SIGMOIDAL

AULA 005

DEEP LEARNING

Praticamente, tudo que se chama deep learning é movido por um algoritmo muito importante: Stochastic Gradient Descent (SGD).

- Goodfellow et al (2016)

MODELOS PARAMÉTRICOS

Tem a capacidade de aprender padrões a partir de dados de input durante uma etapa conhecida como treinamento.

MODELOS PARAMÉTRICOS

Esse modelo é representado não por todos os dados que o alimentam, mas por um número limitado de parâmetros.

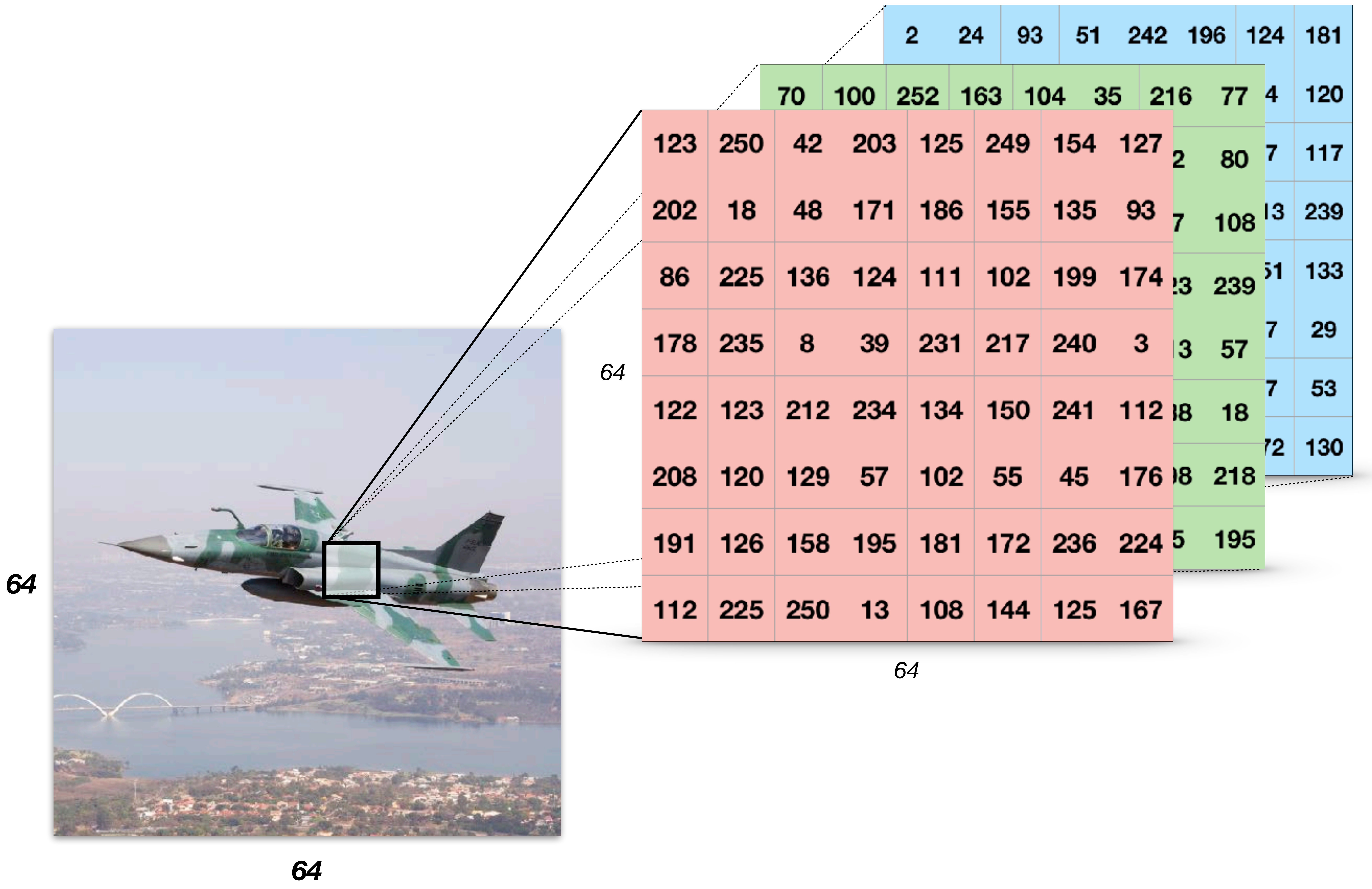
MODELOS PARAMÉTRICOS

Independente do tamanho do meus dados de treino, é possível generalizar, representar os mesmos por meio de parâmetros.

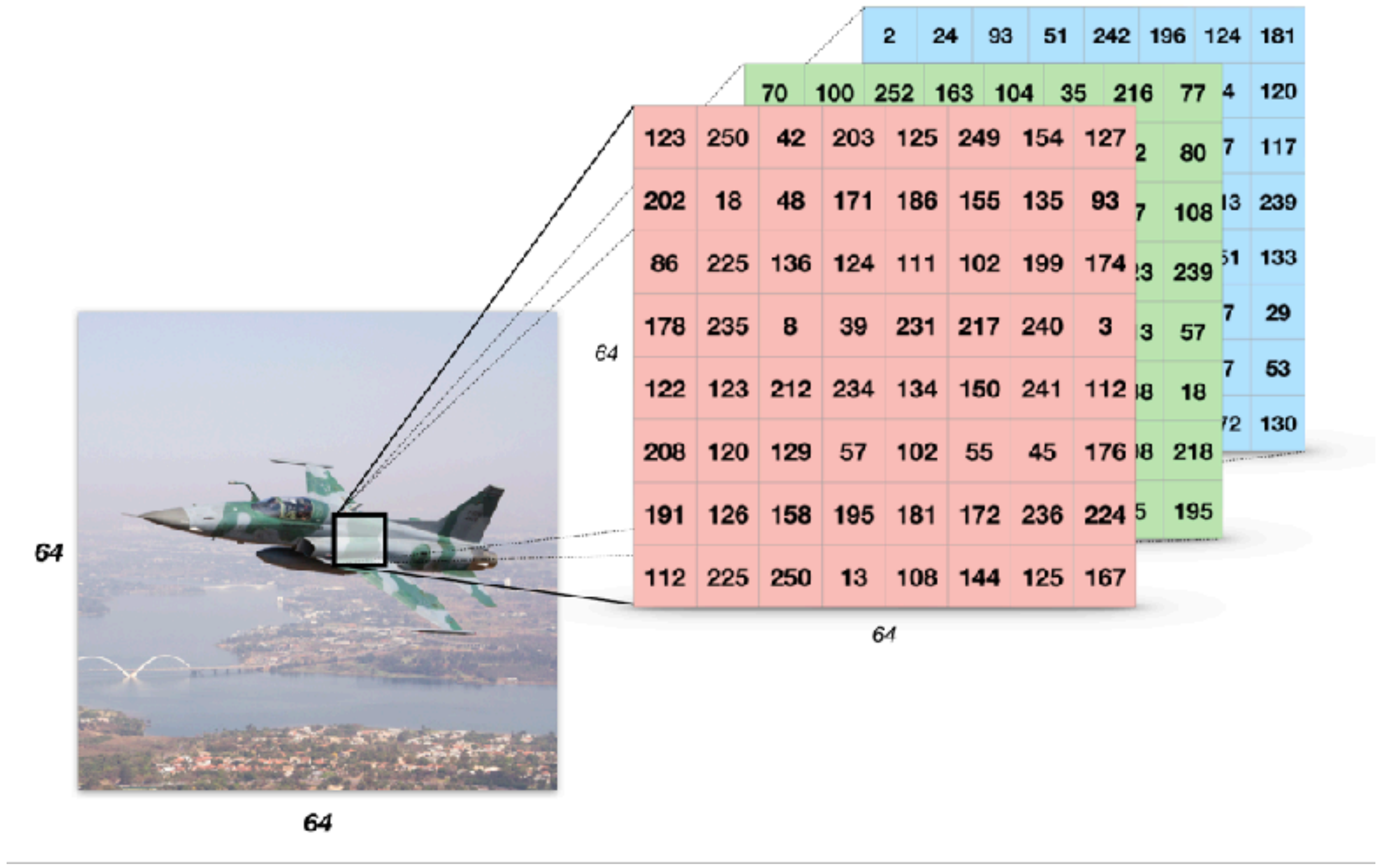
1 (avião)

0 (não-avião)

1 (avião)



2 24 93 51 242 196 124 181
120
117
239
133
29
53
130
70 100 252 163 104 35 216 77
80
108
239
57
18
218
195
123 250 42 203 125 249 154 127
202 18 48 171 186 155 135 93
86 225 136 124 111 102 199 174
178 235 8 39 231 217 240 3
122 123 212 234 134 150 241 112
208 120 129 57 102 55 45 176
191 126 158 195 181 172 236 224
112 225 250 13 108 144 125 167
64
64
64
64



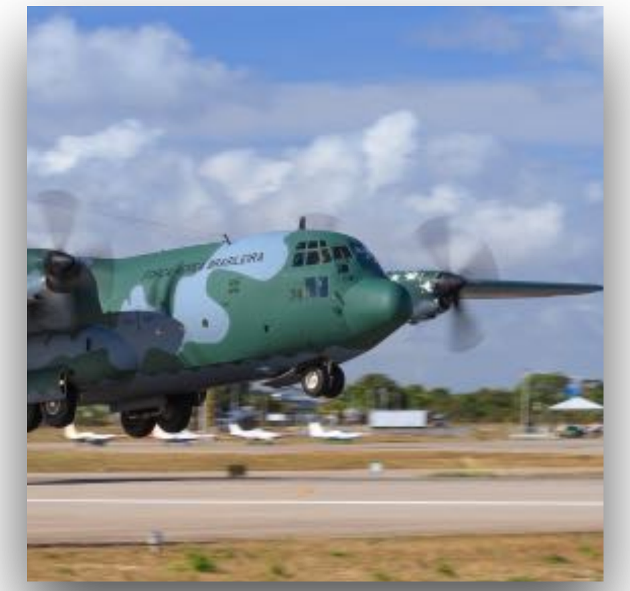
$$x = \begin{bmatrix} 123 \\ 250 \\ \vdots \\ 70 \\ 100 \\ \vdots \\ 130 \end{bmatrix}$$

$$y = 1$$

$$n_x = 64 \times 64 \times 3 = 12288$$
$$m = 1$$

$$x^{(1)} = \begin{bmatrix} 12 \\ 122 \\ \vdots \\ 163 \end{bmatrix}$$

$$x^{(2)} = \begin{bmatrix} 231 \\ 146 \\ \vdots \\ 0 \end{bmatrix}$$



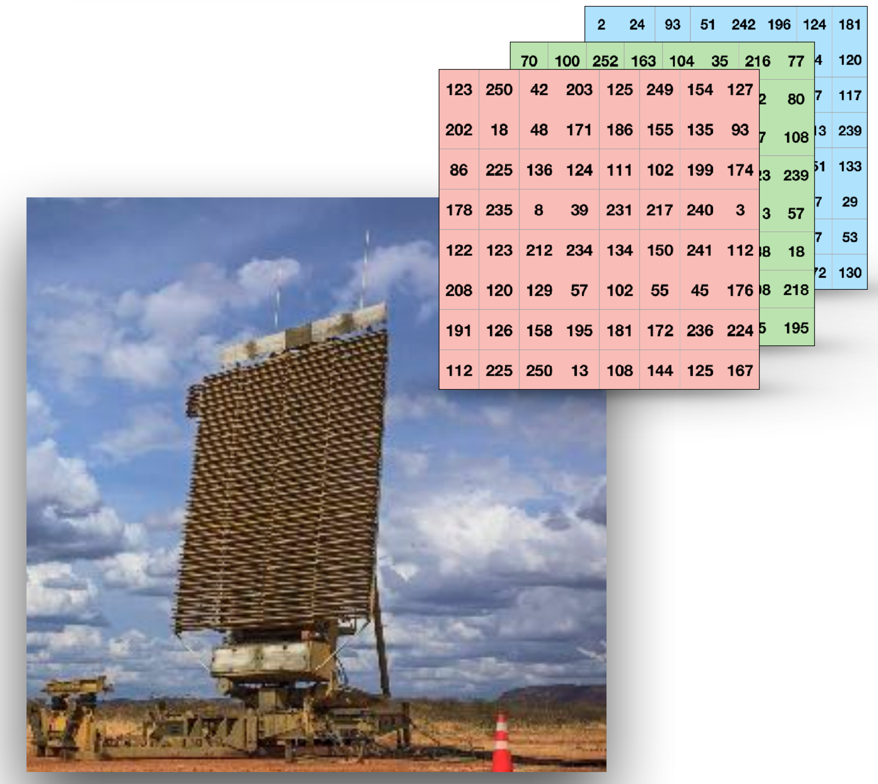
$$x^{(3)} = \begin{bmatrix} 22 \\ 13 \\ \vdots \\ 89 \end{bmatrix}$$

$$X = \begin{bmatrix} 12 & 231 & 22 \\ 122 & 146 & 13 \\ \vdots & \vdots & \vdots \\ 163 & 0 & 89 \end{bmatrix}$$

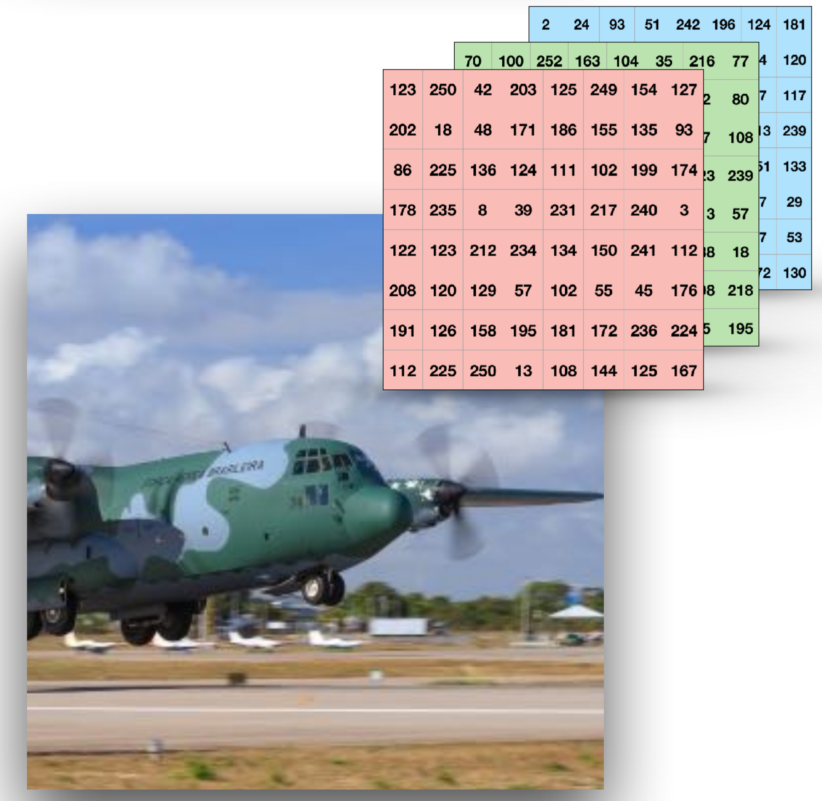
$$Y = [1,0,1]$$

$$n_x = 12288$$
$$m = 3$$

$$x^{(1)} = \begin{bmatrix} 12 \\ 122 \\ \vdots \\ 163 \end{bmatrix}$$



$$x^{(2)} = \begin{bmatrix} 231 \\ 146 \\ \vdots \\ 0 \end{bmatrix}$$



$$x^{(3)} = \begin{bmatrix} 22 \\ 13 \\ \vdots \\ 89 \end{bmatrix}$$

$$X = \begin{bmatrix} 12 & 231 & 22 \\ 122 & 146 & 13 \\ \vdots & \vdots & \vdots \\ 163 & 0 & 89 \end{bmatrix}$$

$$Y = [1,0,1]$$

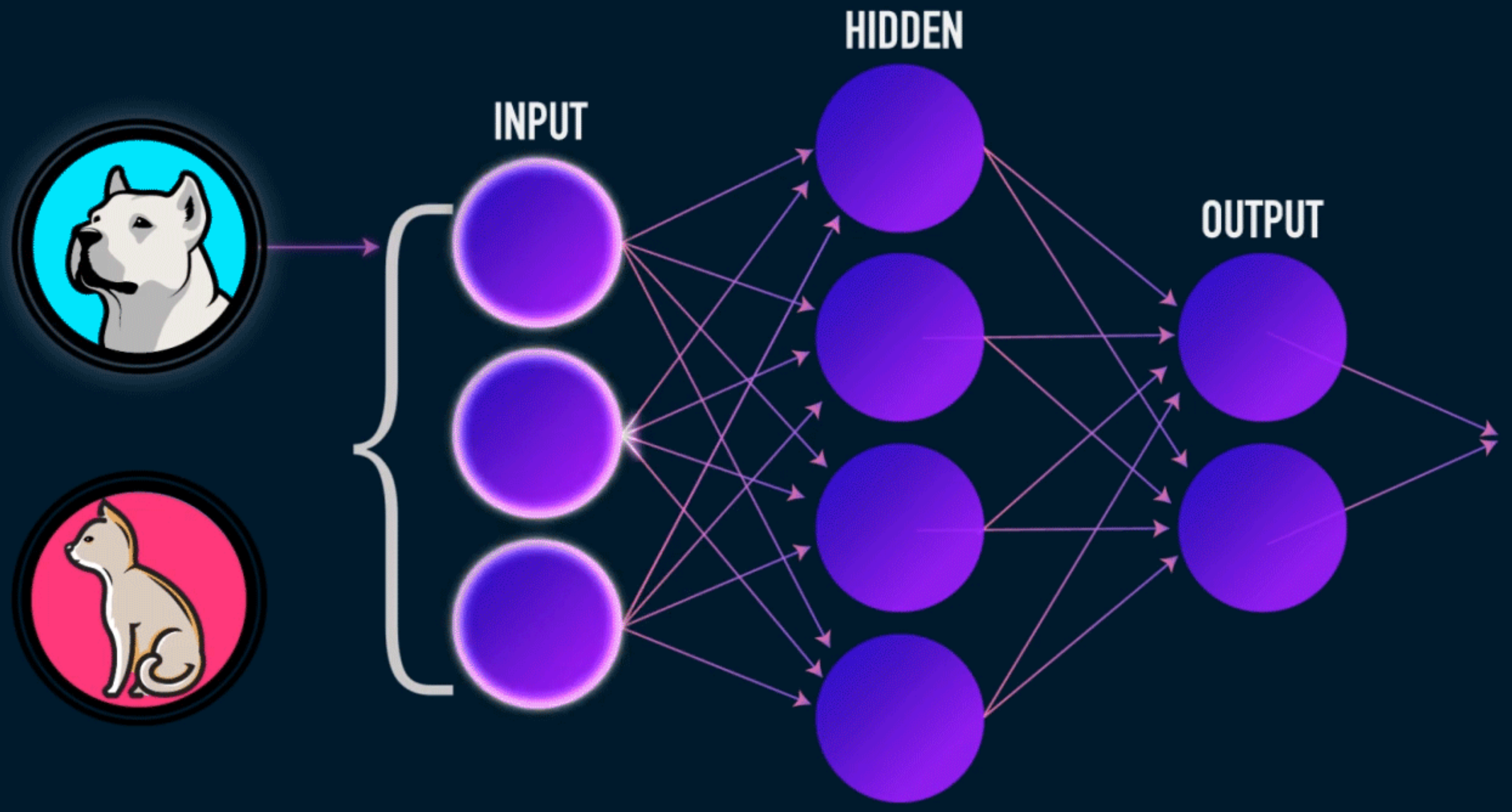
$$n_x = 12288$$
$$m = 3$$

Para um conjunto de *m* amostras de treinamento

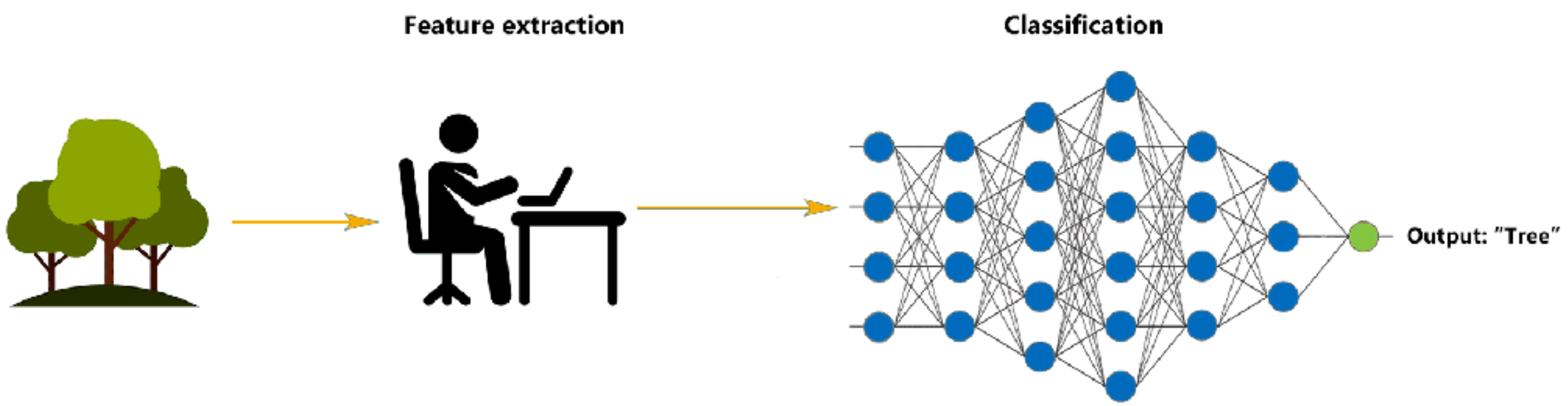
m**-amostras:** $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$

$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \ldots & x^{(m)} \\ | & | & & | \end{bmatrix}$$

