruido, pero sigue siendo una actualización razonable, ya que se basa en parte de los datos.
- Utiliza todos los datos por igual
- Menos costoso desde el punto de vista computacional
- Parece encontrar mejores soluciones

- No converge en el sentido tradicional
- Programación del threshold (α) de aprendizaje: disminución del threshold de aprendizaje a lo largo del tiempo

Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam

Momentum

$i \in \{1, \ldots, N\}$

$|Batch| = 100$

$N = 1000$

$\frac{N}{100} = n.$

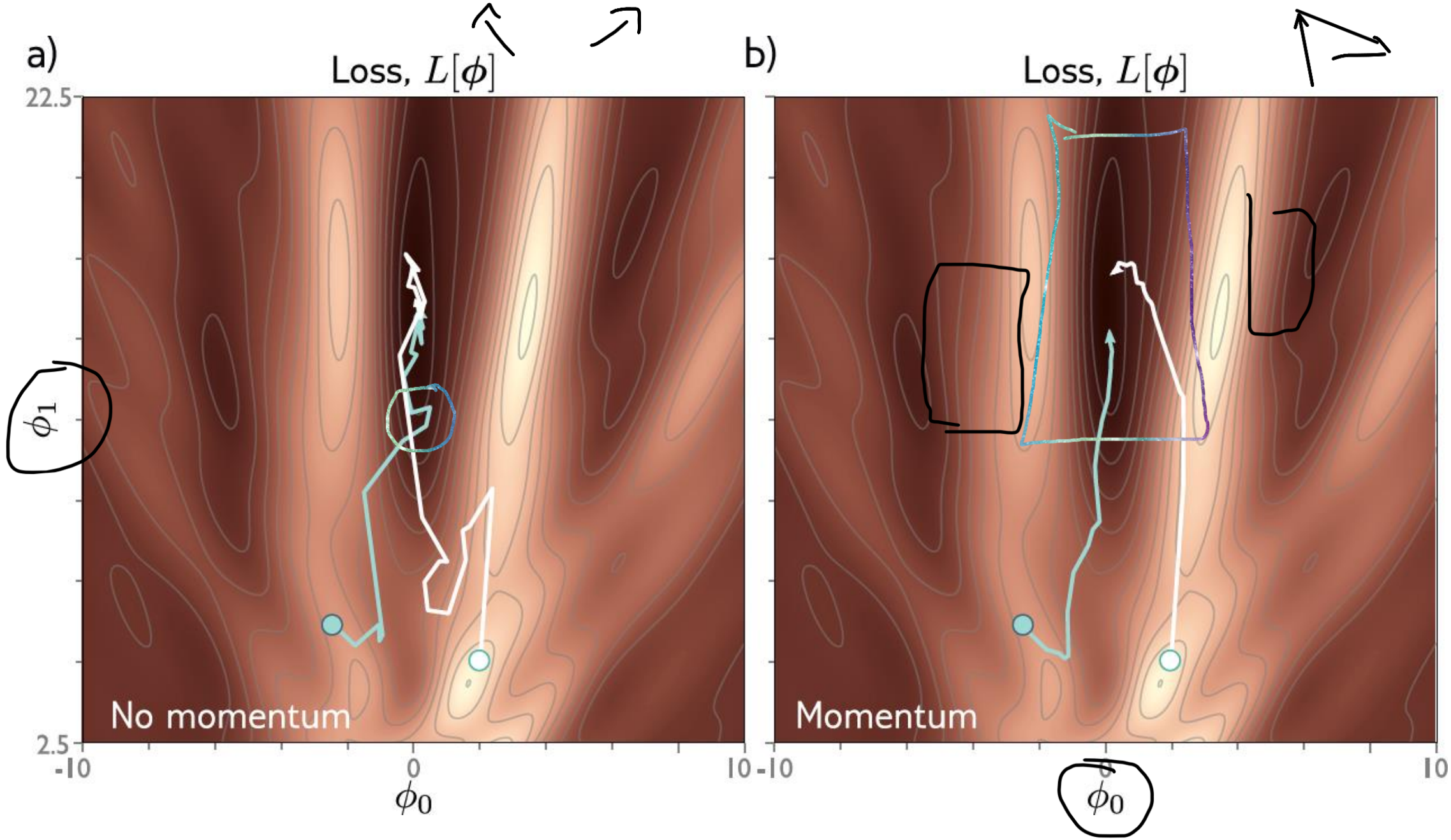
- Suma ponderada de este gradiente y el gradiente anterior

$$\mathbf{m}_{t+1} \leftarrow \beta \cdot \mathbf{m}_t + (1 - \beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_{t+1}$$

$\beta \cdot x + (1 - \beta)\, y$

$\beta \in [0, 1]$


a)
Loss, $L[\phi]$
22.5
2.5
ϕ_1
No momentum
-10
0
10
ϕ_0
b)
Loss, $L[\phi]$
Momentum
-10
0
10
ϕ_0

Nesterov acelera el momentum

- El momentum es una especie de predicción de hacia dónde vamos.

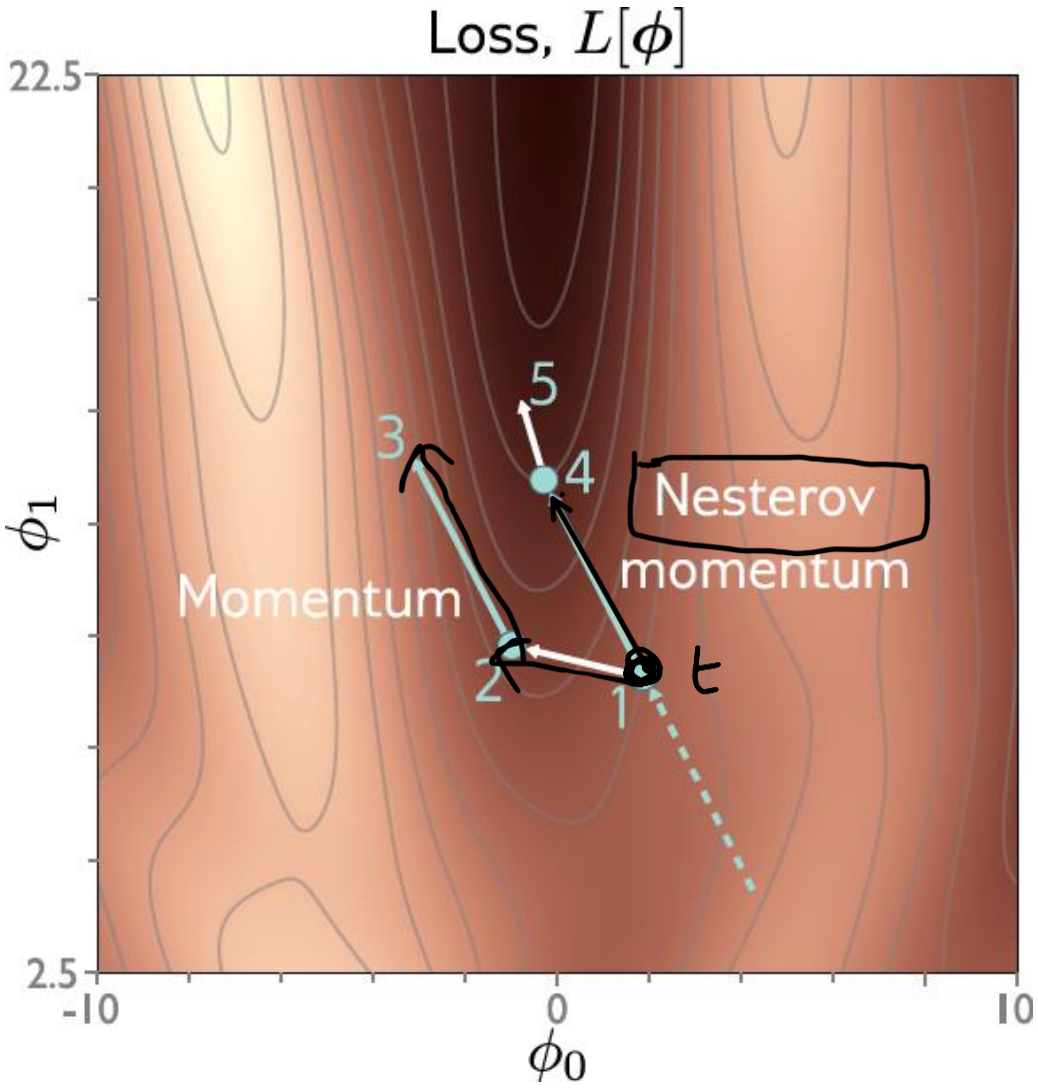
$$\mathbf{m}_{t+1} \leftarrow \beta \cdot \mathbf{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_{t+1}$$