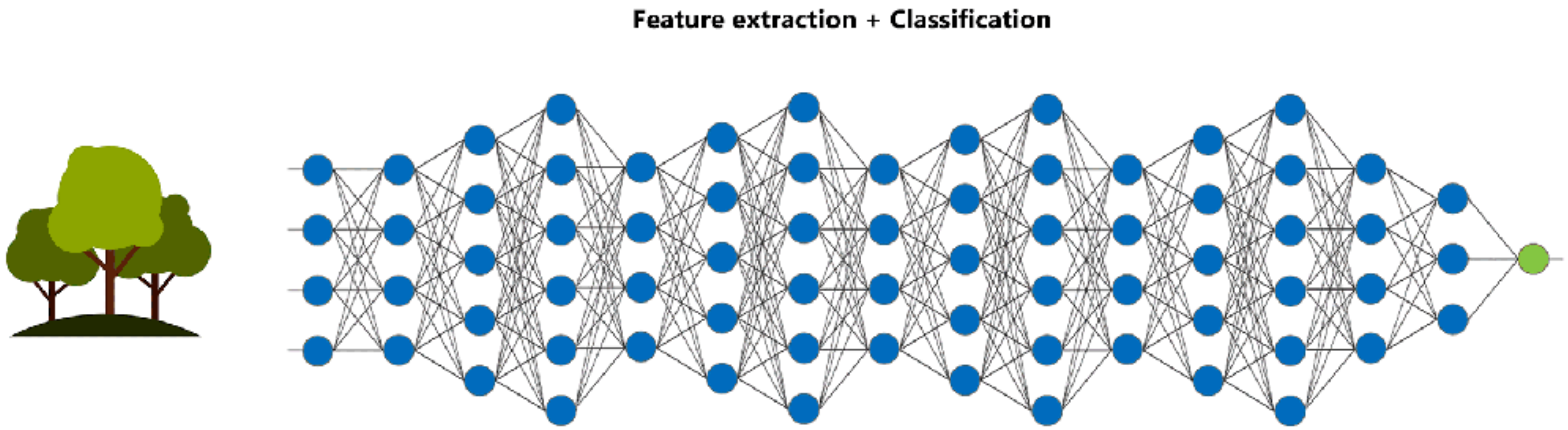
$$Y = \left[y^{(1)}, y^{(1)}, \ldots, y^{(m)}\right]$$

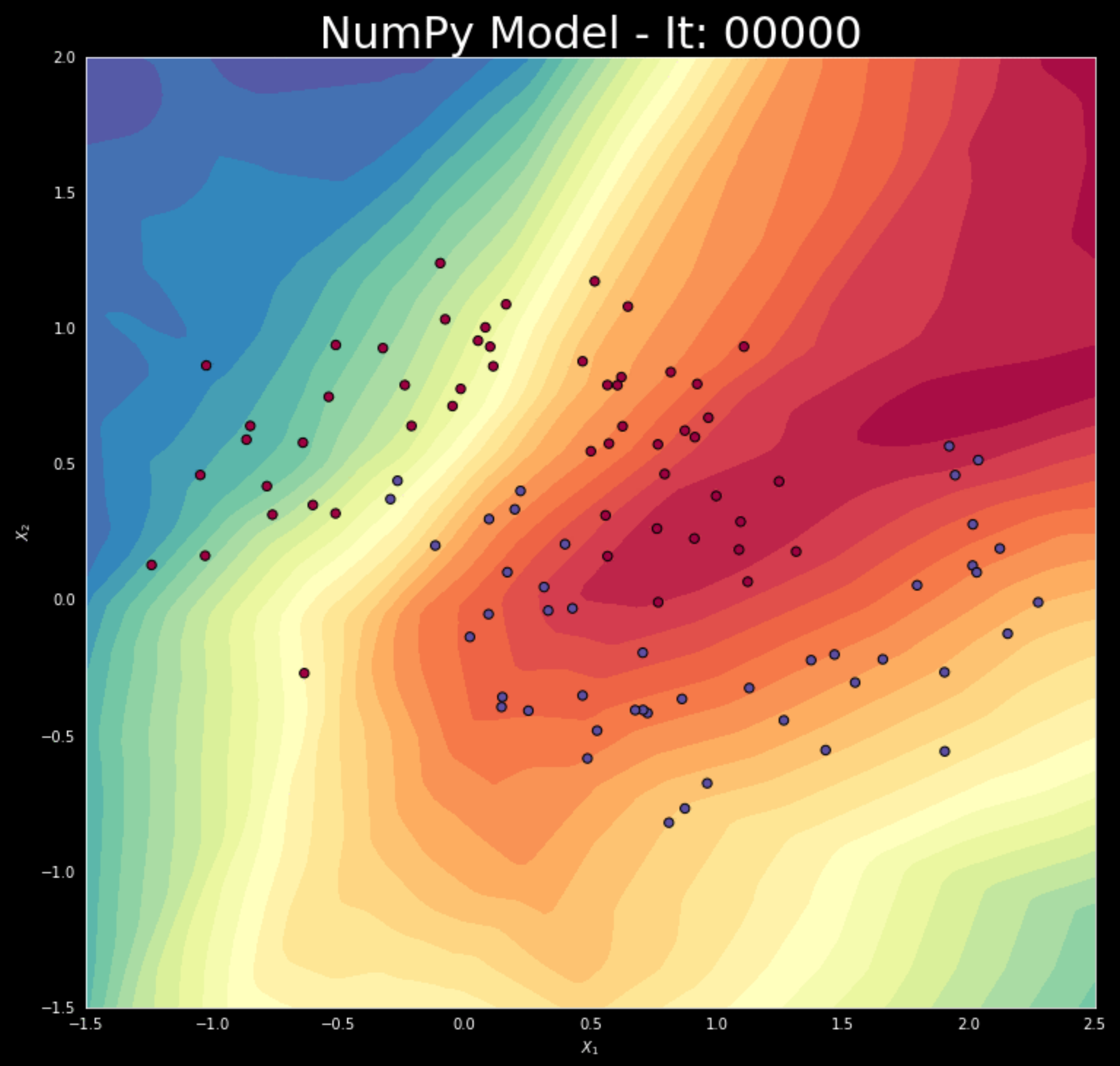

INPUT
HIDDEN
OUTPUT

Machine Learning



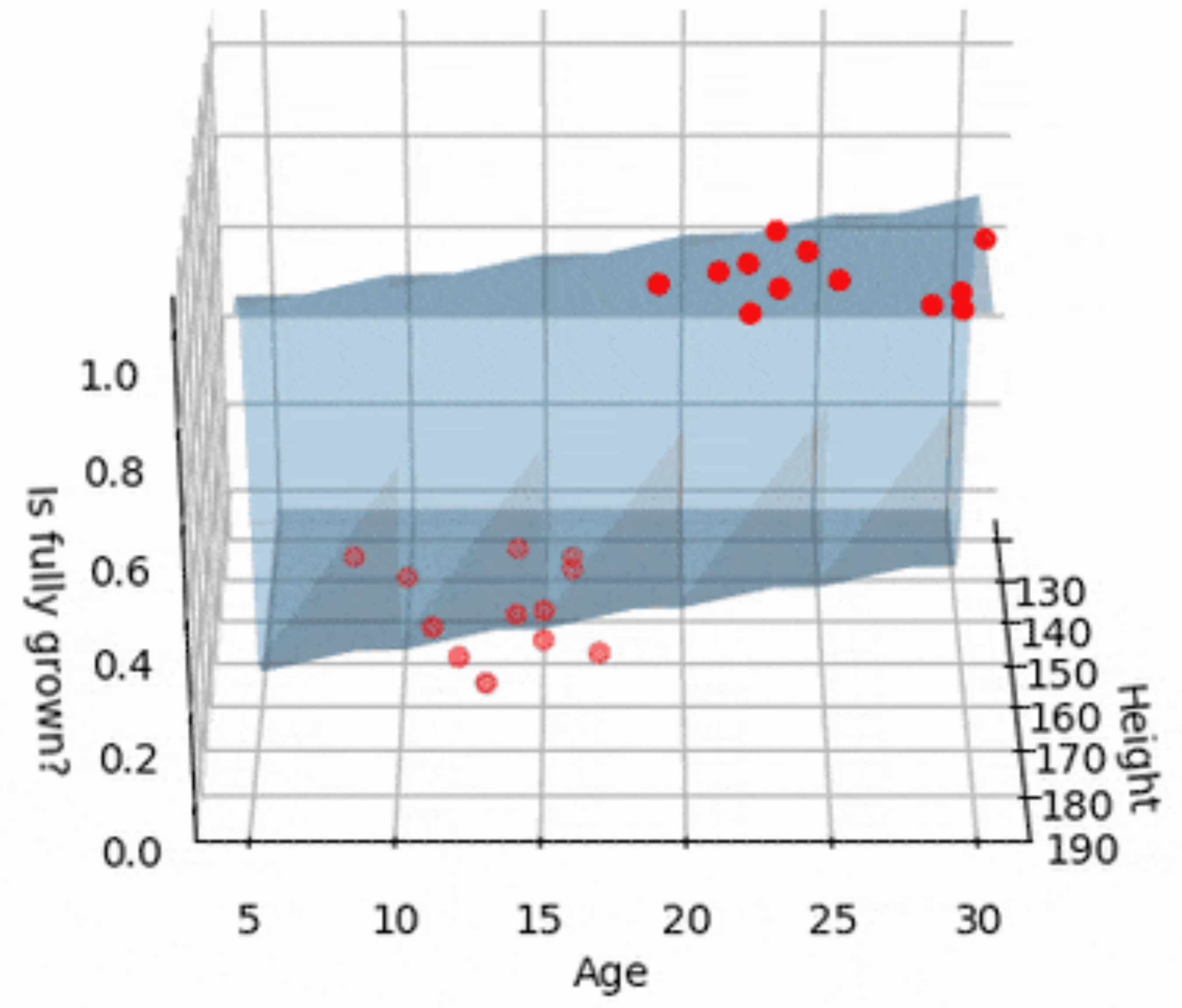
Deep Learning



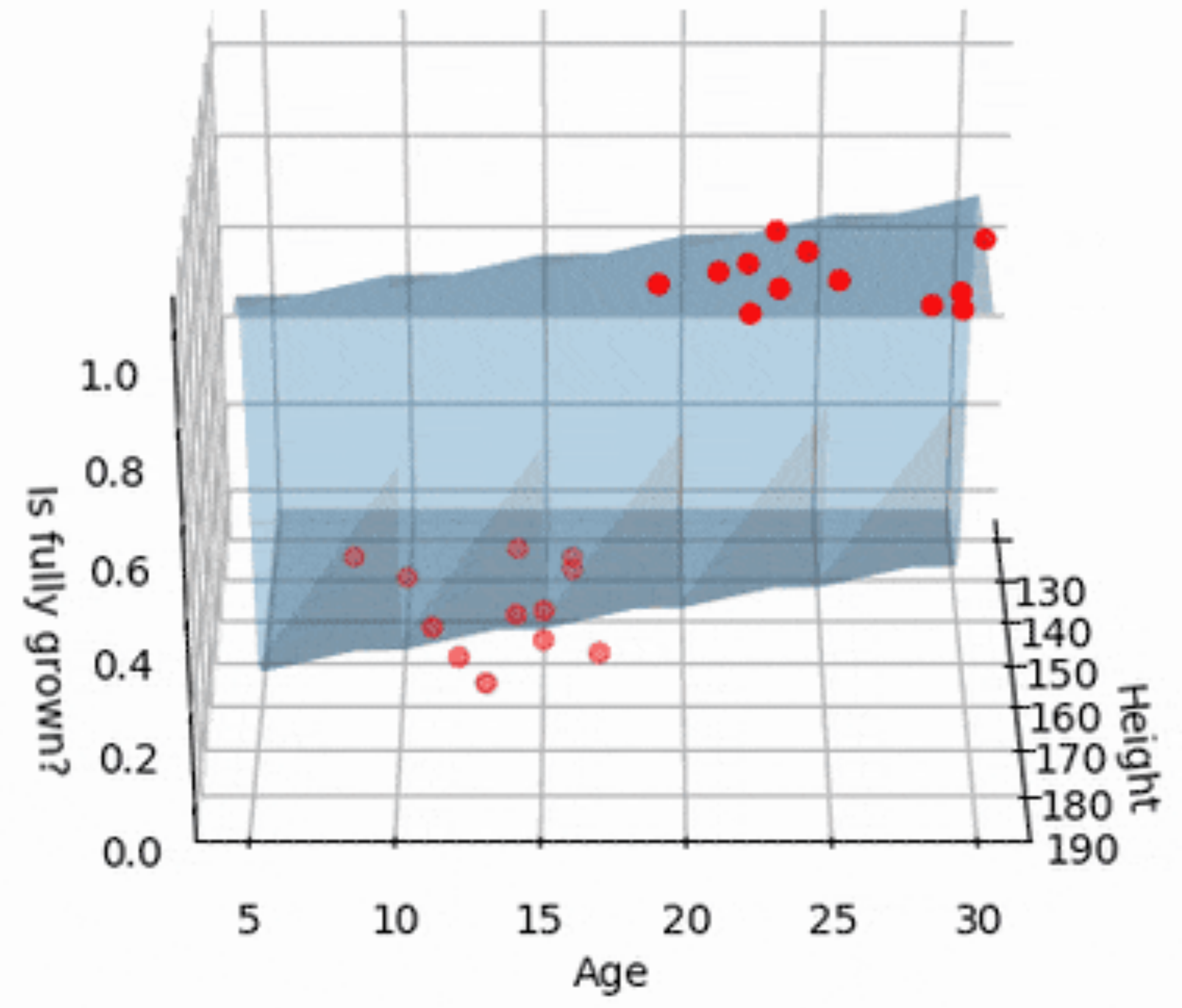


NumPy Model - It: 00000
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
X_2
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
X_1

REGRESSÃO LOGÍSTICA



REGRESSÃO LOGÍSTICA



$$\hat{y} = P(y = 1 \mid x)$$

1. SCORE FUNCTION

A score function é uma função que vai receber o nosso input (no exemplo usado aqui, uma imagem) e mapear os dados para as classes de labels.

$$f(x, \theta, b) = \hat{y} = \theta^T x + b$$

Lembra que estamos falando de classificação de imagens, e da probabilidade dessa imagem ser um avião ou não?

O problema é que a nossa score function, do jeito que está, pode retornar qualquer valor.

$$0 \leq \hat{y} \leq 1$$

FUNÇÕES DE ATIVAÇÃO

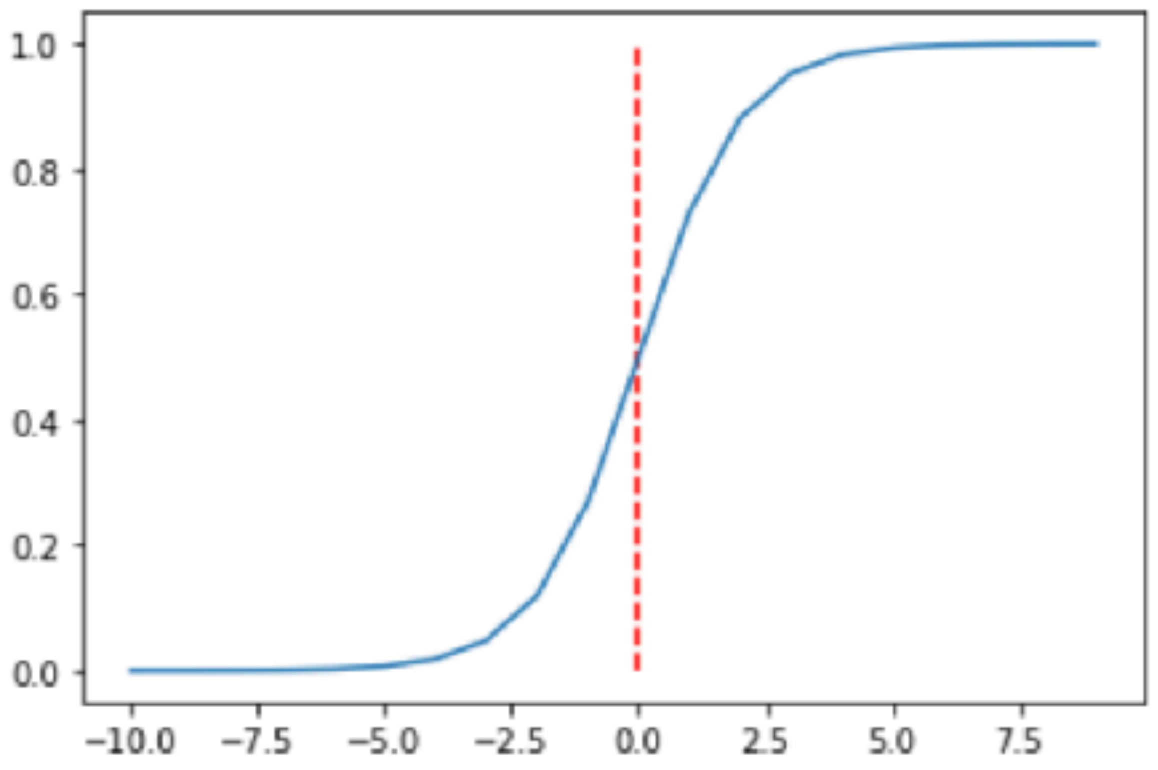
FUNÇÃO SIGMÓIDE

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

```
[ ]  # importar pacotes necessários
     import numpy as np
     import matplotlib.pyplot as plt

     # criar um set entre -10 e 10 e aplicar a função sigmóide
     x = np.arange(-10, 10)
     y = 1 / (1 + np.exp(-x))

     # plotar a curva sigmoidal
     plt.plot(x, y)
     plt.vlines(0, 0, 1, colors='r', linestyles='dashed')
     plt.show()
```

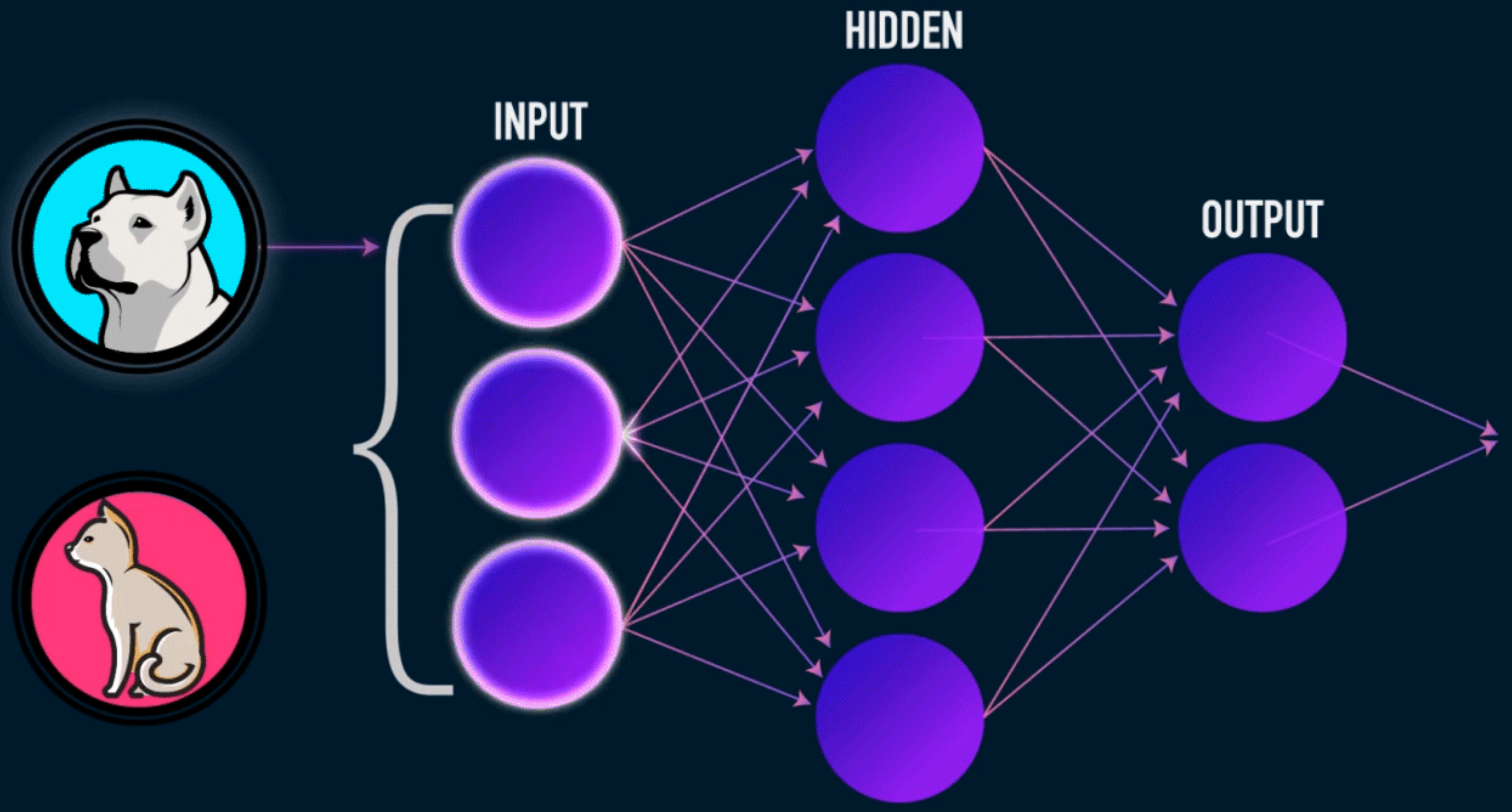


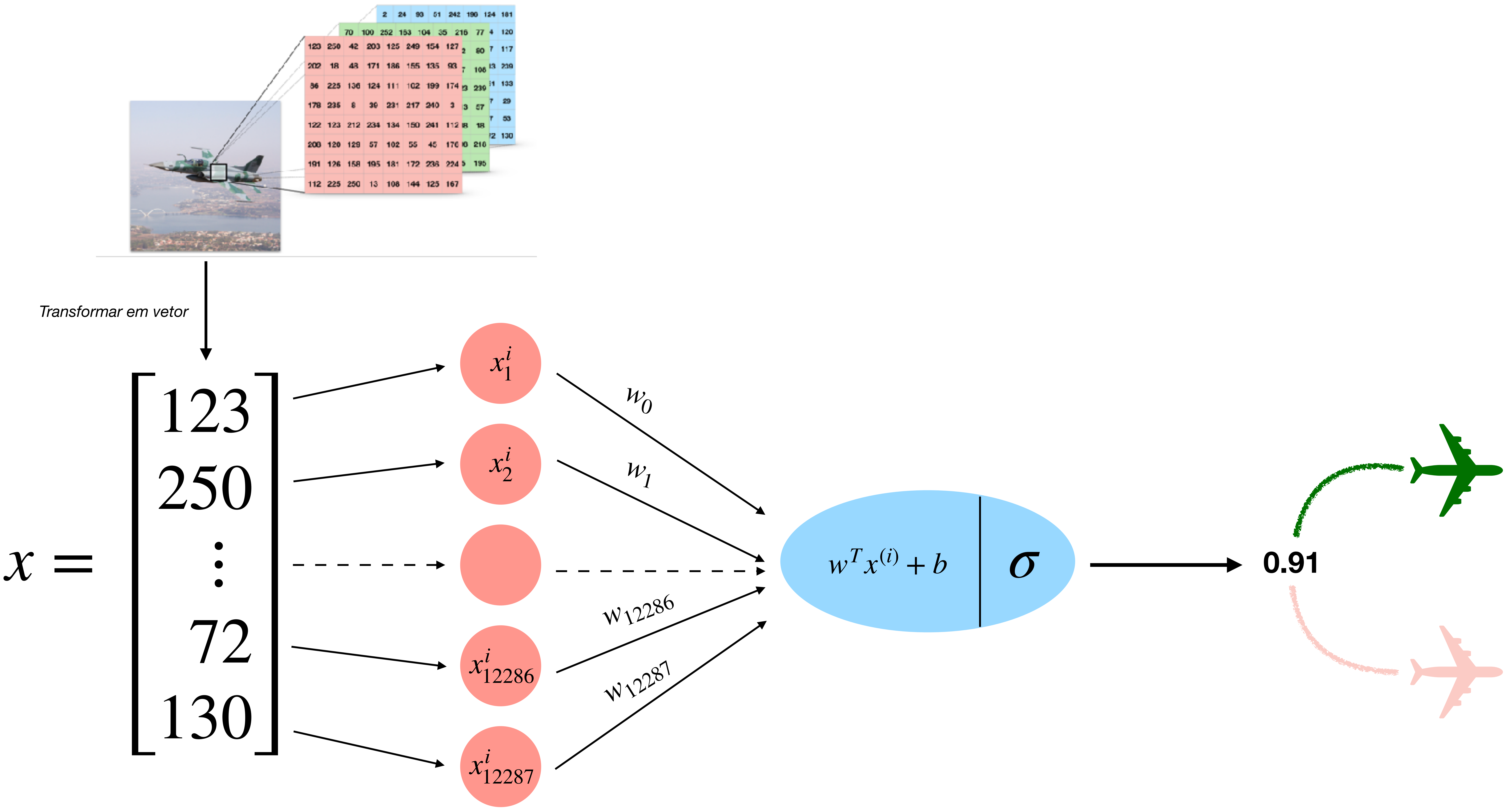
1. SCORE FUNCTION

A score function é uma função que vai receber o nosso input (no exemplo usado aqui, uma imagem) e mapear os dados para as classes de labels.