- Muévase en la dirección prevista, ENTONCES, mida el gradiente

$$\mathbf{m}_{t+1} \leftarrow \beta \cdot \mathbf{m}_t + (1-\beta) \sum_{i \in \mathcal{B}_t} \frac{\partial \ell_i[\boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_t]}{\partial \boldsymbol{\phi}}$$

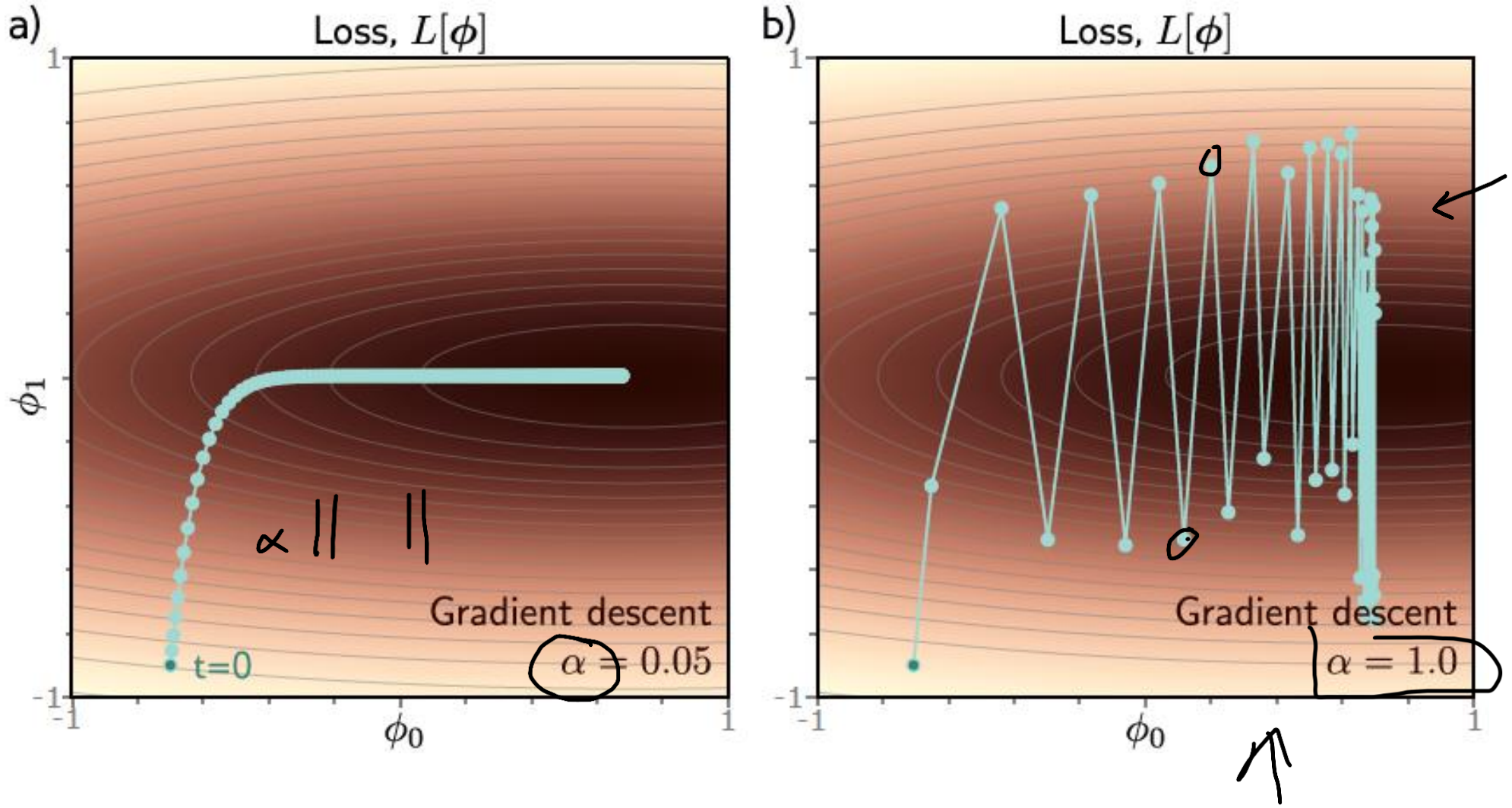
$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \mathbf{m}_{t+1}$$



Modelos de ajuste

- Matemáticas
- Algoritmo de descenso gradiente
- Ejemplo de regresión lineal
- Ejemplo de modelo de Gabor
- Descenso de gradiente estocástico
- Momentum
- Adam

Estimación adaptativa del momento (Adam)



Gradientes normalizados

- Medir la media y el gradiente cuadrático puntual

$$\mathbf{m}_{t+1} \leftarrow \frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

$$\mathbf{v}_{t+1} \leftarrow \frac{\partial L[\boldsymbol{\phi}_t]^2}{\partial \boldsymbol{\phi}}$$

$$\propto \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$$

- Normalizar:

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v}_{t+1}} + \epsilon}$$

Gradientes normalizados

- Medir la media y el gradiente cuadrático puntual

$$\mathbf{m}_{t+1} \leftarrow \frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

$$\mathbf{v}_{t+1} \leftarrow \frac{\partial L[\boldsymbol{\phi}_t]^2}{\partial \boldsymbol{\phi}}$$

$$\mathbf{m}_{t+1} = \begin{bmatrix} 3.0 \\ 2.0 \\ 5.0 \end{bmatrix}$$

$$\mathbf{v}_{t+1} = \begin{bmatrix} 9.0 \\ 4.0 \\ 25.0 \end{bmatrix}$$

- Normalizar:

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v}_{t+1}} + \epsilon}$$

$$\frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v}_{t+1}} + \epsilon} = \begin{bmatrix} 1.0 \\ -1.0 \\ 1.0 \end{bmatrix}$$