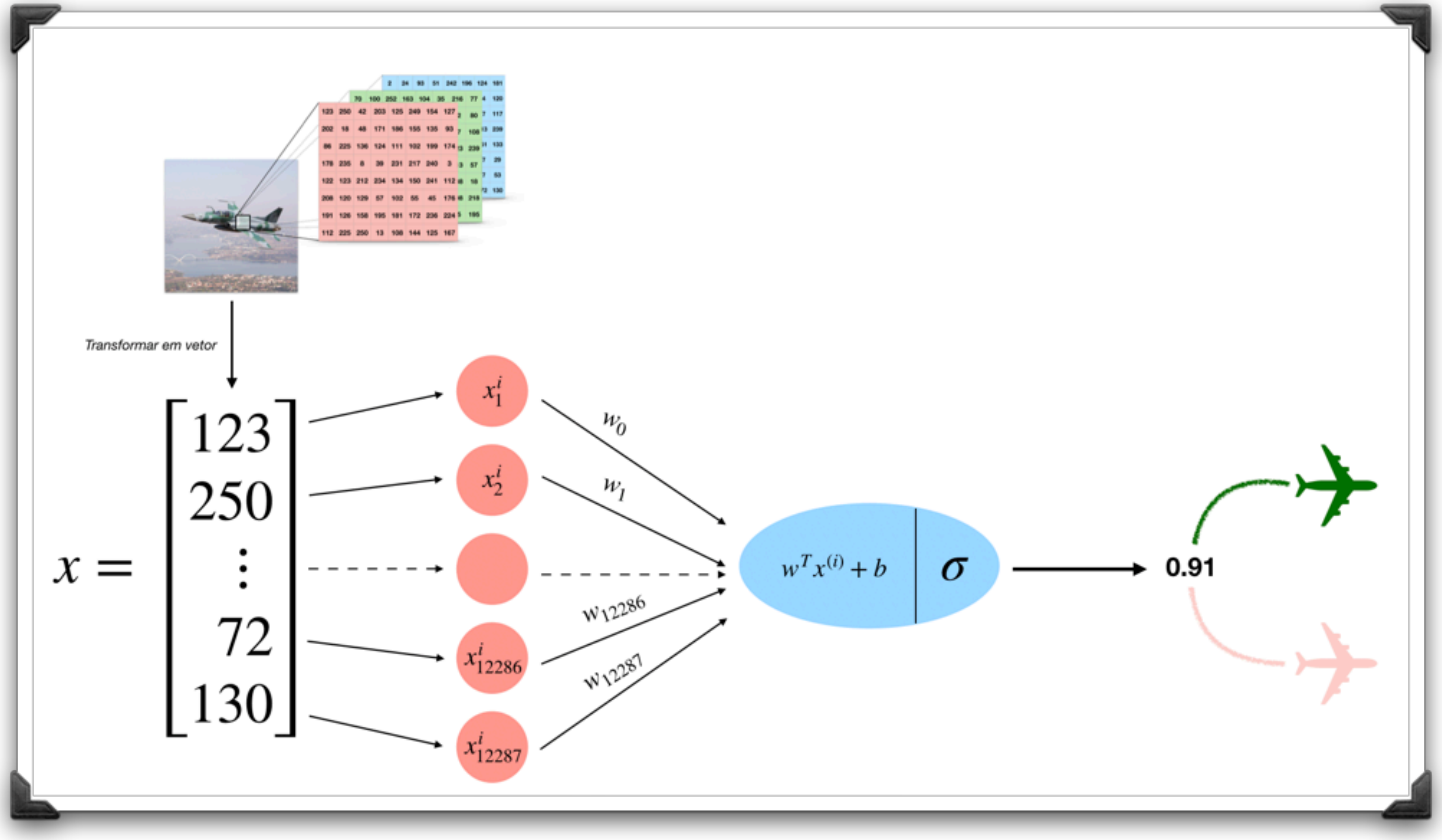
$$f(x, \theta, b) = \hat{y} = \theta^T x + b$$

$$\hat{y} = \sigma(\theta^T x + b) = \frac{1}{1 + e^{-(\theta^T x + b)}}$$


HIDDEN
INPUT
OUTPUT



Transformar em vetor
$x = \begin{bmatrix} 123 \\ 250 \\ \vdots \\ 72 \\ 130 \end{bmatrix}$
x_1^i
x_2^i
x_{12286}^i
x_{12287}^i
w_0
w_1
w_{12286}
w_{12287}
$w^T x^{(i)} + b$
σ
0.91



1. Transformar a imagem em um vetor.
2. Multiplicar o valor de cada pixel de x pelo seu peso w.
3. Obter o valor de z.
4. Obter a probabilidade de ser um avião. Ou seja, obter um valor entre 0 e 1.
5. Classificar a imagem com o label "*avião*".

2. COST FUNCTION

Como saber se nosso algoritmo está indo bem?

$$\mathcal{L}(\hat{y}, y)$$

LOSS FUNCTION

1. COST FUNCTION

$$\mathscr{L}(\hat{y}, y) = -\left(y \log \hat{y} + (1 - y) log(1 - \hat{y})\right)$$

LOSS FUNCTION

1. COST FUNCTION

$$\mathscr{L}(\hat{y}, y) = -\,(y \log \hat{y} + (1 - y) log(1 - \hat{y}))$$

LOSS FUNCTION

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}(\hat{y}^{(i)}, y^{(i)})$$

COST FUNCTION

1. COST FUNCTION

$$\mathscr{L}(\hat{y}, y) = -(y \log \hat{y} + (1-y) log(1-\hat{y}))$$

LOSS FUNCTION

$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}\mathscr{L}(\hat{y}^{(i)}, y^{(i)})$$

COST FUNCTION

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) log(1-\hat{y}^{(i)})]$$

DIMENSÕES ERRADAS!

MAIORES PROBLEMAS NA IMPLEMENTAÇÃO DE REDES NEURASI

$$z = w^T x + b$$

$$z = wx + b$$

1 x m

num_pixels x 1 —— num_pixels x m

escalar

$$z = wx + b$$

1 x m — num_pixels x 1 — num_pixels x m — escalar

$$z = w^T x + b$$

1 x m — 1 x num_pixels — num_pixels x m — escalar

$$z = w^T x + b$$

1 x m

1 x num_pixels num_pixels x m

escalar

$$\sigma(-z) = \frac{1}{1 + e^{(-z)}}$$

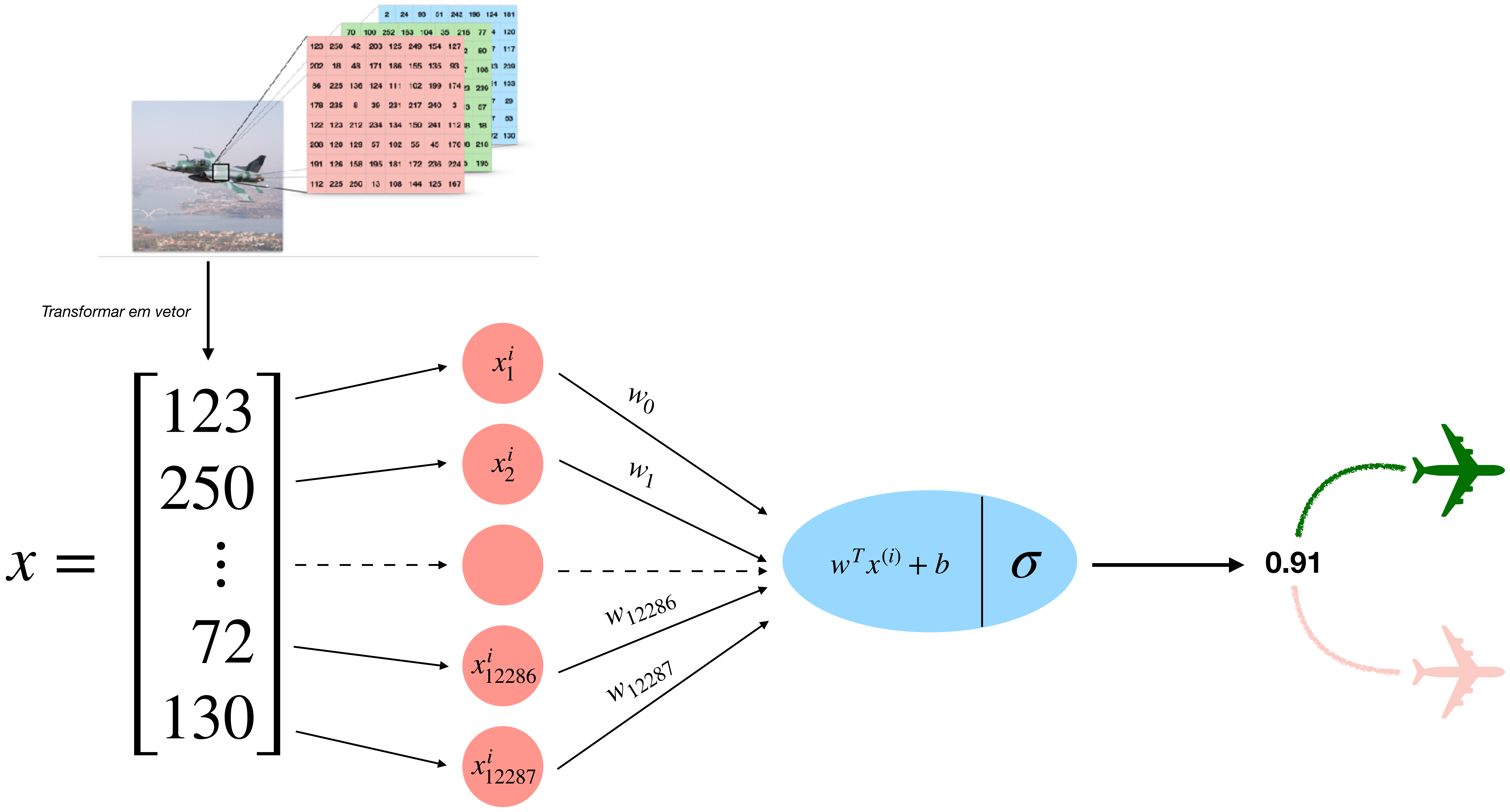
$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log\hat{y}^{(i)} + (1 - y^{(i)})log(1 - \hat{y}^{(i)})]$$

1 x m

1 x m

1 x m

1 x m


Transformar em vetor
x =
123
250
72
130
x_1^i
x_2^i
x_12286^i
x_12287^i
w_0
w_1
w_12286
w_12287
w^T x^(i) + b
σ
0.91

m**-amostras:** $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \ldots, (x^{(m)}, y^{(m)})\}$

$$X = \begin{bmatrix} | & | & & | \\ x^{(1)} & x^{(2)} & \ldots & x^{(m)} \\ | & | & & | \end{bmatrix}$$

$$\hat{Y} = \left[y^{(1)}, y^{(1)}, \ldots, y^{(m)}\right]$$

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})log(1-\hat{y}^{(i)})]$$

1 x m

1 x m

1 x m

1 x m

1 x m

$$a = \begin{bmatrix} 2 & 3 & 4 \end{bmatrix}$$

$$b = \begin{bmatrix} 6 & 4 & 4 \end{bmatrix}$$

$$
a \circ b = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}_{(n \times 1)} \circ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}_{(n \times 1)} = \begin{bmatrix} a_1 b_1 \\ a_2 b_2 \\ a_3 b_3 \\ a_4 b_4 \\ a_5 b_5 \end{bmatrix}_{(n \times 1)}
$$

Element wise Product

$$a \cdot b = \underset{(1 \times n)}{\begin{bmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \end{bmatrix}} \underset{(n \times 1)}{\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix}} = \left\{ a_1b_1 + a_2b_2 + a_3b_3 + a_4b_4 + a_5b_5 \right\}$$

Dot Product

$$a \otimes b = ab^T = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}_{(n \times 1)} \begin{bmatrix} b_1 & b_2 & b_3 & b_4 & b_5 \end{bmatrix}_{(1 \times n)} = \begin{Bmatrix} a_1b_1 & a_1b_2 & a_1b_3 & a_1b_4 & a_1b_5 \\ a_2b_1 & a_2b_2 & a_2b_3 & a_2b_4 & a_2b_5 \\ a_3b_1 & a_3b_2 & a_3b_3 & a_3b_4 & a_3b_5 \\ a_4b_1 & a_4b_2 & a_4b_3 & a_4b_4 & a_4b_5 \\ a_5b_1 & a_5b_2 & a_5b_3 & a_5b_4 & a_5b_5 \end{Bmatrix}_{(n \times n)}$$

Outer Product

$$\hat{y} = \sigma(w^T x + b)$$

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

$$\hat{y} = \sigma(w^T x + b) \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})]$$

$$\hat{y} = \sigma(w^T x + b) \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

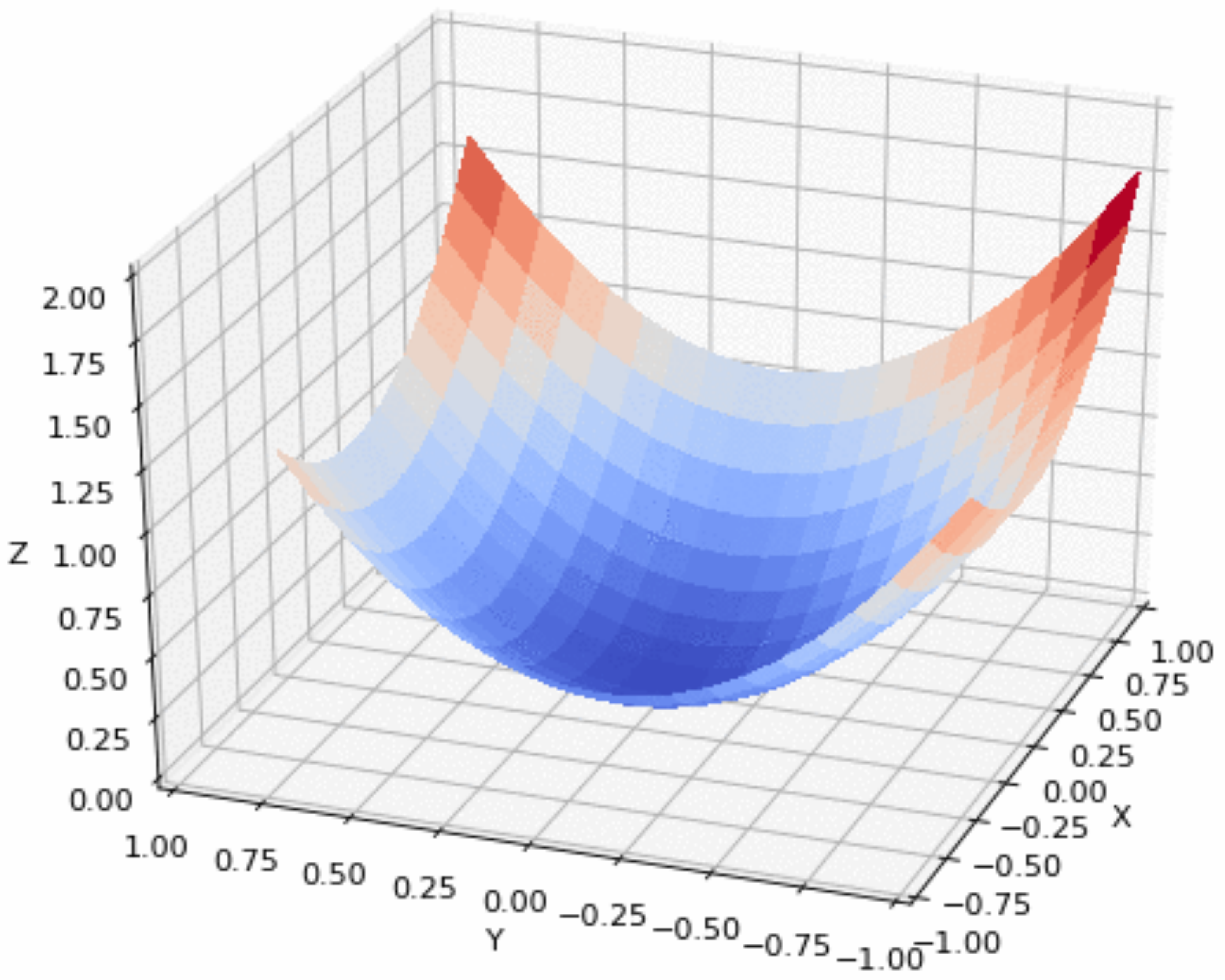
$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})]$$

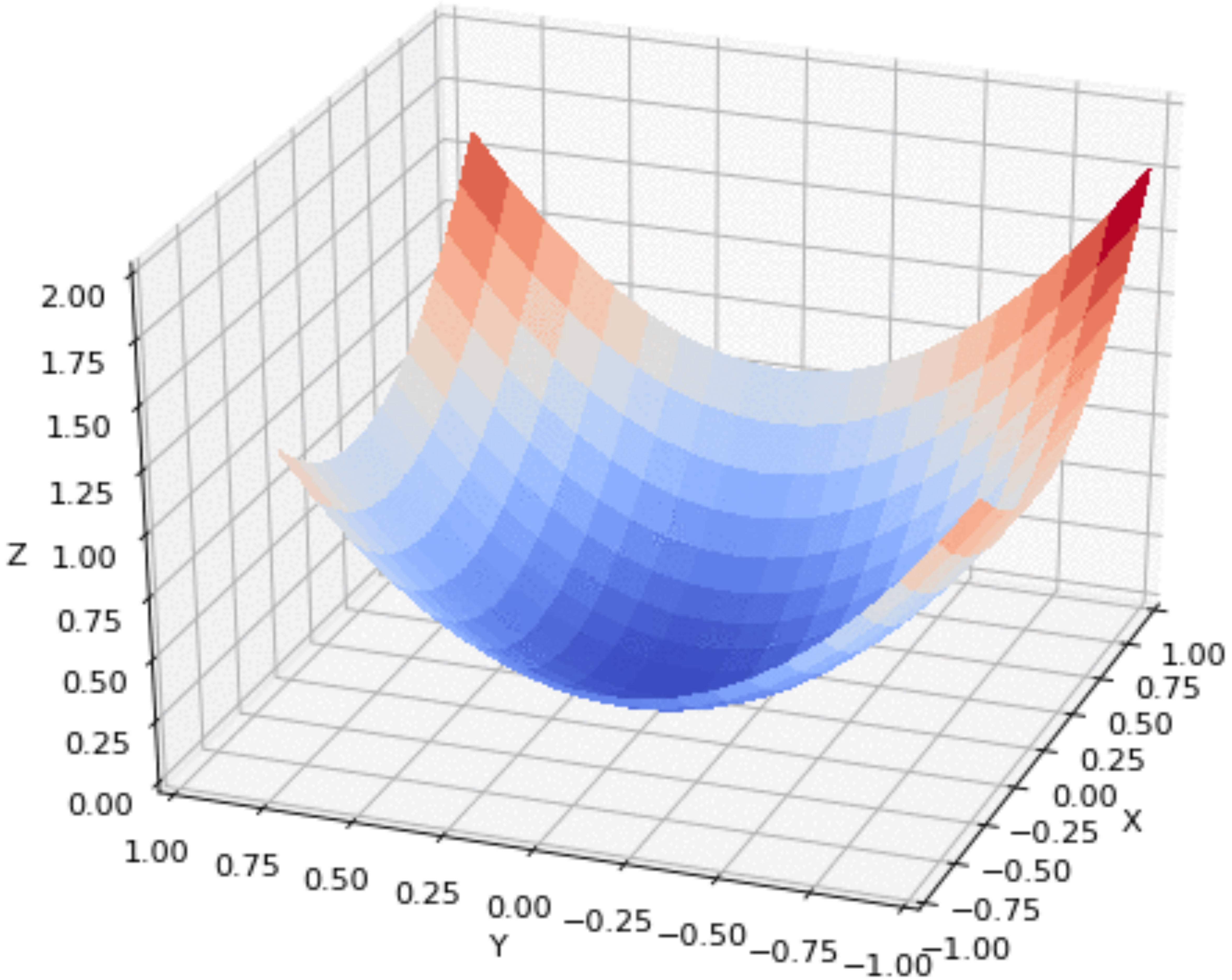
Queremos minimizar J(w, b)

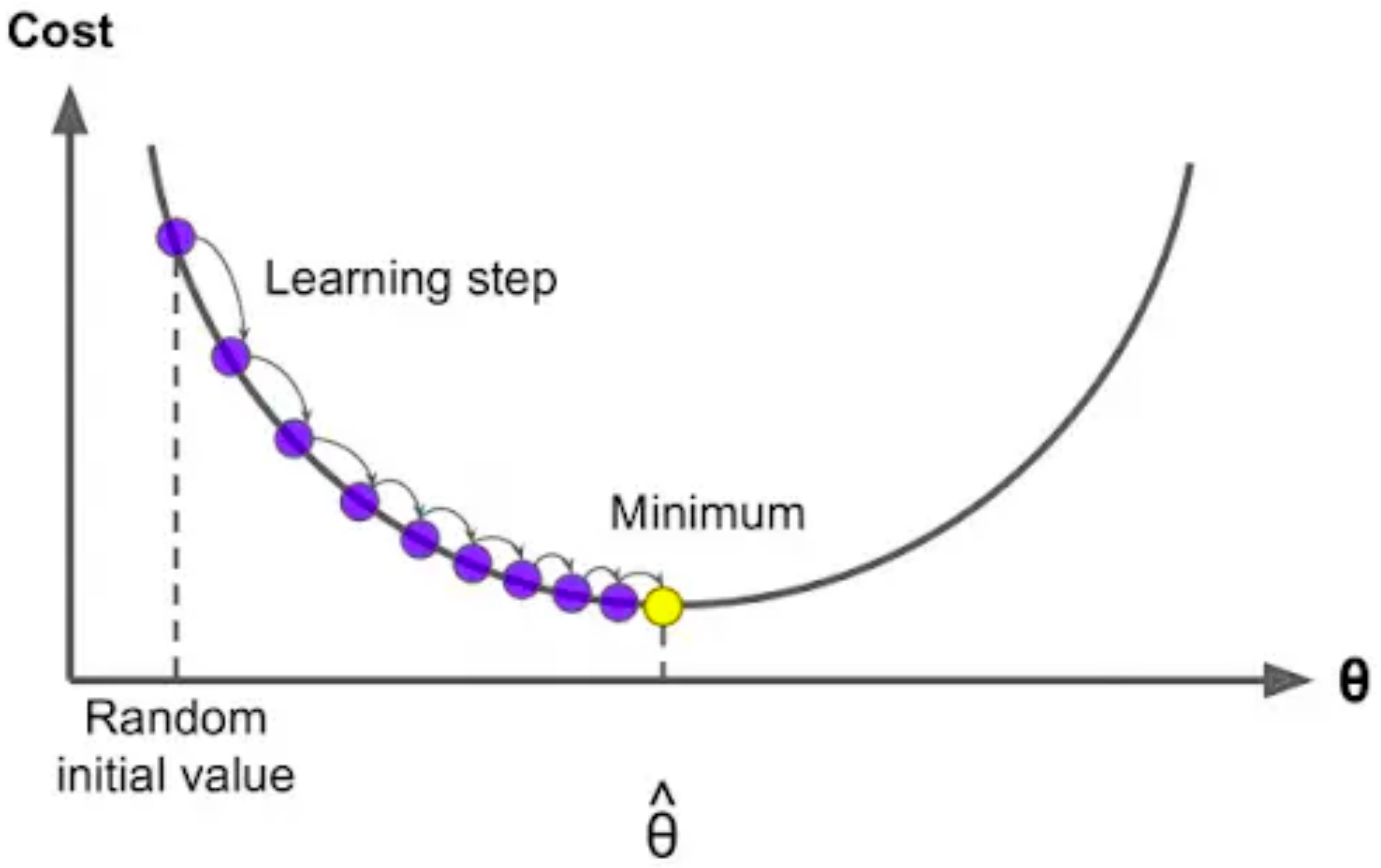
- **Encontrar w e b que representem o menor custo**

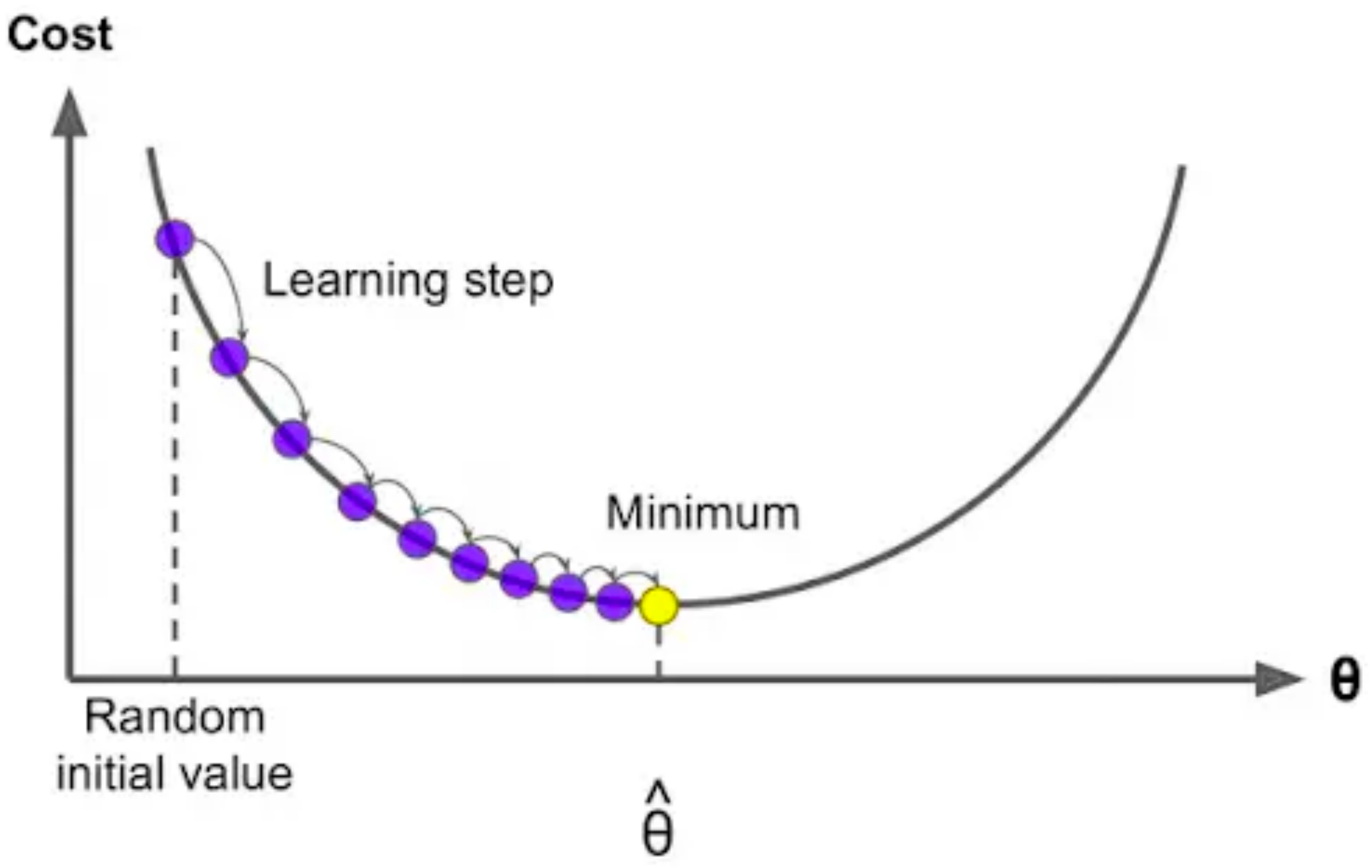
$$\hat{y} = \sigma(w^T x + b) \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

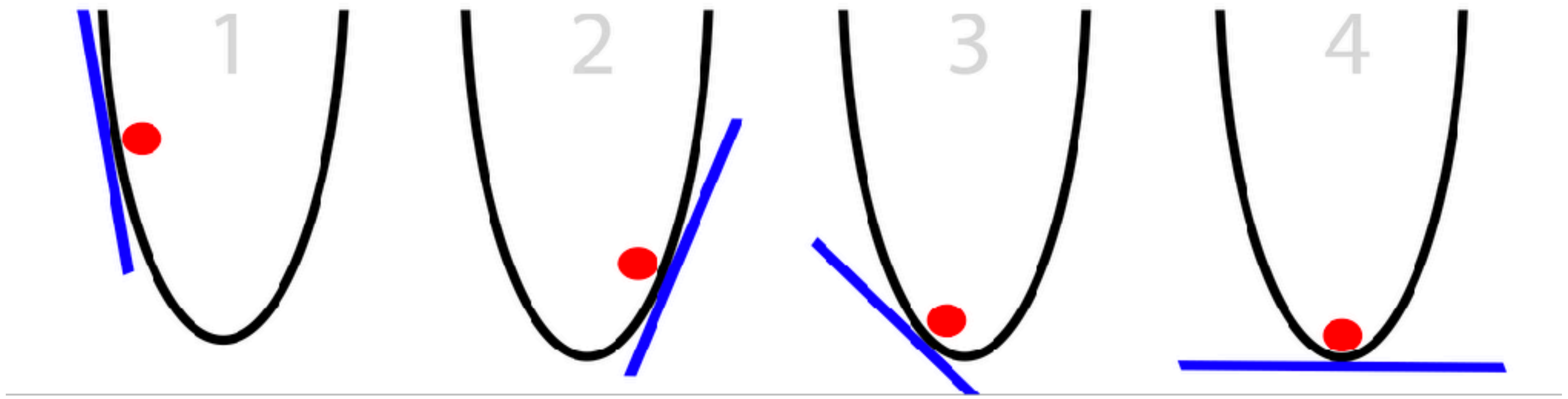
$$J(w,b) = \frac{1}{m}\sum_{i=1}^{m}\mathscr{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{m}[y^{(i)}\log\hat{y}^{(i)} + (1 - y^{(i)})log(1 - \hat{y}^{(i)})]$$

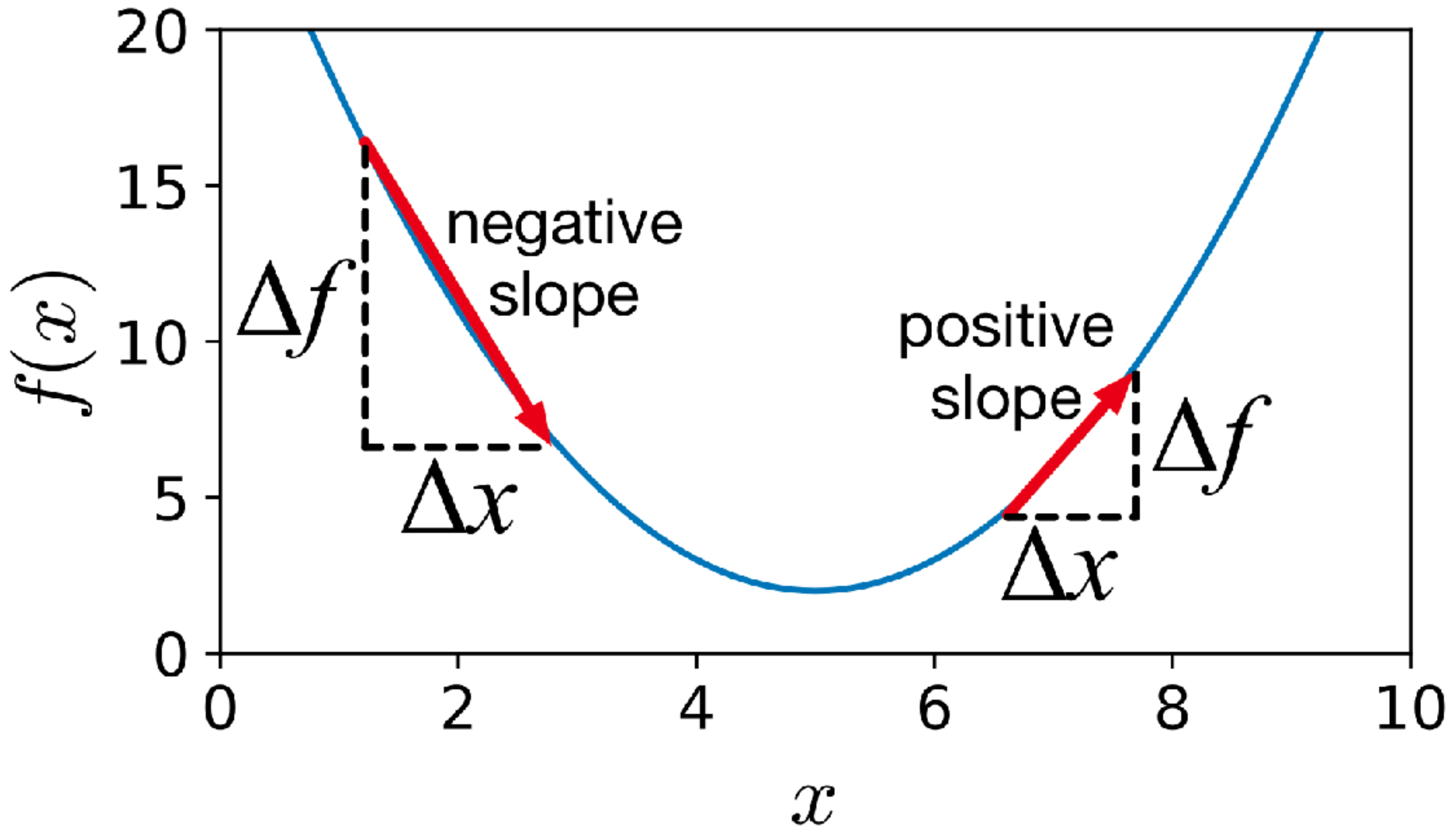



2.00
1.75
1.50
1.25
Z
1.00
0.75
0.50
0.25
0.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Y
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
X


Cost
Learning step
Minimum
θ
Random
initial value
$\hat{\theta}$

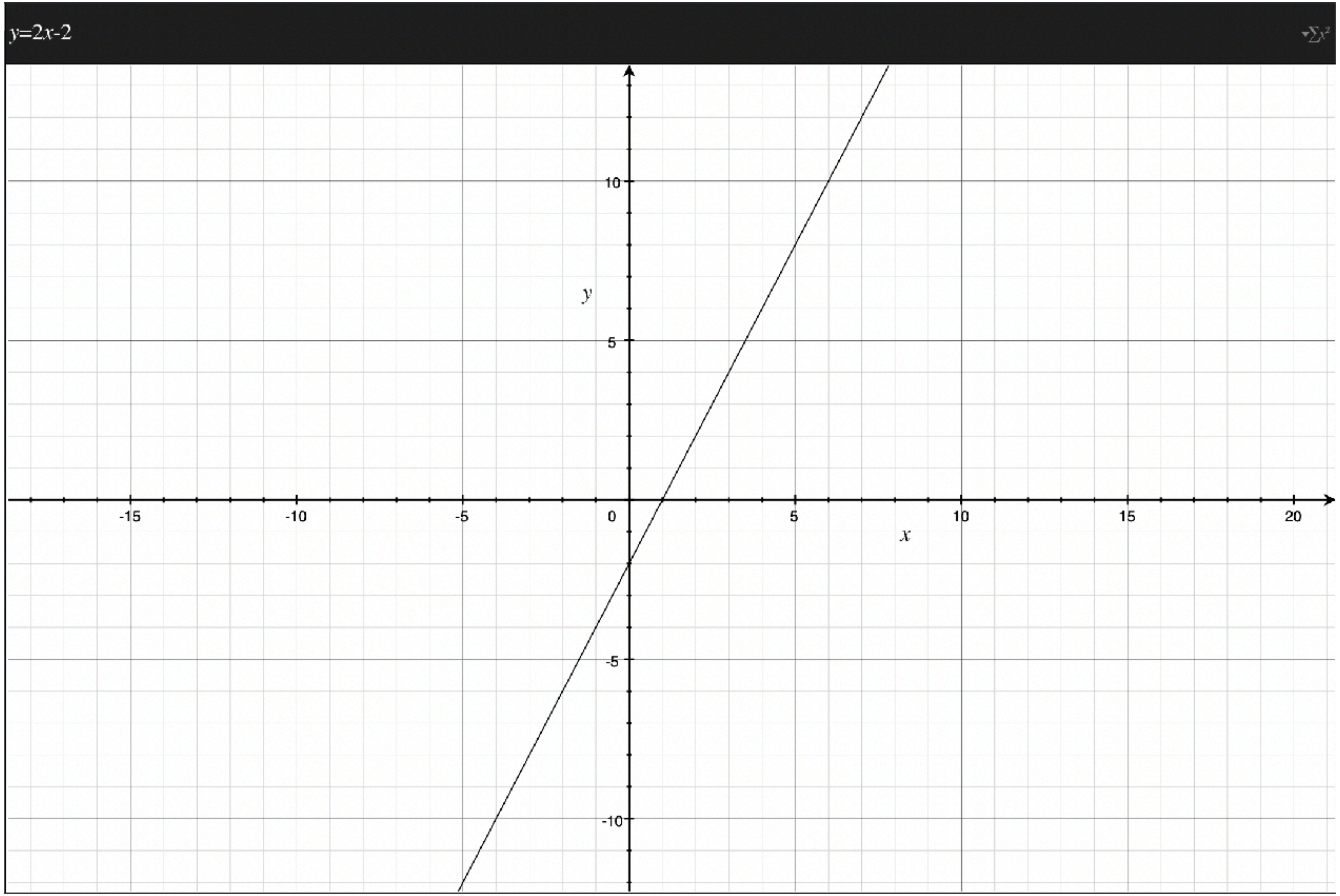

Cost
Learning step
Minimum
θ
Random
initial value
$\hat{\theta}$


1
2
3
4

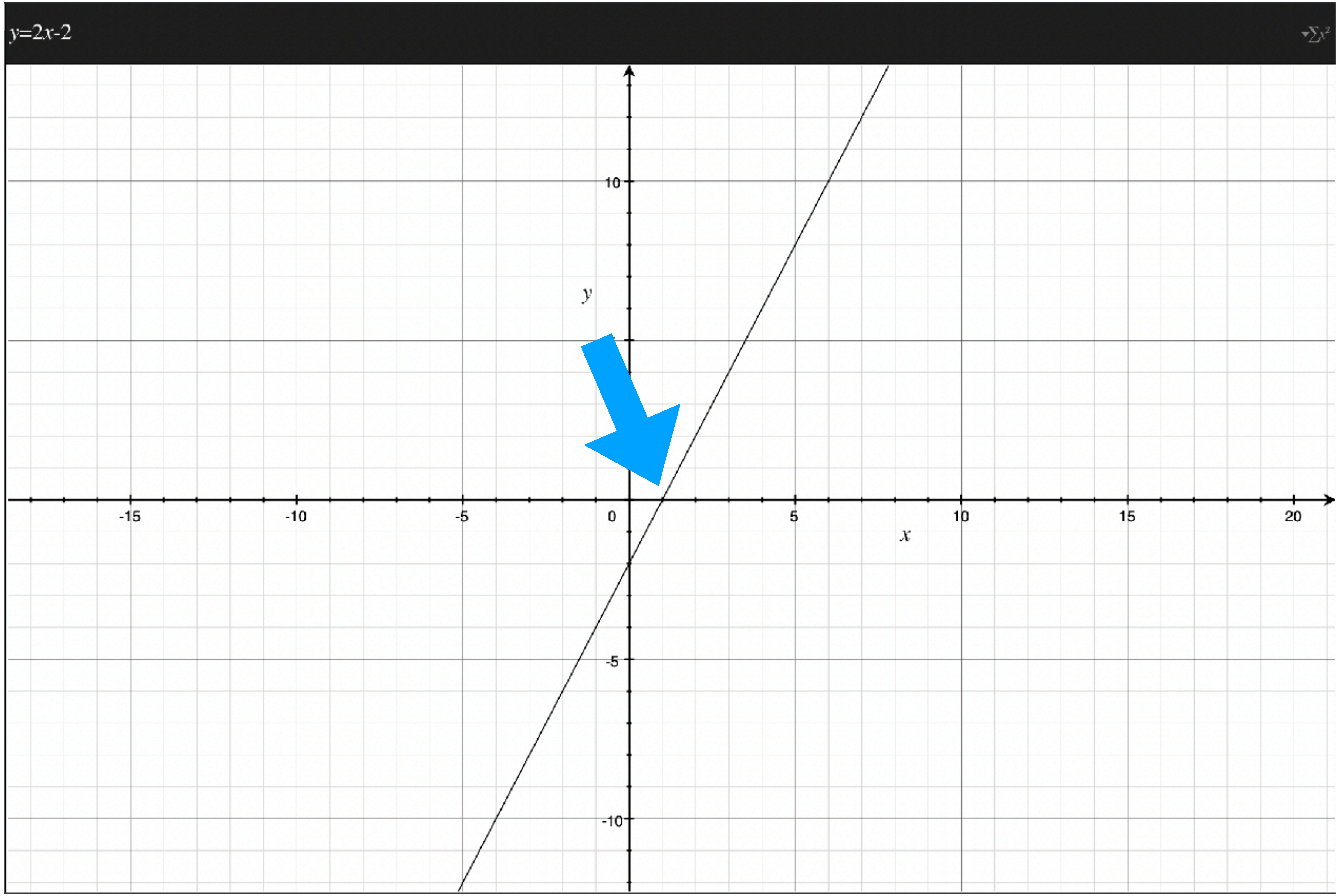


1. CALCULAR A INCLINAÇÃO NO PONTO
2. SE A INCLINAÇÃO FOR POSITIVA, DESLOQUE PARA A ESQUERDA
3. SE A INCLINAÇÃO FOR NEGATIVA, DESLOQUE PARA A DIREITA
4. CONTINUE ATÉ ENCONTRAR INCLINAÇÃO == 0