Gradientes normalizados

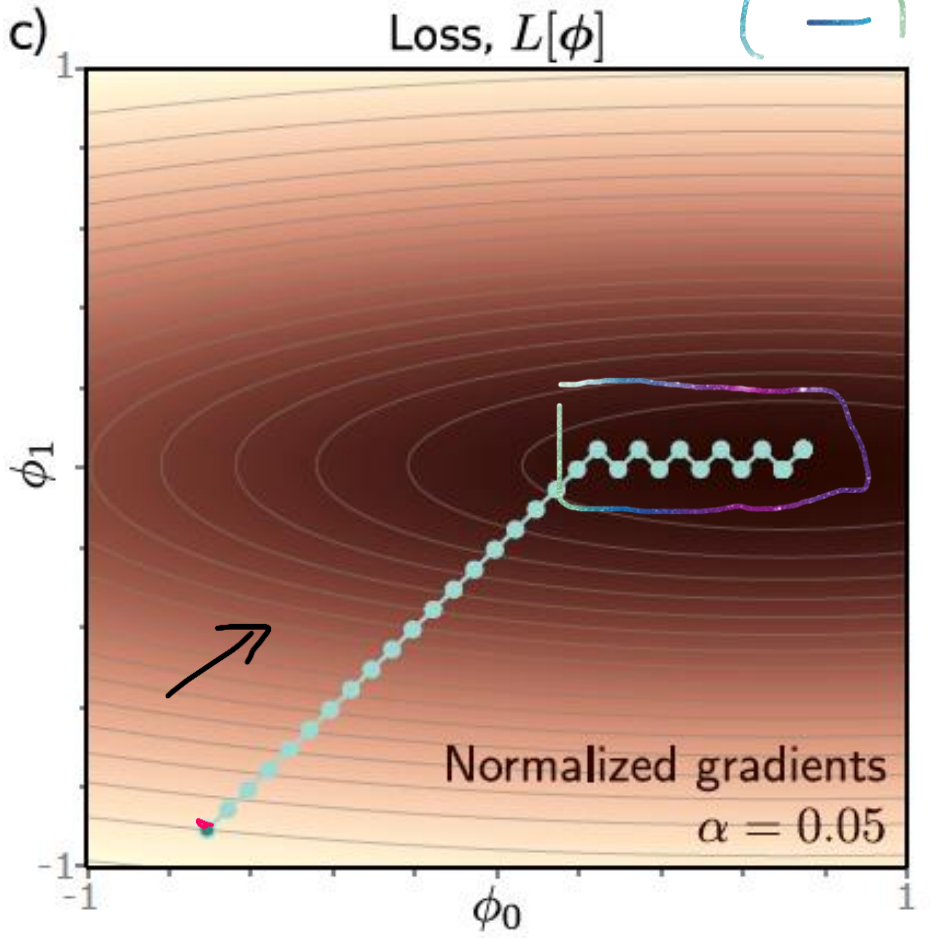
- Medir la media y el gradiente cuadrático puntual

$$\mathbf{m}_{t+1} \leftarrow \frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

$$\mathbf{v}_{t+1} \leftarrow \frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}^2$$

- Normalizar:

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \frac{\mathbf{m}_{t+1}}{\sqrt{\mathbf{v}_{t+1}} + \epsilon}$$



Estimación adaptativa del momento (Adam)

- Calcular la media y los gradientes puntuales al cuadrado con el impulso

$$\mathbf{m}_{t+1} \leftarrow \beta \cdot \mathbf{m}_t + (1-\beta)\frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}$$

$$\mathbf{v}_{t+1} \leftarrow \gamma \cdot \mathbf{v}_t + (1-\gamma)\left(\frac{\partial L[\boldsymbol{\phi}_t]}{\partial \boldsymbol{\phi}}\right)^2$$