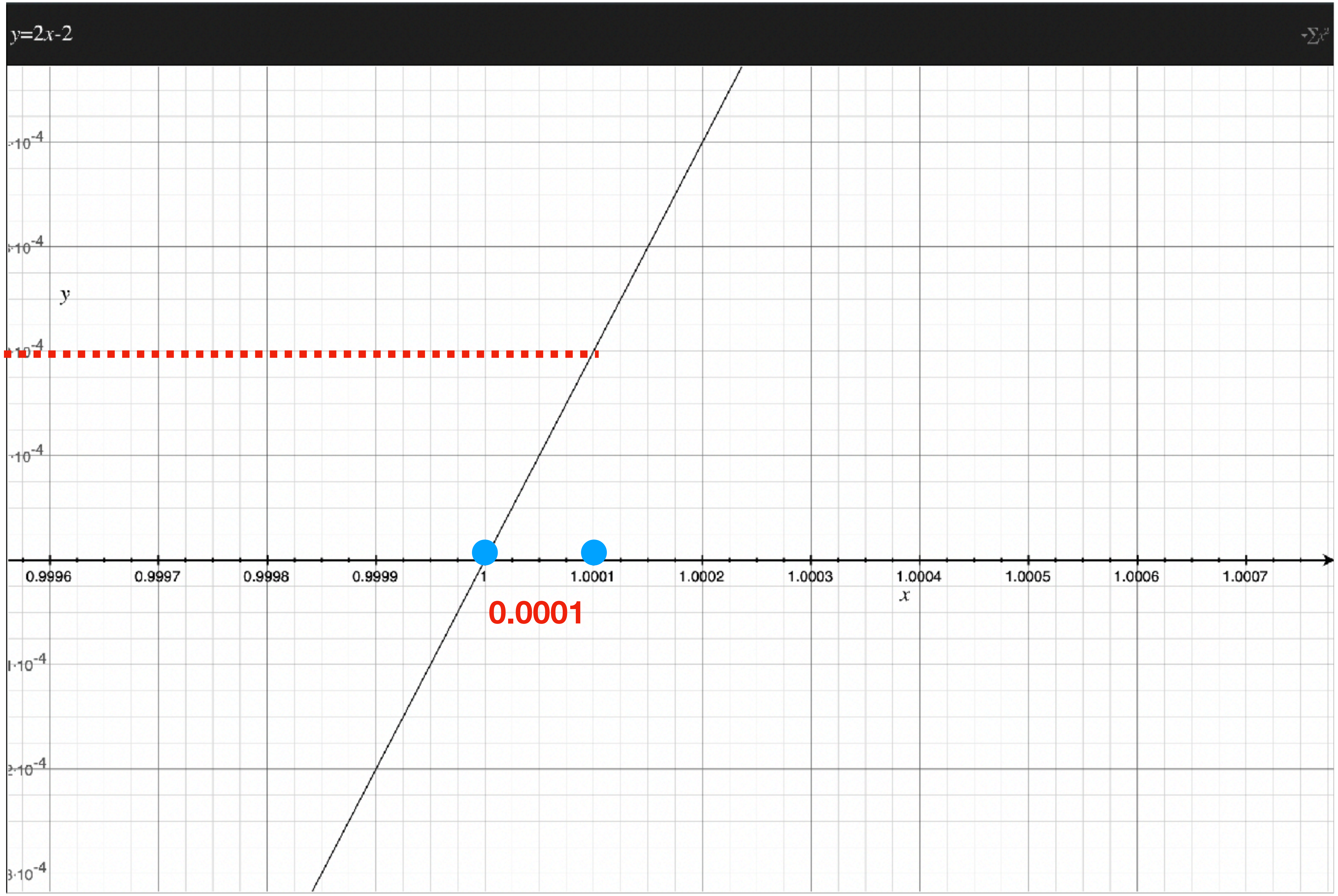
$$f(x) = y = 2x - 2$$



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



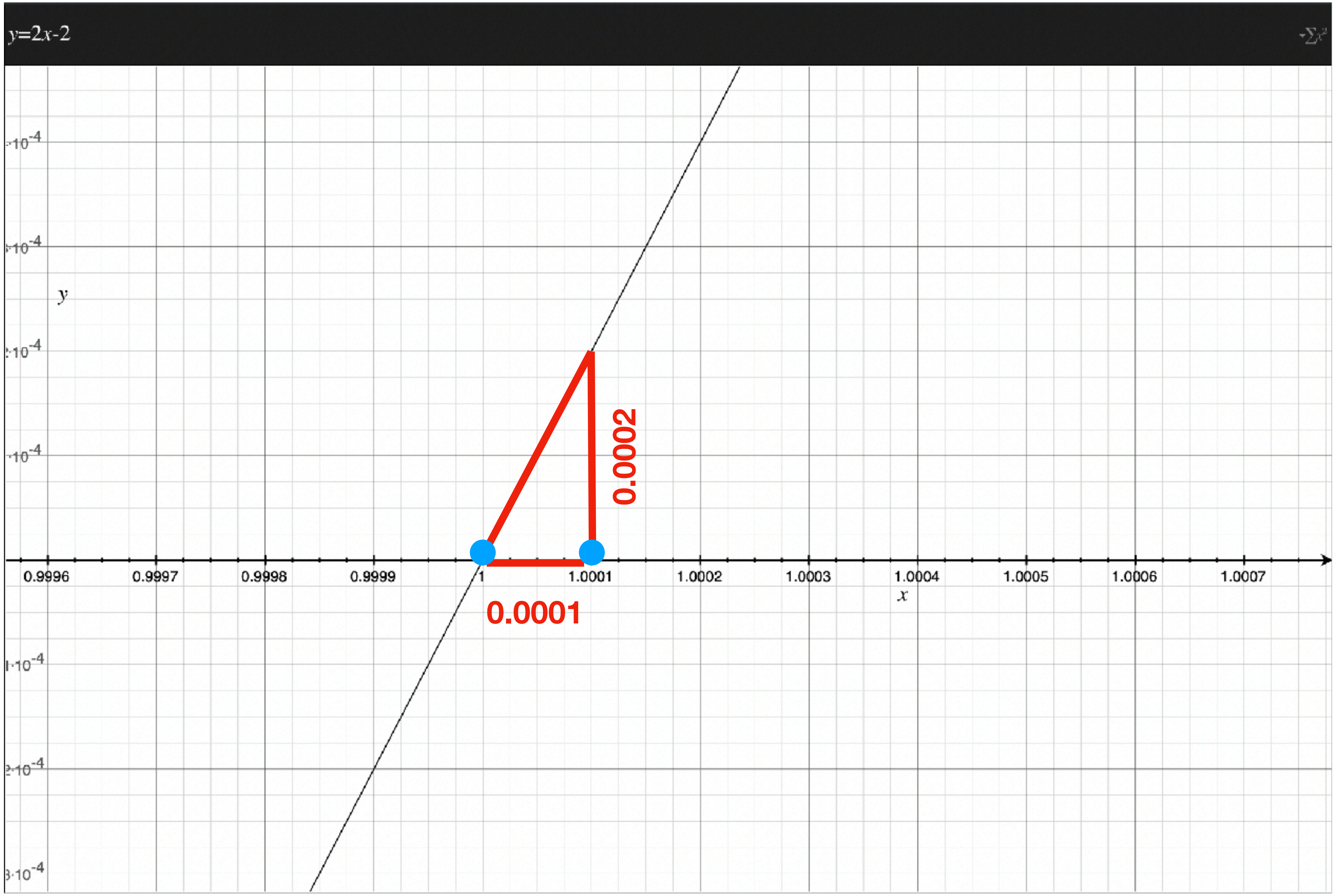
$$y = 2x - 2$$



$$x = 1$$
$$y = 0$$

$$x = 1.0001$$
$$y = 0.0002$$

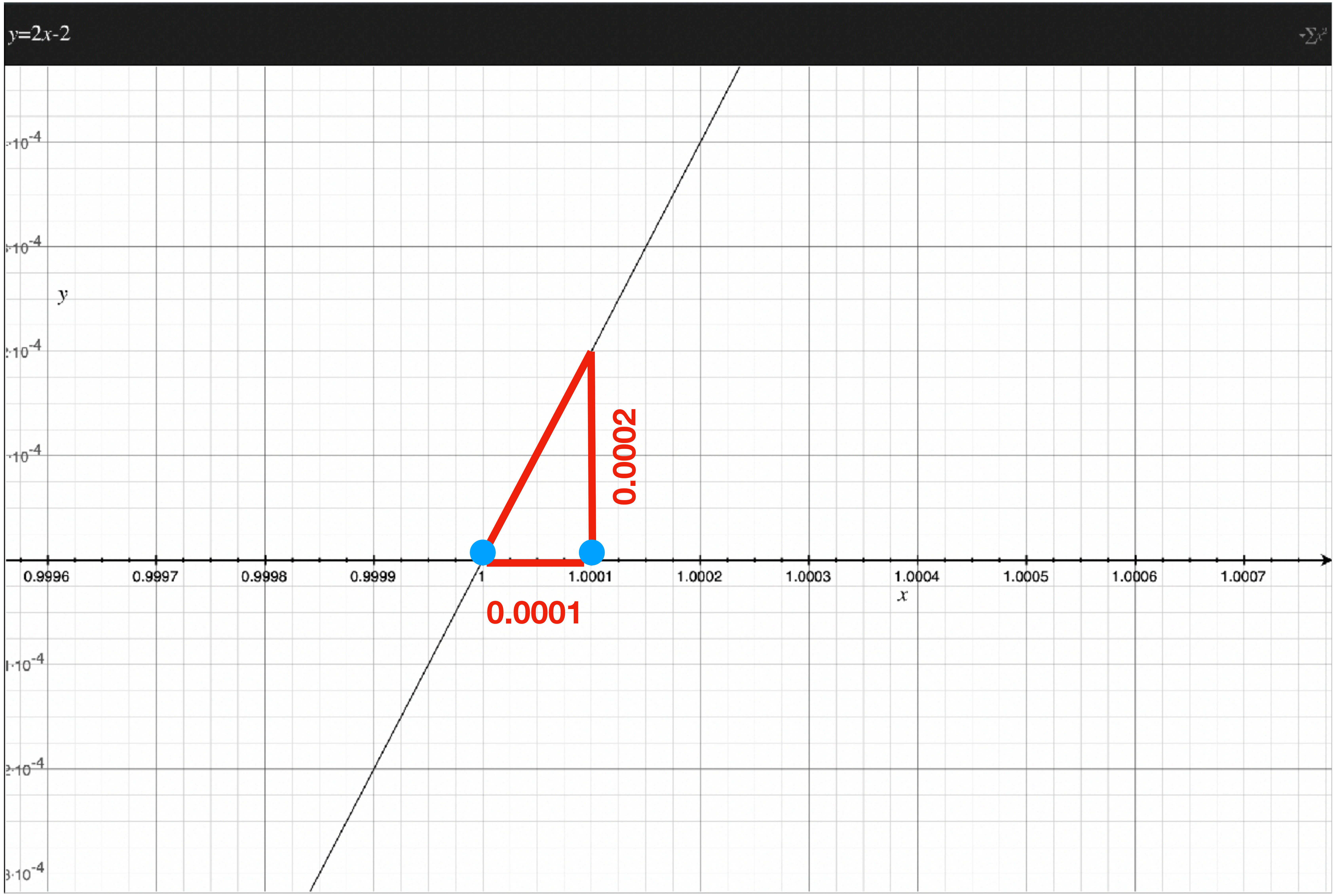
$$y = 2x - 2$$



$$x = 1$$
$$y = 0$$

$$x = 1.0001$$
$$y = 0.0002$$

$$y = 2x - 2$$

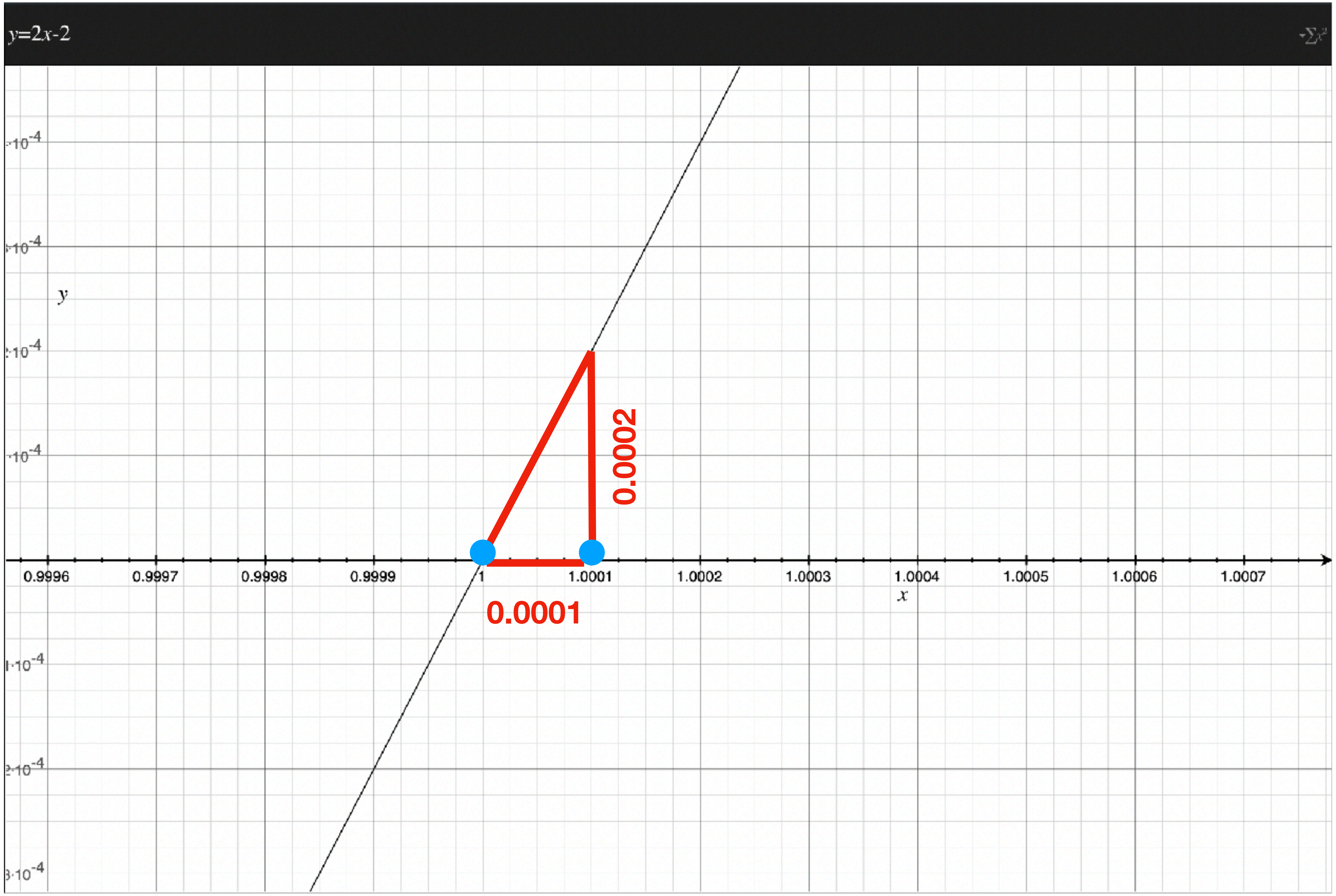


$$x = 1$$
$$y = 0$$

$$x = 1.0001$$
$$y = 0.0002$$

$$\frac{0.0002}{0.0001} = 2$$

$$y = 2x - 2$$



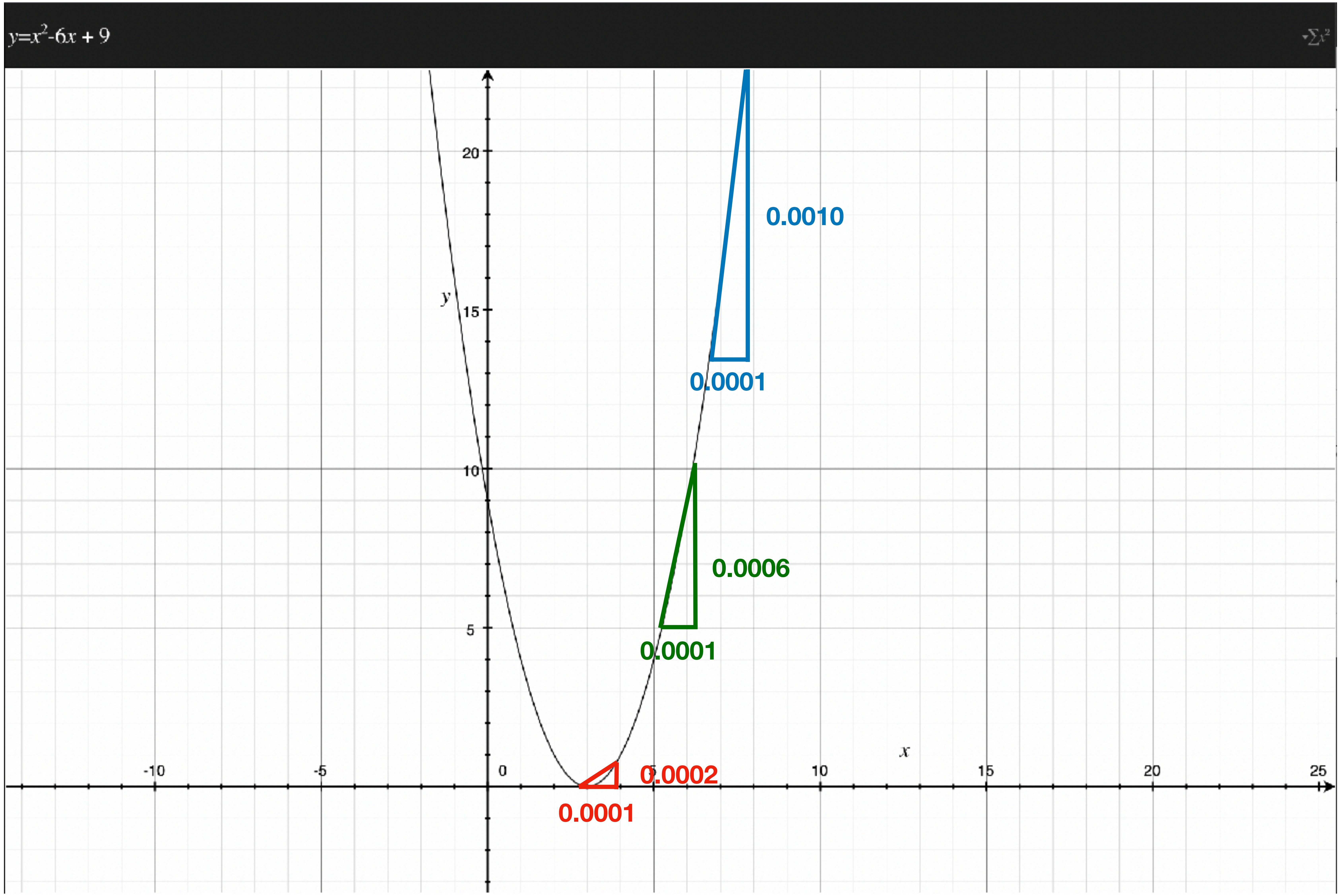
$$x = 1$$
$$y = 0$$

$$x = 1.0001$$
$$y = 0.0002$$

$$\frac{0.0002}{0.0001} = 2$$

$$\frac{dy}{dx} = 2$$

$$y = x^2 - 6x + 9$$



$$\frac{dy}{dx} = 2x - 6$$

$$\frac{dy}{dx} = 2$$ quando *x=4*

$$\frac{dy}{dx} = 6$$ quando *x=6*

$$\frac{dy}{dx} = 10$$ quando *x=8*

CALCULUS

DERIVATIVES AND LIMITS

DERIVATIVE DEFINITION

$$\frac{d}{dx}(f(x)) = f'(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

BASIC PROPERTIES

$$(cf(x))' = c(f'(x))$$

$$(f(x) \pm g(x))' = f'(x) \pm g'(x)$$

$$\frac{d}{dx}(c) = 0$$

MEAN VALUE THEOREM

If f is differentiable on the interval (a,b) and continuous at the end points there exists a c in (a,b) such that

$$f'(c) = \frac{f(b)-f(a)}{b-a}$$

PRODUCT RULE

$$(f(x)g(x))' = f(x)'g(x) + f(x)g(x)'$$

QUOTIENT RULE

$$\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2}$$

POWER RULE

$$\frac{d}{dx}(x^n) = nx^{n-1}$$

CHAIN RULE

$$\frac{d}{dx}\big(f(g(x))\big) = f'(g(x))g'(x)$$

COMMON DERIVATIVES

$$\frac{d}{dx}(x) = 1$$

$$\frac{d}{dx}(\sin x) = \cos x$$

$$\frac{d}{dx}(\cos x) = -\sin x$$

$$\frac{d}{dx}(\tan x) = \sec^2 x$$

$$\frac{d}{dx}(\sec x) = \sec x \tan x$$

$$\frac{d}{dx}(\csc x) = -\csc x \cot x$$

$$\frac{d}{dx}(\cot x) = -\csc^2 x$$

$$\frac{d}{dx}(\sin^{-1} x) = \frac{1}{\sqrt{1-x^2}}$$

$$\frac{d}{dx}(\cos^{-1} x) = -\frac{1}{\sqrt{1-x^2}}$$

$$\frac{d}{dx}(\tan^{-1} x) = \frac{1}{1+x^2}$$

$$\frac{d}{dx}(a^x) = a^x \ln(a)$$

$$\frac{d}{dx}(e^x) = e^x$$

$$\frac{d}{dx}(\ln(x)) = \frac{1}{x}, x > 0$$

$$\frac{d}{dx}(\ln|x|) = \frac{1}{x}$$

$$\frac{d}{dx}(\log_a(x)) = \frac{1}{x\ln(a)}$$

CHAIN RULE AND OTHER EXAMPLES

$$\frac{d}{dx}([f(x)]^n) = n[f(x)]^{n-1}f'(x)$$

$$\frac{d}{dx}(e^{f(x)}) = f'(x)e^{f(x)}$$

$$\frac{d}{dx}(\ln[f(x)]) = \frac{f'(x)}{f(x)}$$

$$\frac{d}{dx}(\sin[f(x)]) = f'(x)\cos[f(x)]$$

$$\frac{d}{dx}(\cos[f(x)]) = -f'(x)\sin[f(x)]$$

$$\frac{d}{dx}(\tan[f(x)]) = f'(x)\sec^2[f(x)]$$

$$\frac{d}{dx}(\sec[f(x)]) = f'(x)\sec[f(x)]\tan[f(x)]$$

$$\frac{d}{dx}(\tan^{-1}[f(x)]) = \frac{f'(x)}{1+[f(x)]^2}$$

$$\frac{d}{dx}\left(f(x)^{g(x)}\right) = f(x)^{g(x)}\left(\frac{g(x)f'(x)}{f(x)} + \ln(f(x))\,g'(x)\right)$$

LIMIT EVALUATION METHOD – FACTOR AND CANCEL

$$\lim_{x\to -3}\frac{x^2-x-12}{x^2+3x} = \lim_{x\to -3}\frac{(x+3)(x-4)}{x(x+3)} = \lim_{x\to -3}\frac{(x-4)}{x} = \frac{7}{3}$$

L'HOPITAL'S RULE

$$\text{If } \lim_{x\to a}\frac{f(x)}{g(x)} = \frac{0}{0} \text{ or } \frac{\pm\infty}{\pm\infty} \text{ then } \lim_{x\to a}\frac{f(x)}{g(x)} = \lim_{x\to a}\frac{f'(x)}{g'(x)}$$

PROPERTIES OF LIMITS

These properties require that the limit of $f(x)$ and $g(x)$ exist

$$\lim_{x\to a}[cf(x)] = c\lim_{x\to a} f(x)$$

$$\lim_{x\to a}[f(x) \pm g(x)] = \lim_{x\to a} f(x) \pm \lim_{x\to a} g(x)$$

$$\lim_{x\to a}[f(x)g(x)] = \lim_{x\to a} f(x)\lim_{x\to a} g(x)$$

$$\lim_{x\to a}\left[\frac{f(x)}{g(x)}\right] = \frac{\lim_{x\to a} f(x)}{\lim_{x\to a} g(x)} \quad \text{if } \lim_{x\to a} g(x) \neq 0$$

$$\lim_{x\to a}[f(x)]^n = \left[\lim_{x\to a} f(x)\right]^n$$

LIMIT EVALUATIONS AT +-∞

$$\lim_{x\to\infty} e^x = \infty \text{ and } \lim_{x\to -\infty} e^x = 0$$

$$\lim_{x\to\infty} \ln(x) = \infty \text{ and } \lim_{x\to 0^+} \ln(x) = -\infty$$

$$\text{If } r > 0 \text{ then } \lim_{x\to\infty}\frac{c}{x^r} = 0$$

$$\text{If } r > 0 \text{ \& } x^r \text{ is real for } x < 0 \text{ then } \lim_{x\to -\infty}\frac{c}{x^r} = 0$$

$$\lim_{x\to\pm\infty} x^r = \infty \text{ for even } r$$

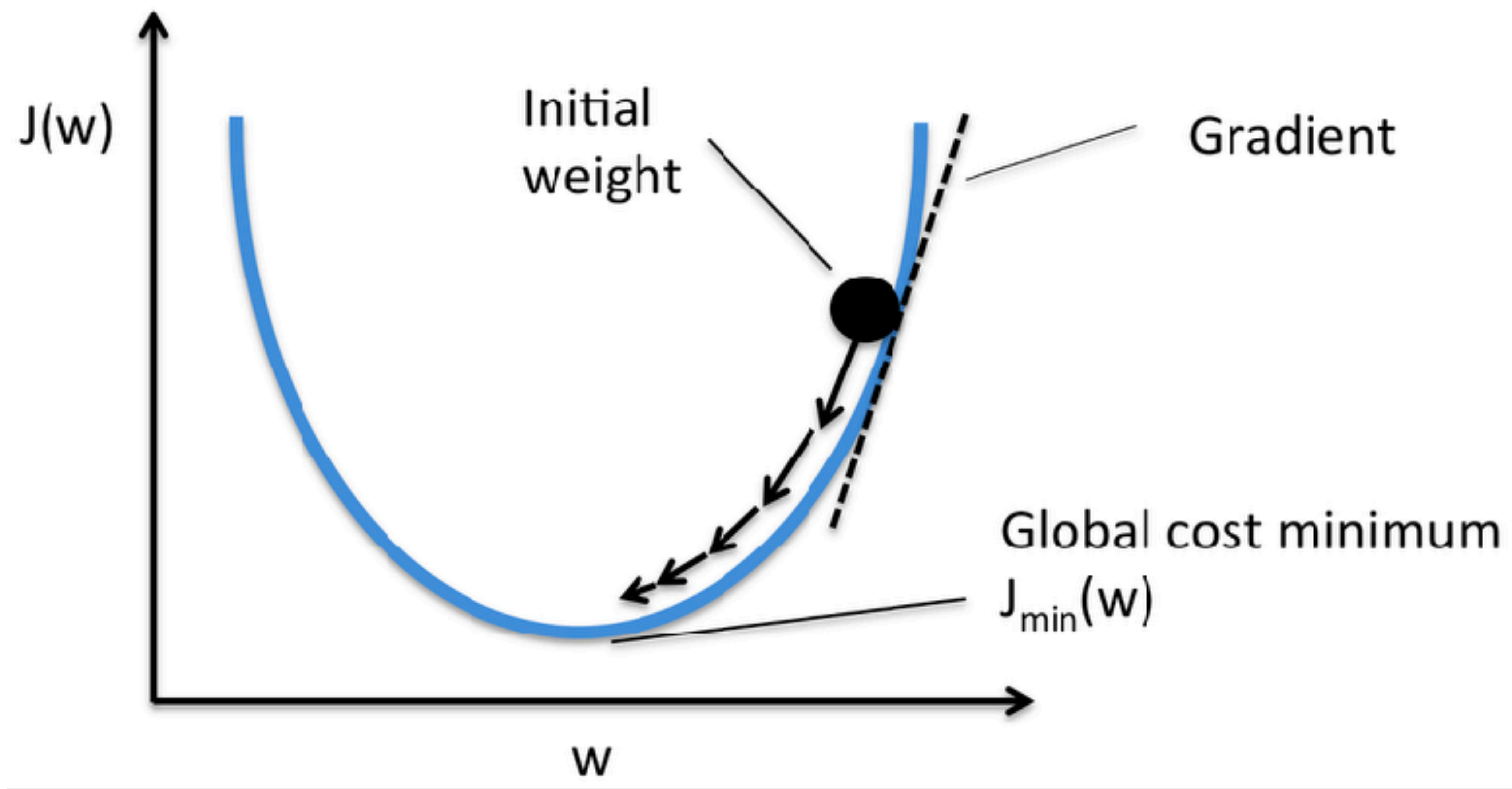
$$\lim_{x\to\infty} x^r = \infty \text{ \& } \lim_{x\to -\infty} x^r = -\infty \text{ for odd } r$$

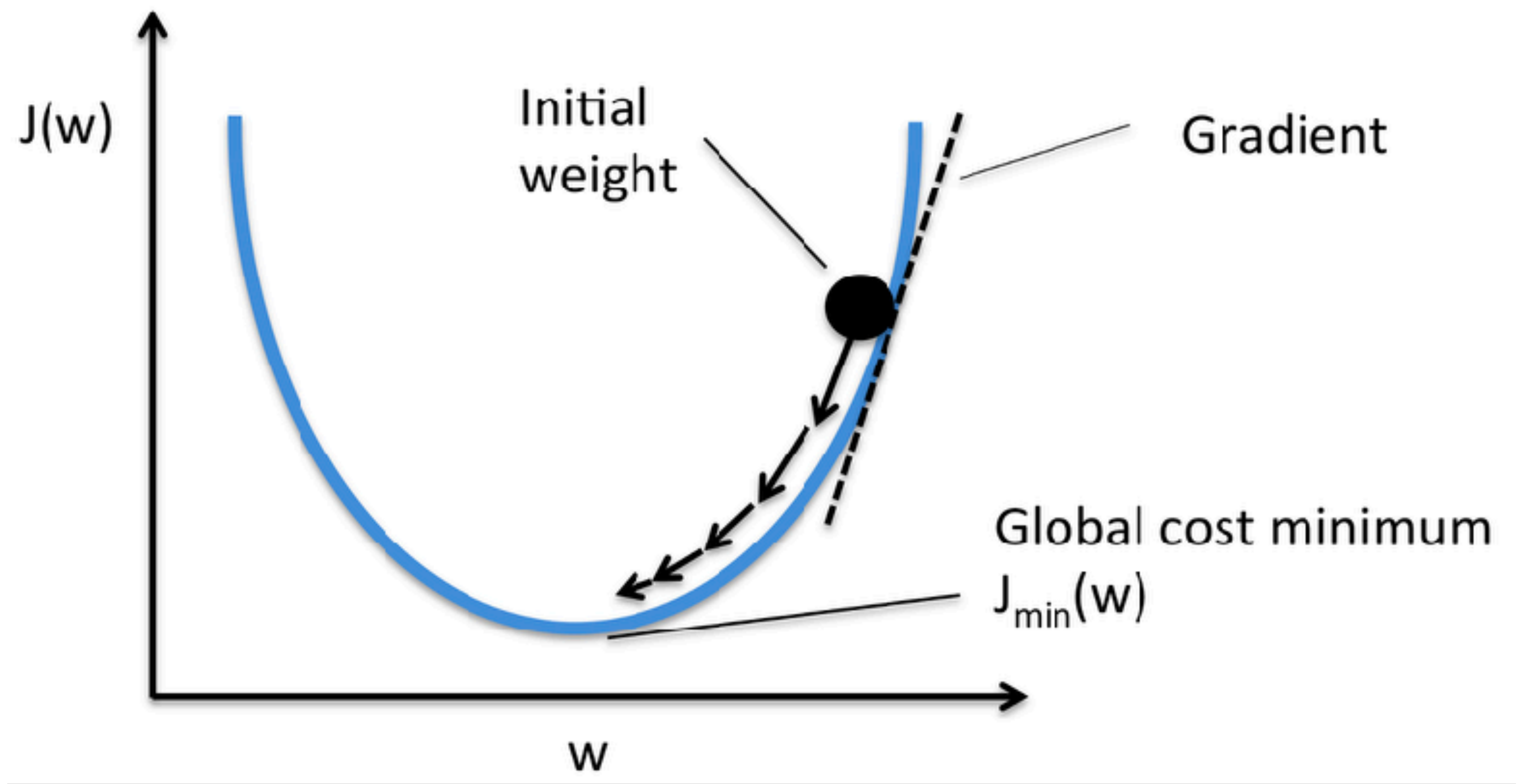
Derivative of Sigmoid function

$$y = \frac{1}{1+e^{-x}}$$

$$\frac{dy}{dx} = -\frac{1}{(1+e^{-x})^2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2}$$

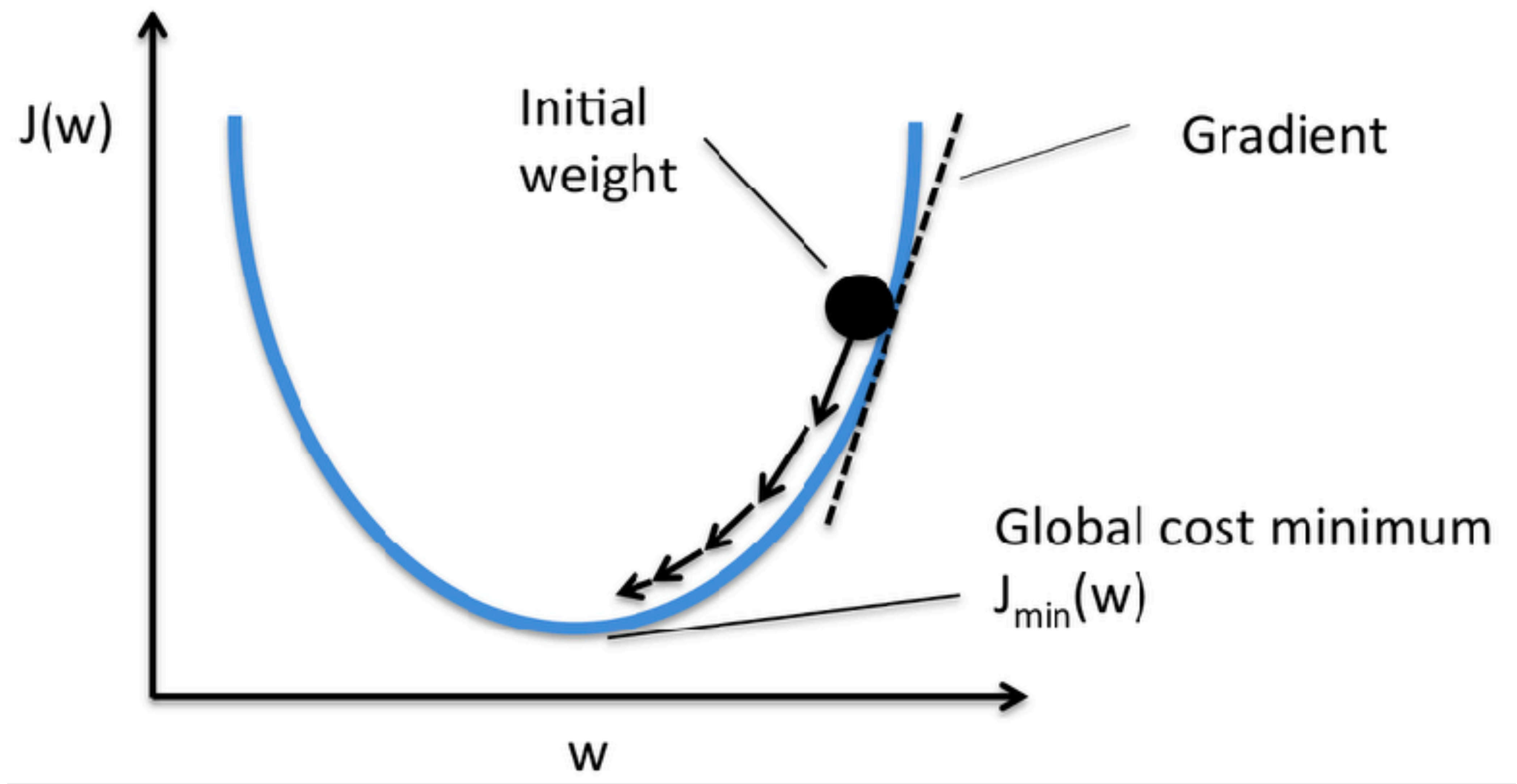
$$= \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right) = y(1-y)$$


J(w)
Initial
weight
Gradient
Global cost minimum
$J_{min}(w)$
w



Repetir até encontrar inclinação == 0:

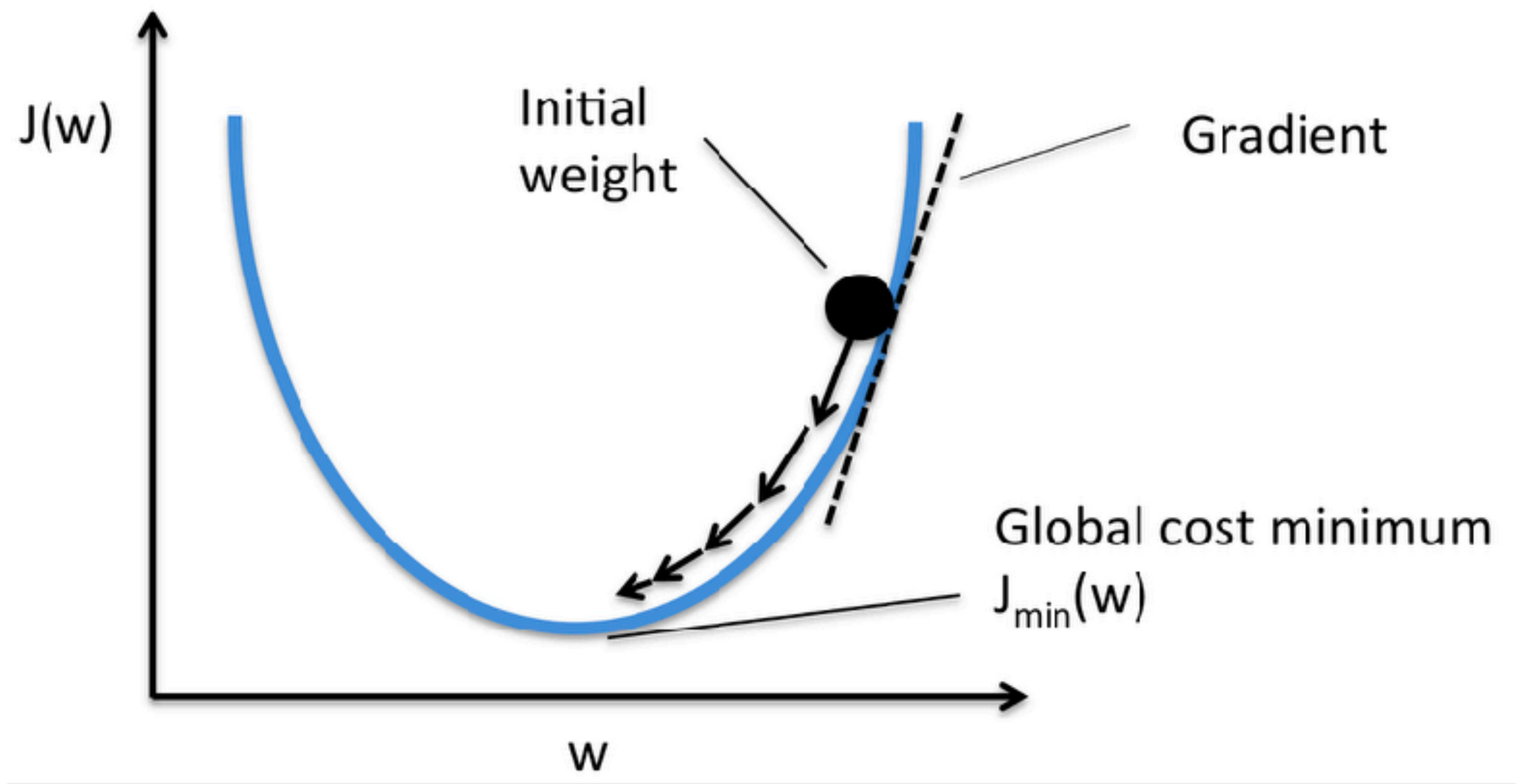
$$w := w - \alpha \frac{dJ(w)}{dw}$$



Repetir até encontrar inclinação == 0:

$$w := w - \alpha \boxed{\frac{dJ(w)}{dw}}$$

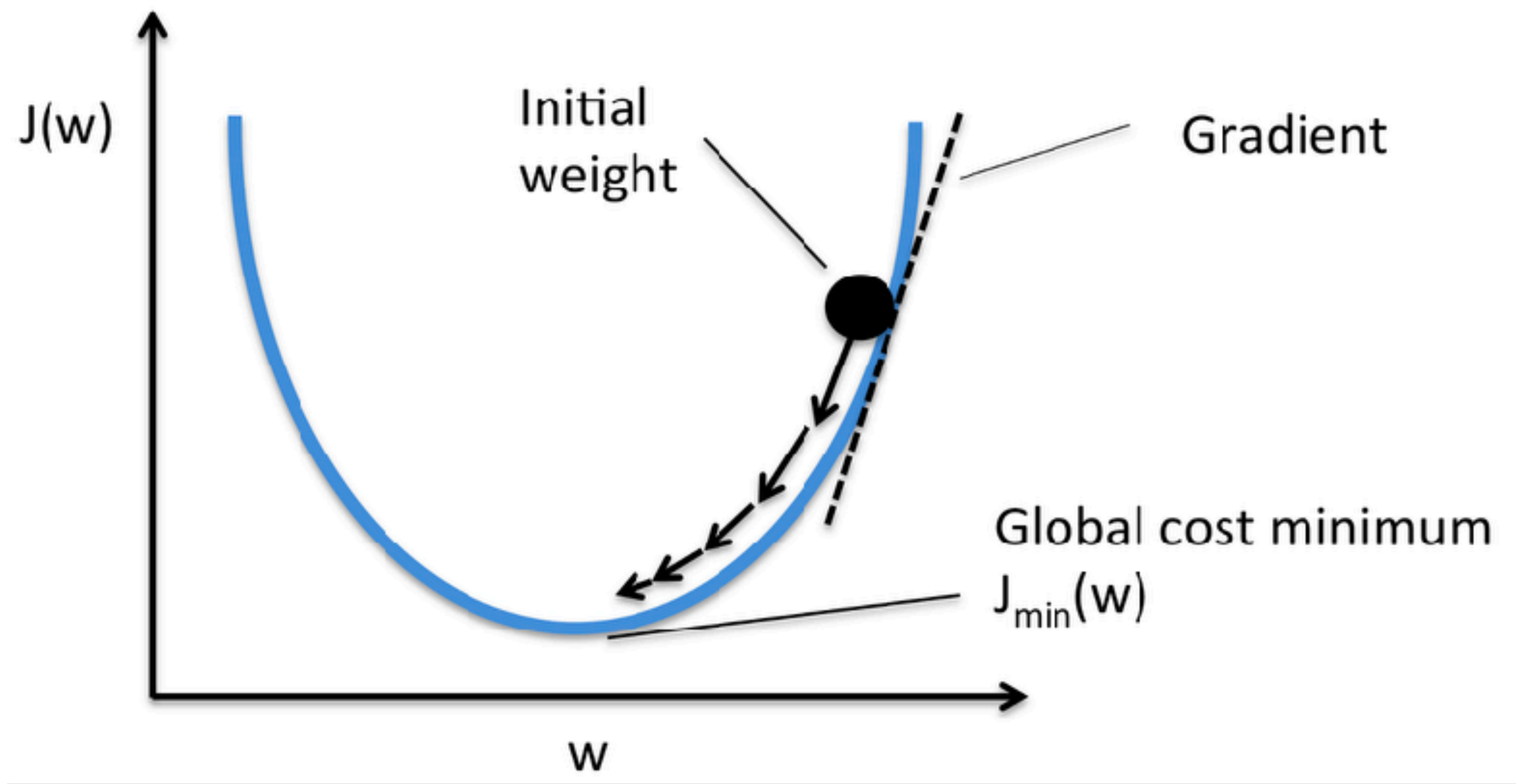
Direção

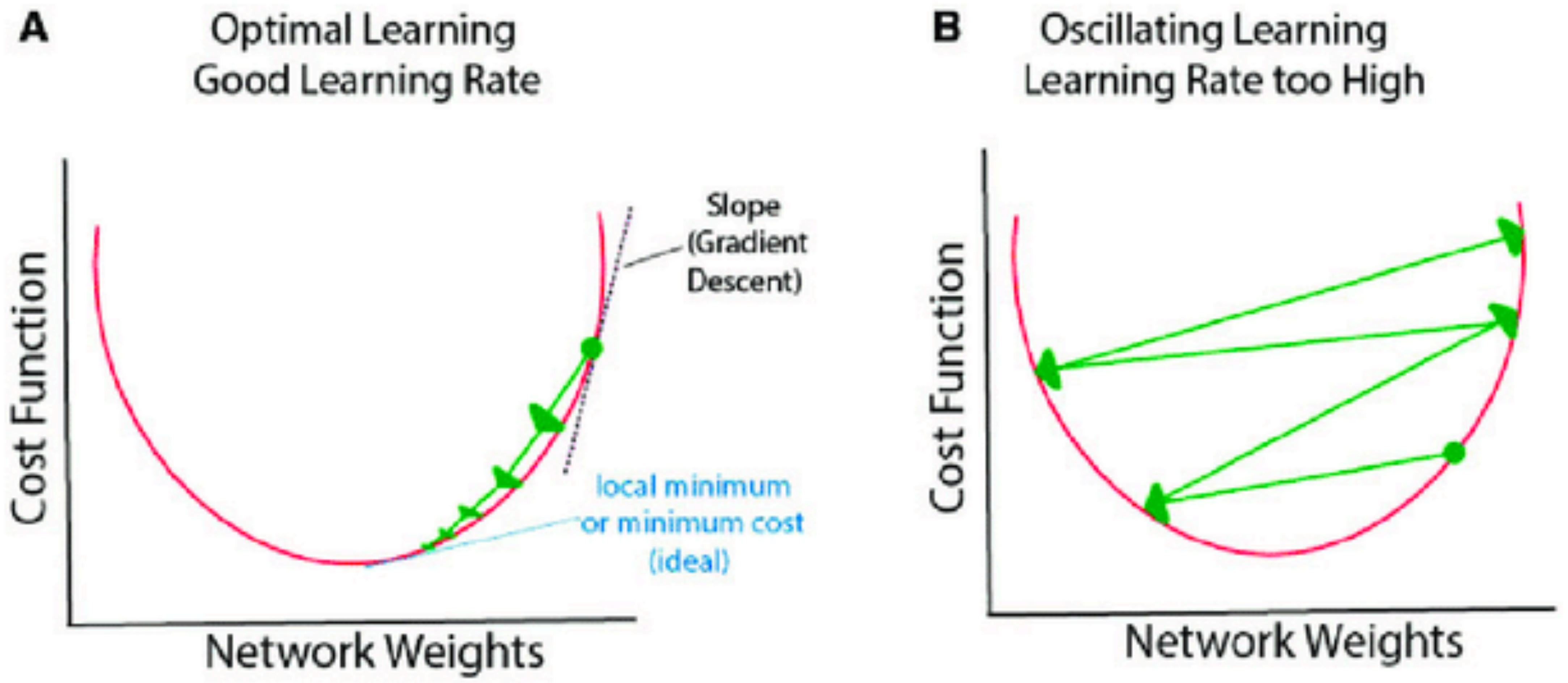


Repetir até encontrar inclinação == 0:

$$w := w - \alpha \frac{dJ(w)}{dw}$$

Passo Direção

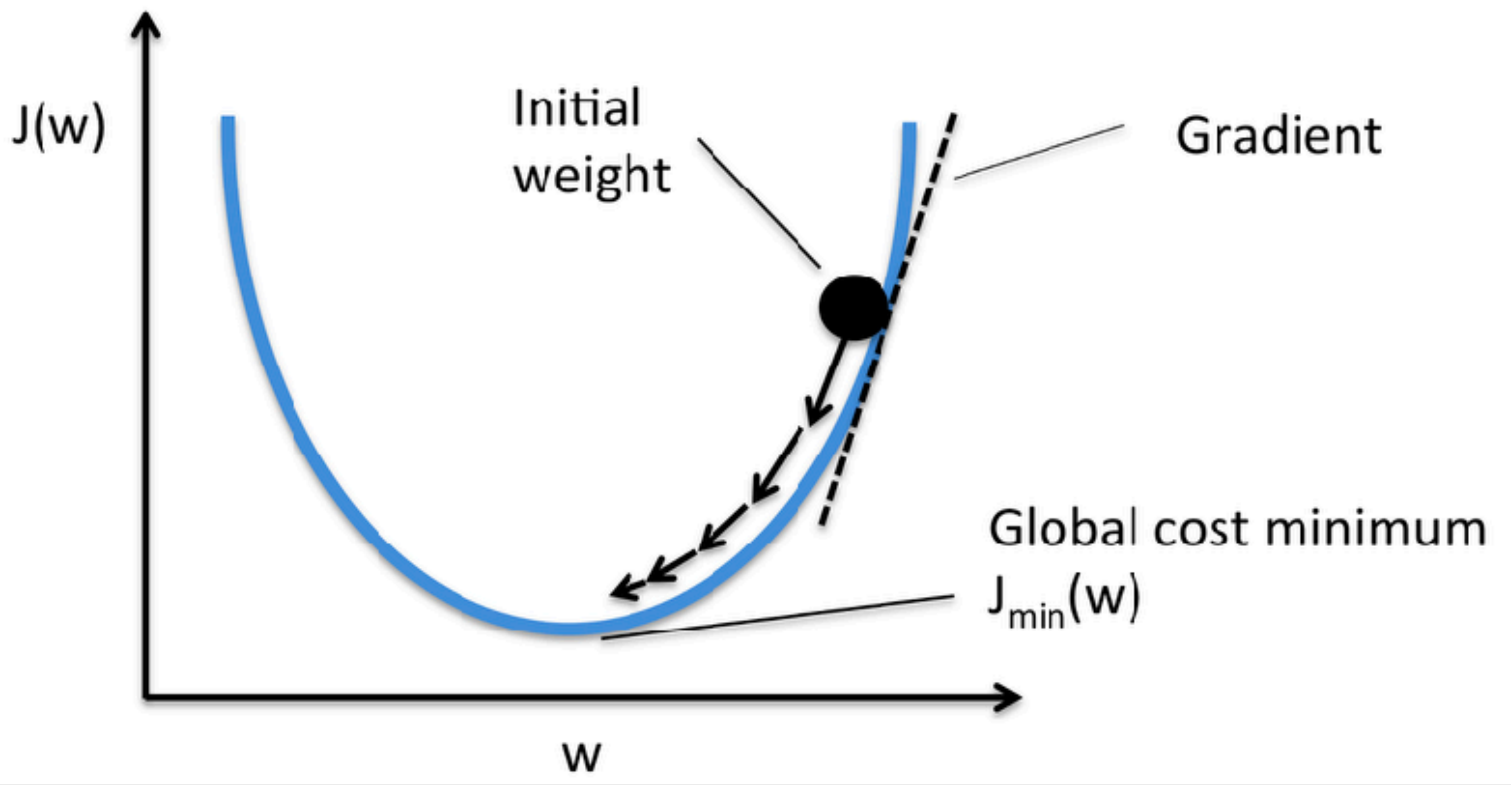


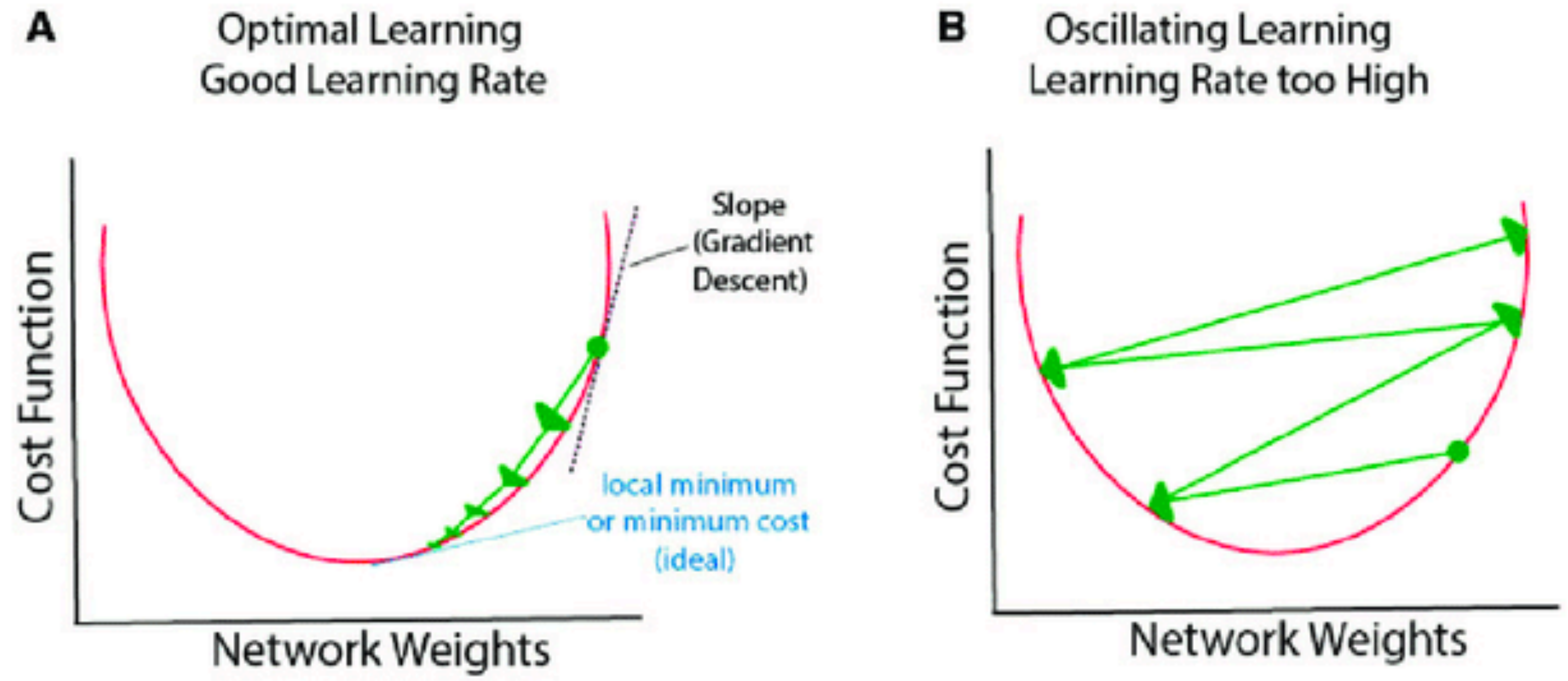


Repetir até encontrar inclinação == 0:

$$w := w - \alpha \frac{dJ(w)}{dw}$$

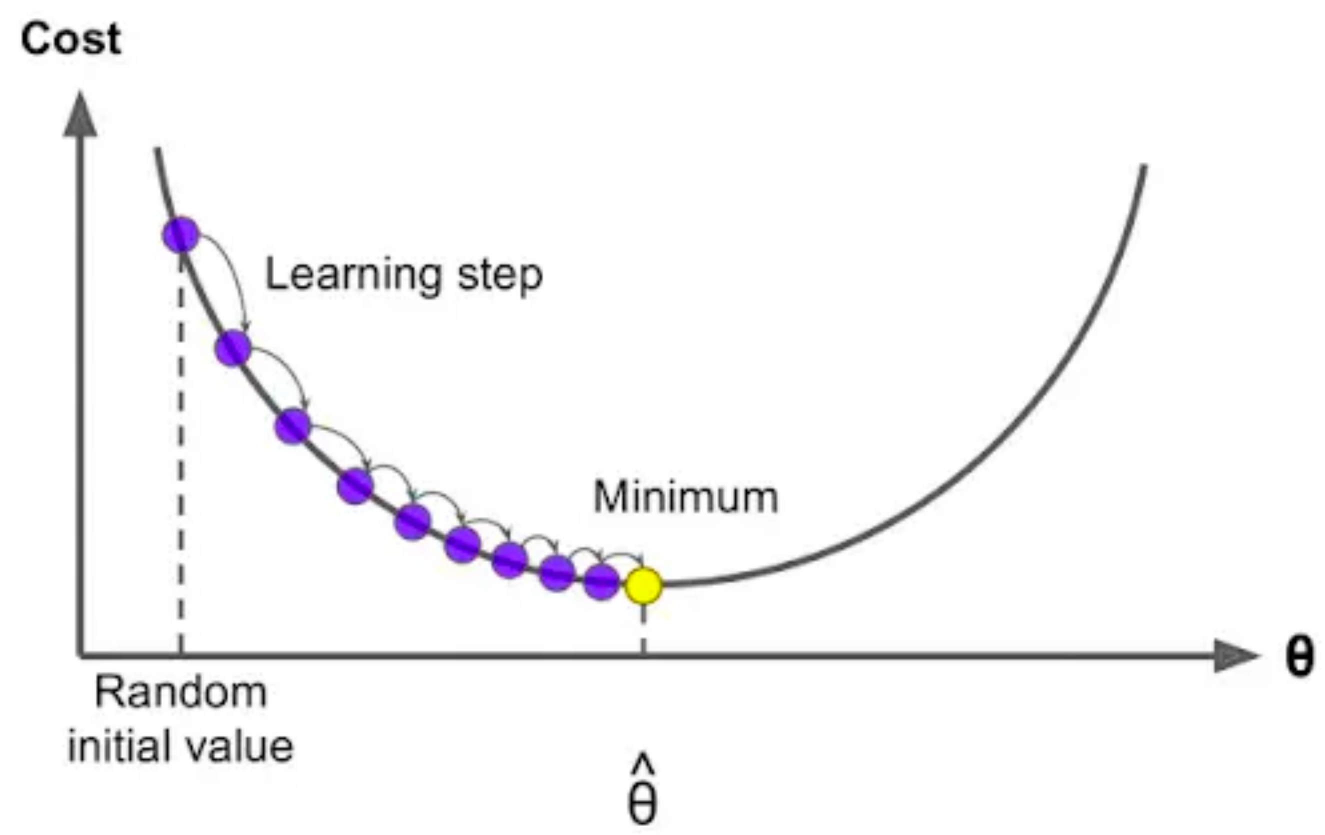
Passo **Direção**





Repetir até encontrar inclinação == 0:

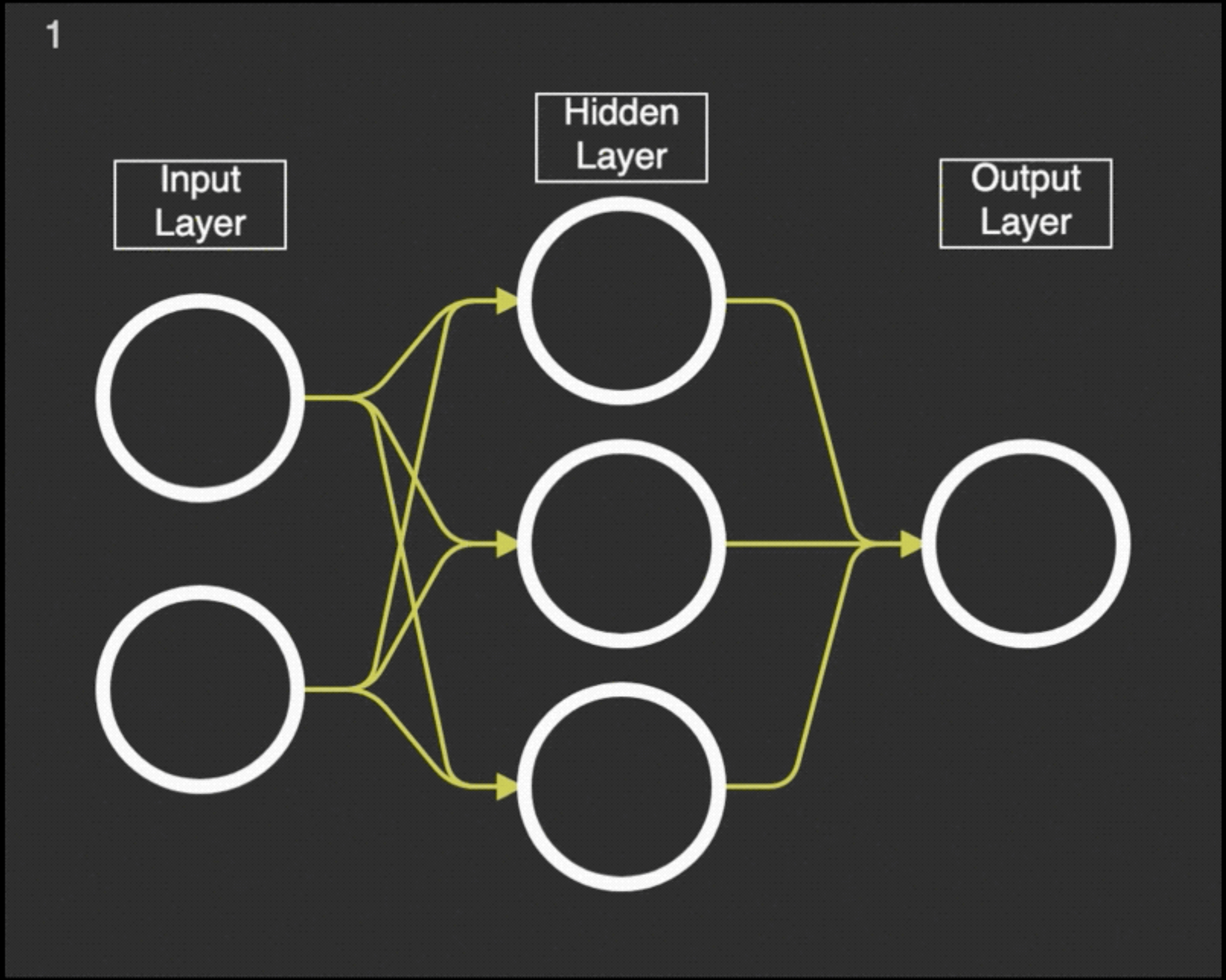
$$w := w - \underbrace{\alpha}_{\text{Passo}} \underbrace{\frac{dJ(w)}{dw}}_{\text{Direção}}$$



Minimizar J(w, b):

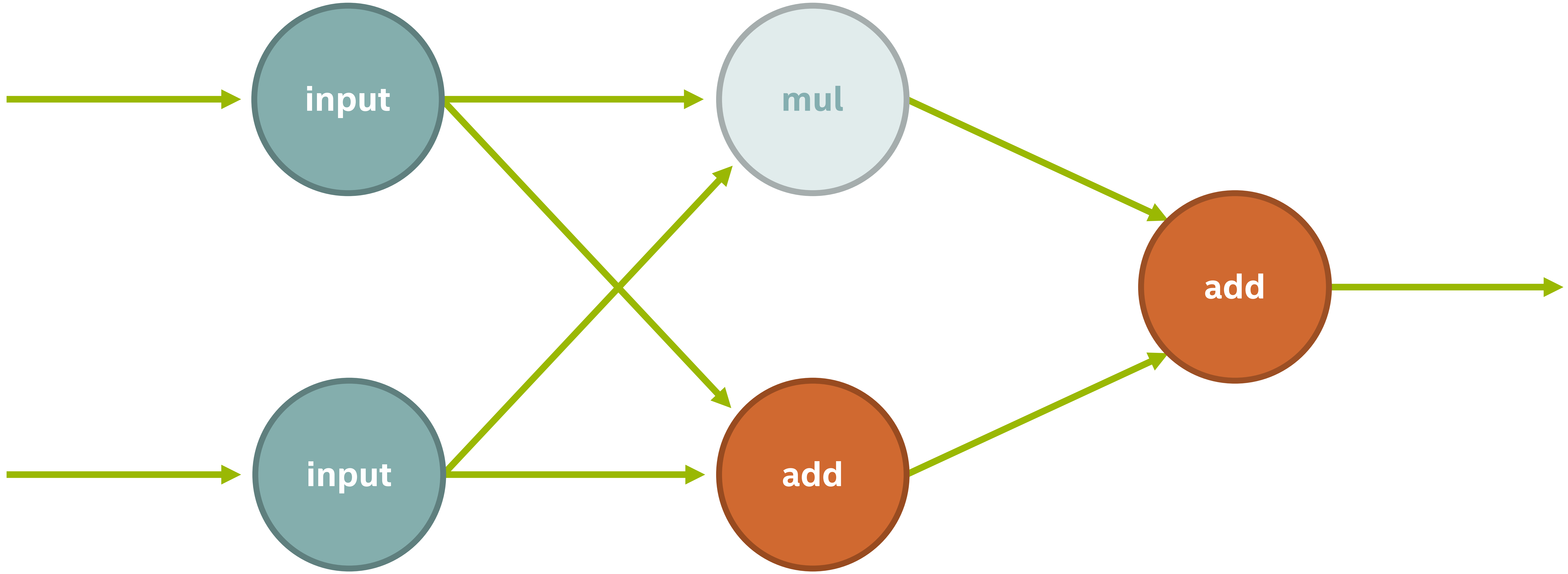
$$w := w - \alpha \frac{\partial J(w, b)}{\partial w}$$