- Moderado cerca del inicio de la secuencia

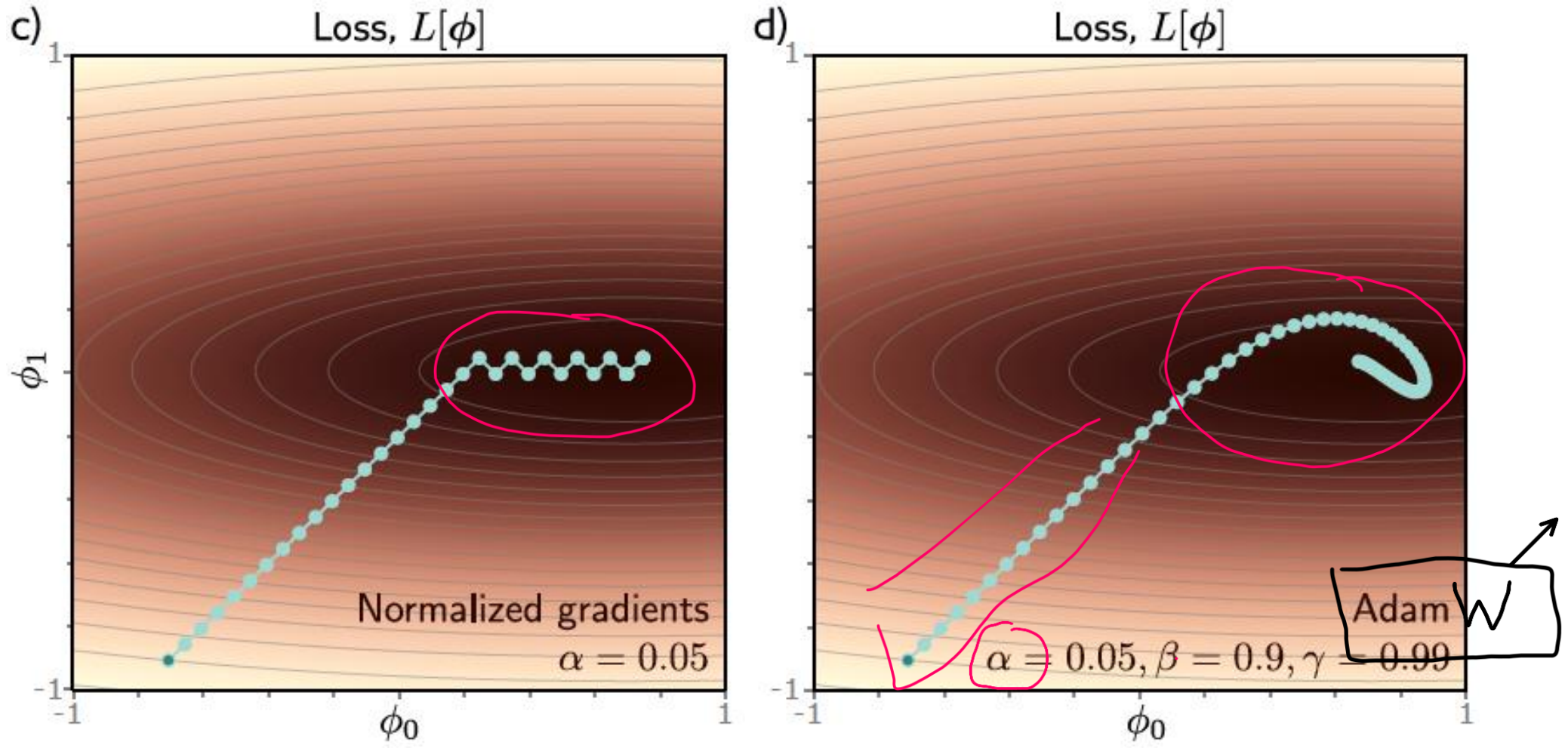
$$\tilde{\mathbf{m}}_{t+1} \leftarrow \frac{\mathbf{m}_{t+1}}{1-\beta^{t+1}}$$

$$\tilde{\mathbf{v}}_{t+1} \leftarrow \frac{\mathbf{v}_{t+1}}{1-\gamma^{t+1}}$$

- Actualizar los parámetros

$$\boldsymbol{\phi}_{t+1} \leftarrow \boldsymbol{\phi}_t - \alpha \cdot \frac{\tilde{\mathbf{m}}_{t+1}}{\sqrt{\tilde{\mathbf{v}}_{t+1}} + \epsilon}$$

Estimación adaptativa del momento (Adam)



Hiperparámetros

- Elección del algoritmo de aprendizaje
- Tasa de aprendizaje α
- Momentum β, β, γ