$$b := b - \alpha \frac{\partial J(w, b)}{\partial b}$$

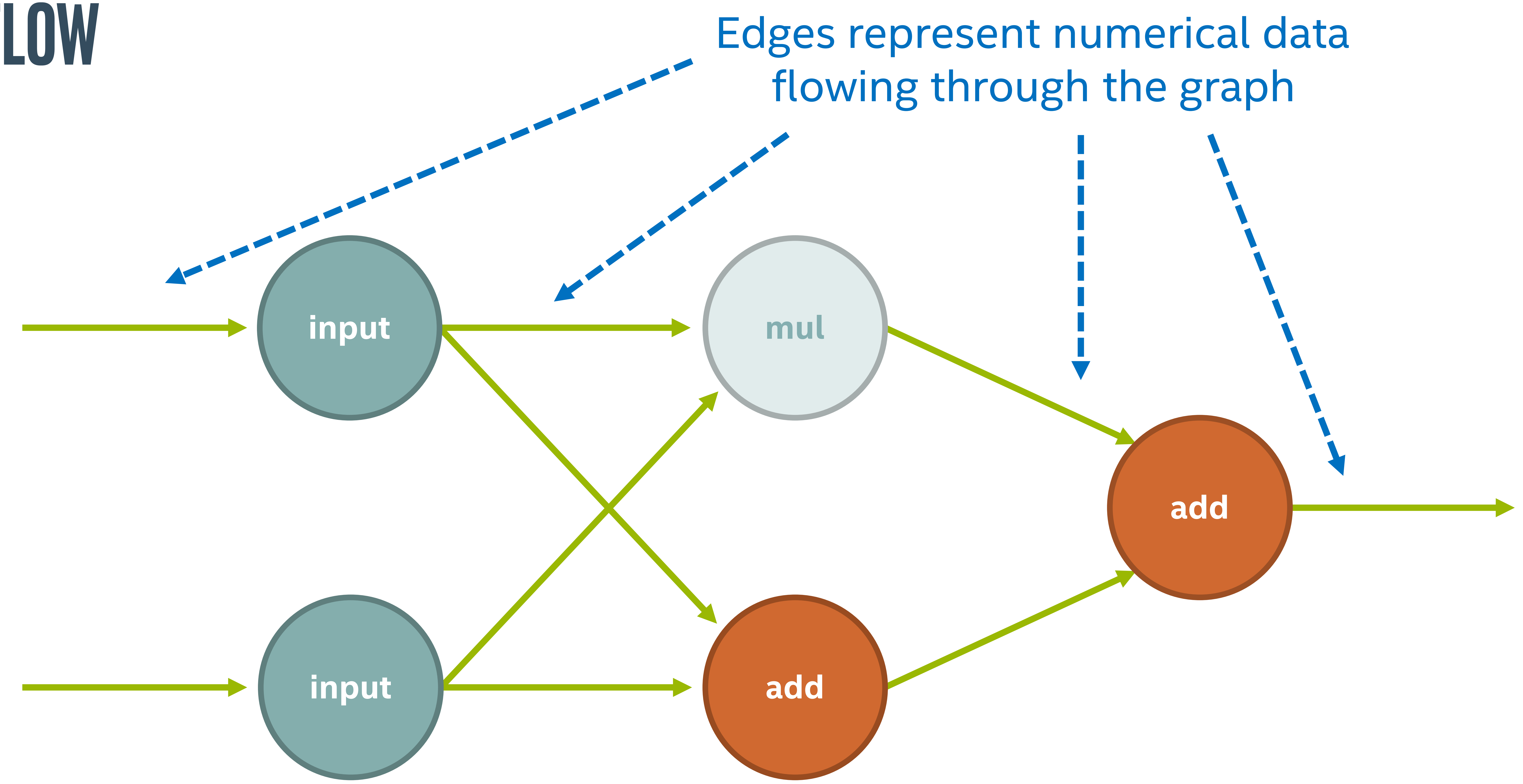

1
Input Layer
Hidden Layer
Output Layer

COMPUTATION GRAPHS

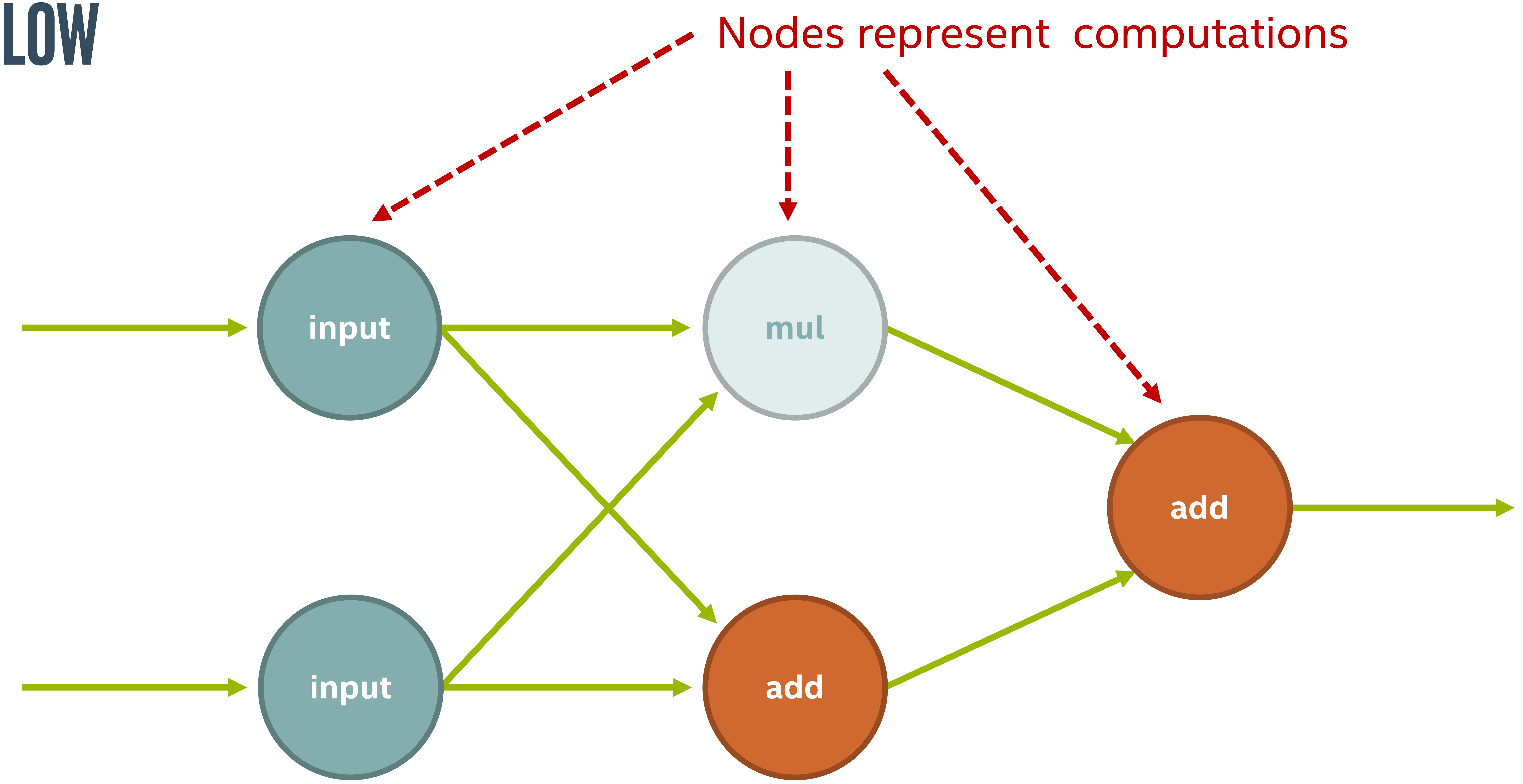
DATA FLOW



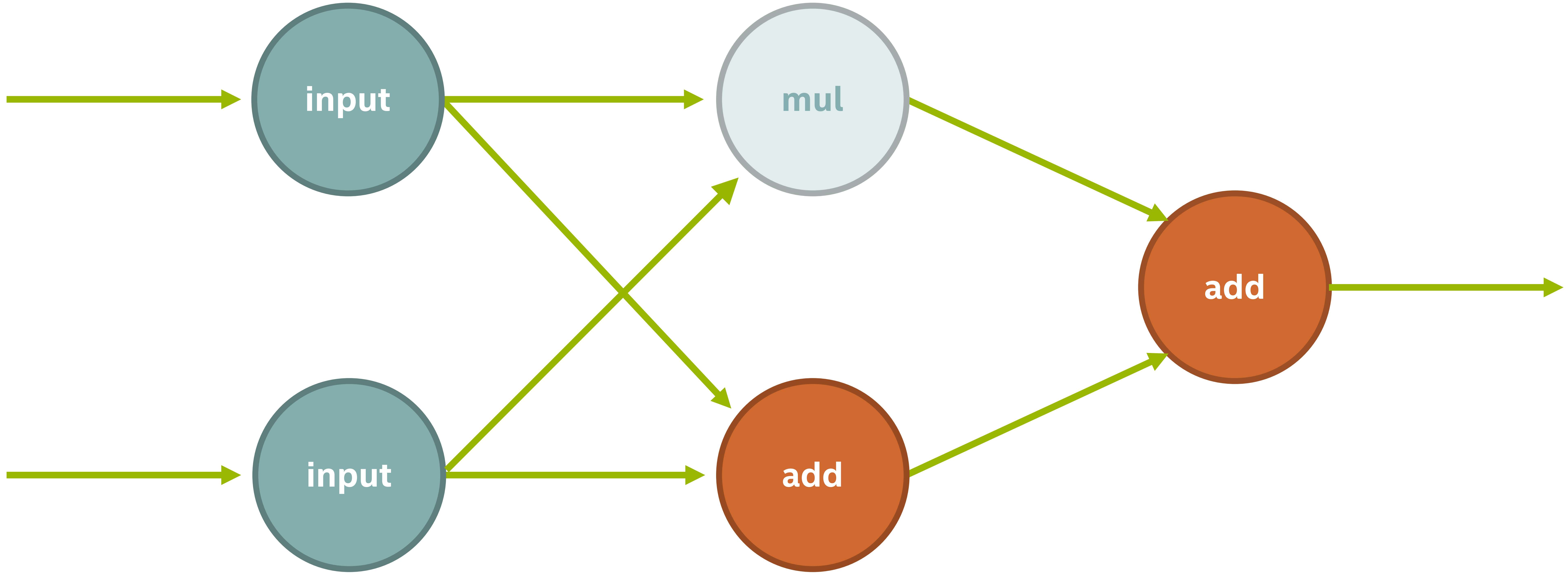
DATA FLOW


Edges represent numerical data flowing through the graph
input
input
mul
add
add

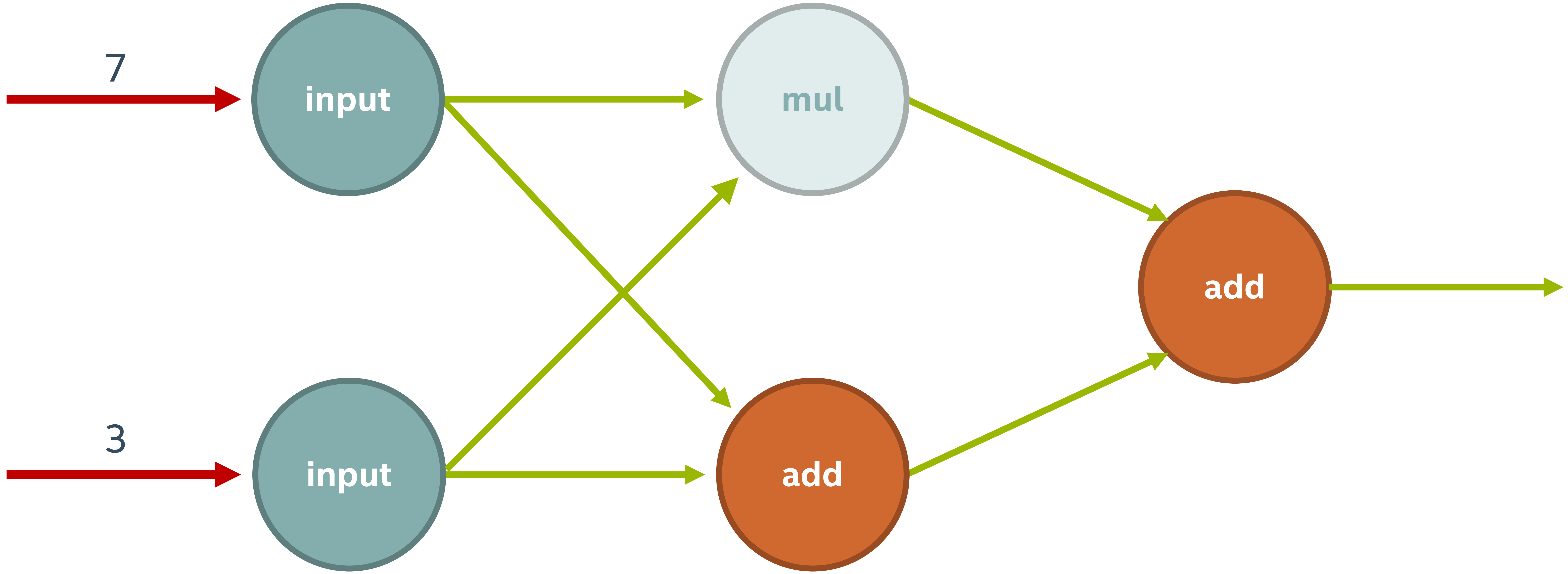
DATA FLOW



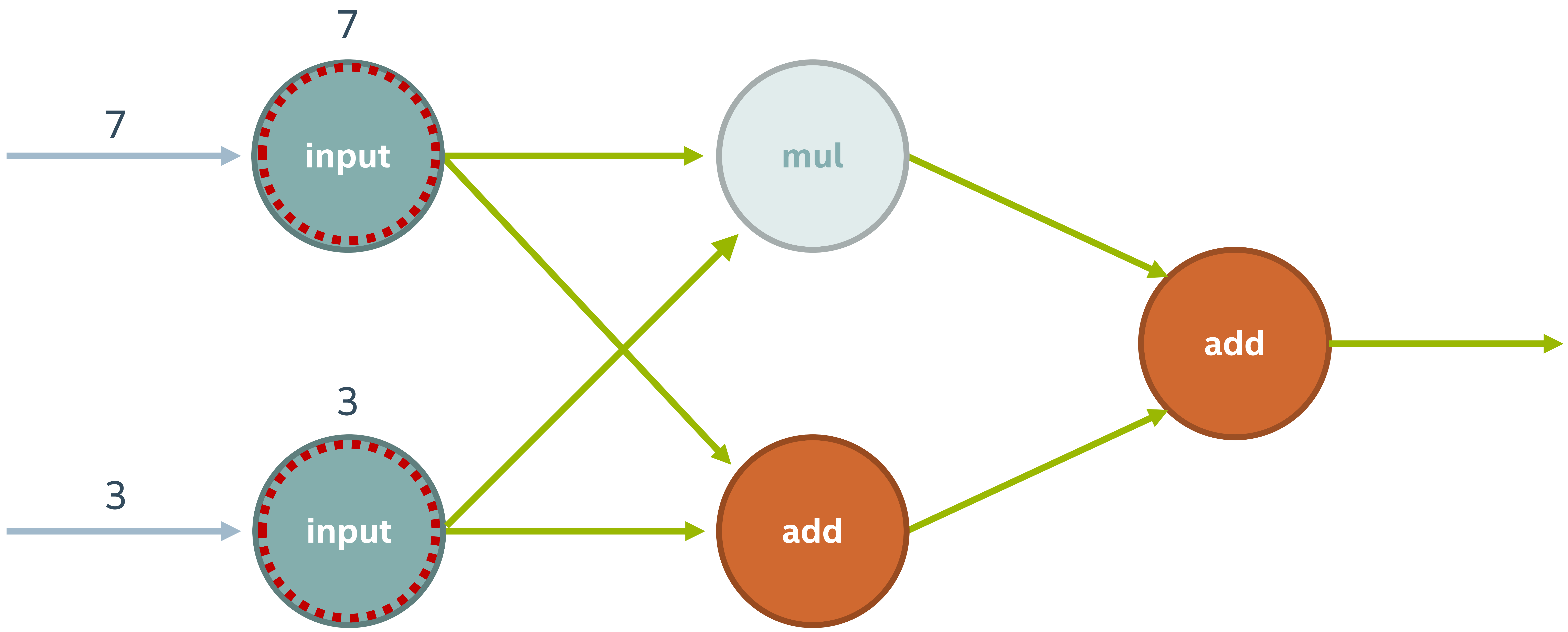
DATA FLOW



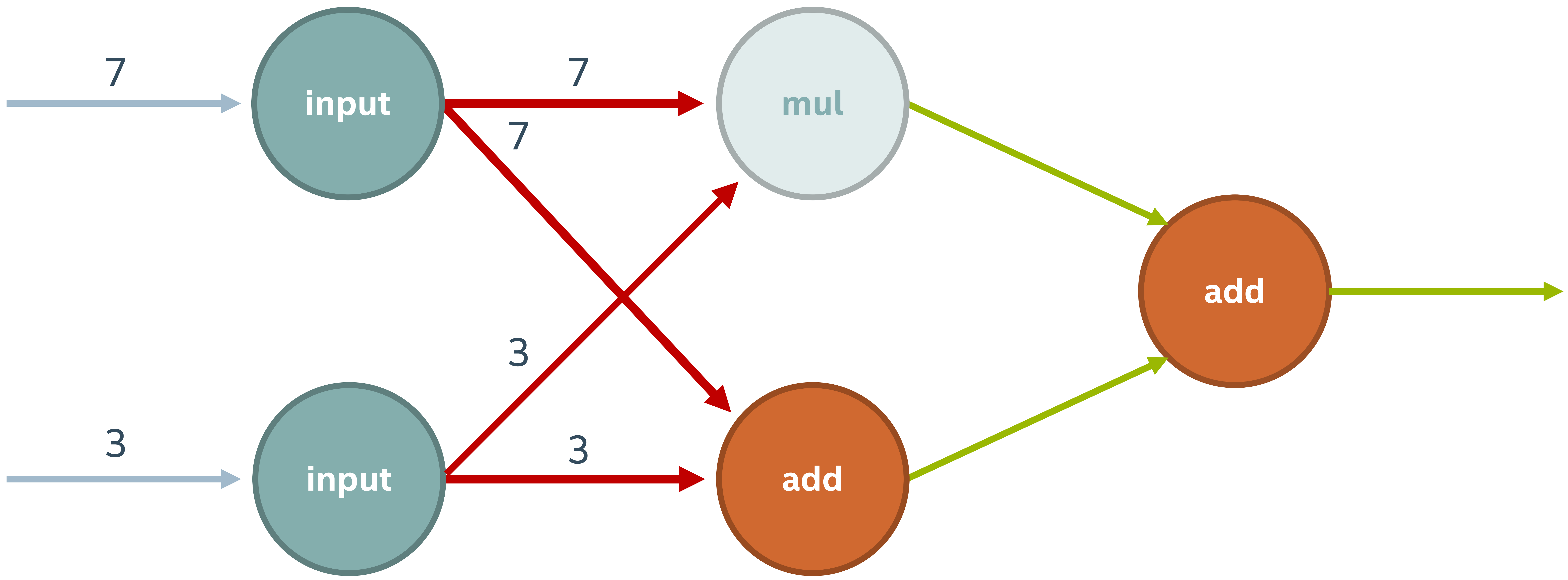
DATA FLOW



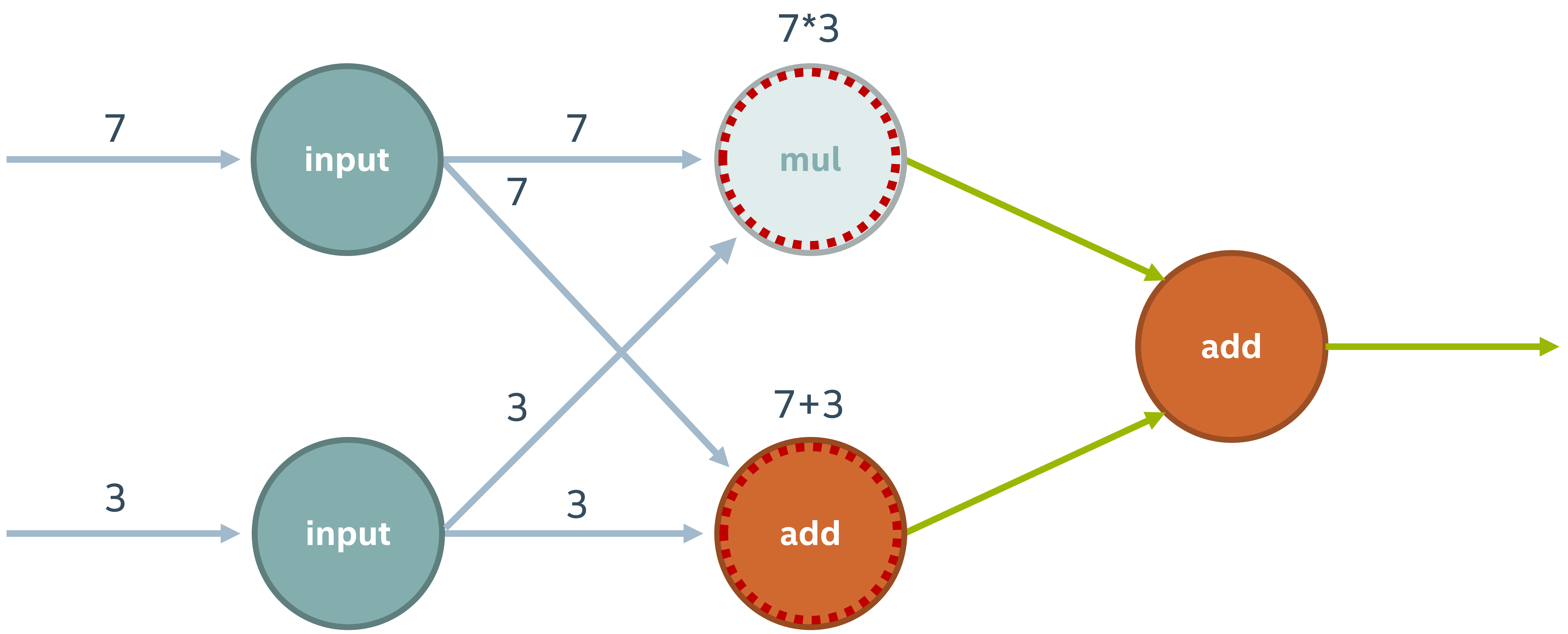
DATA FLOW



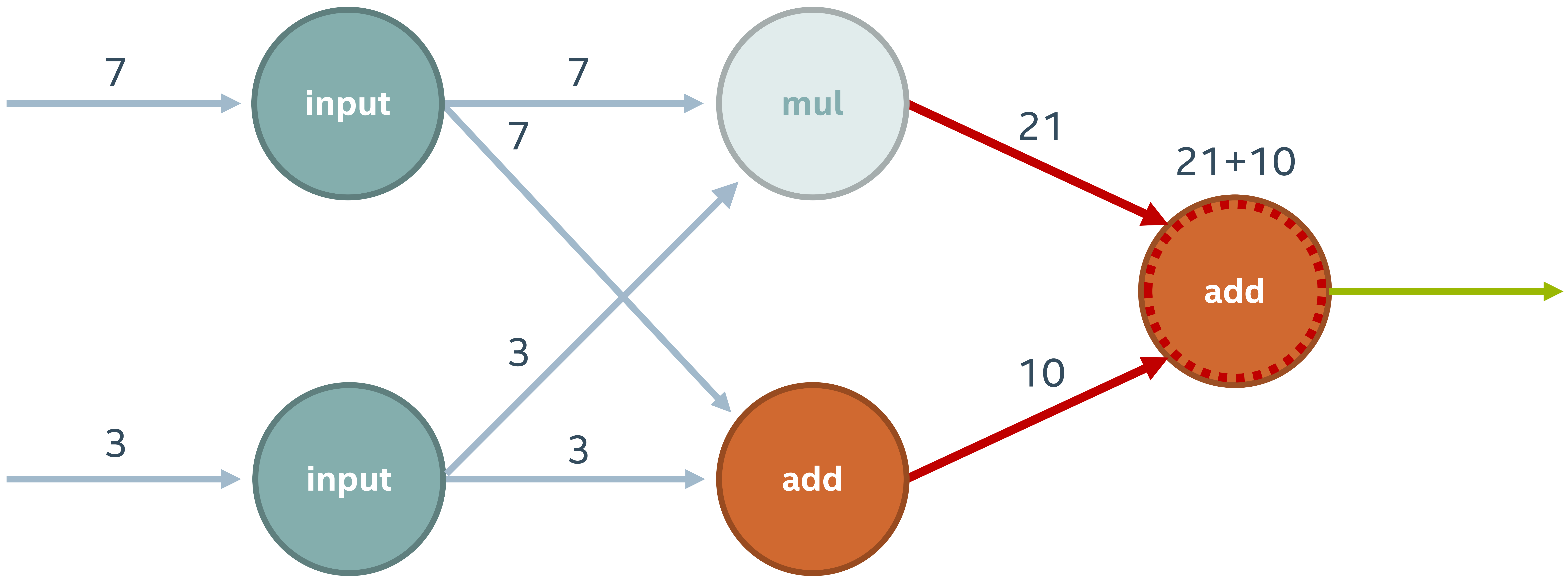
DATA FLOW



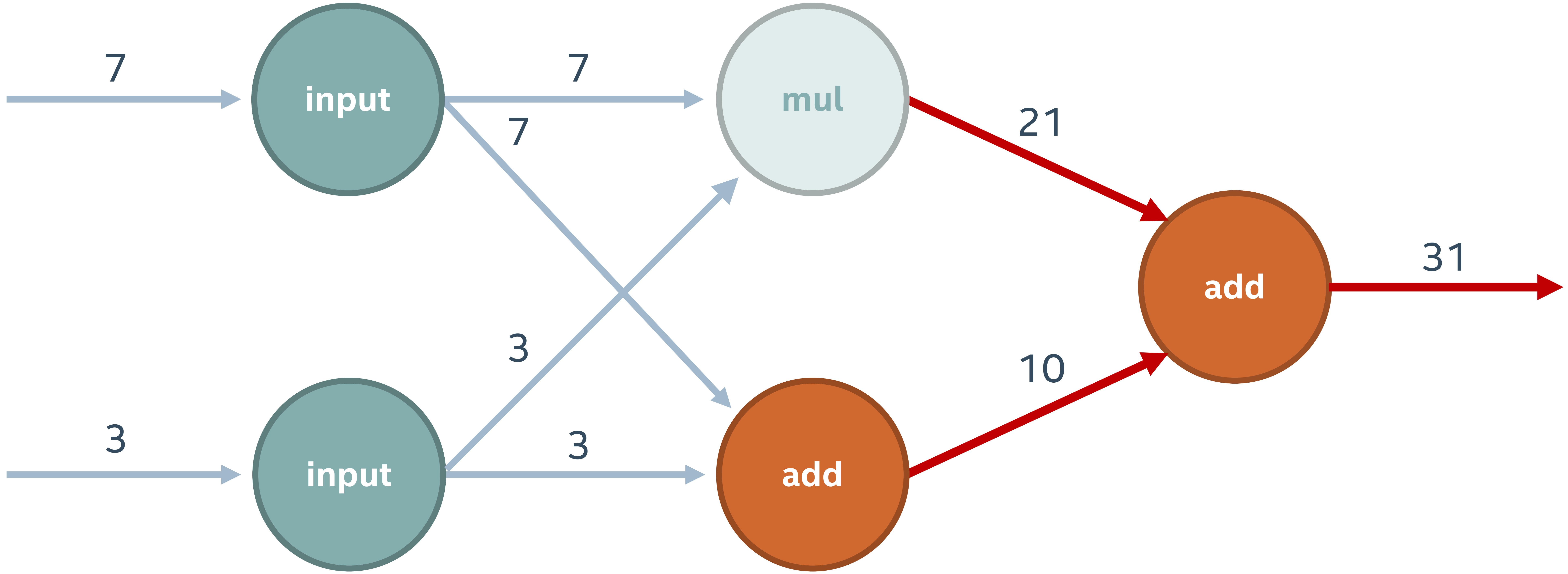
DATA FLOW

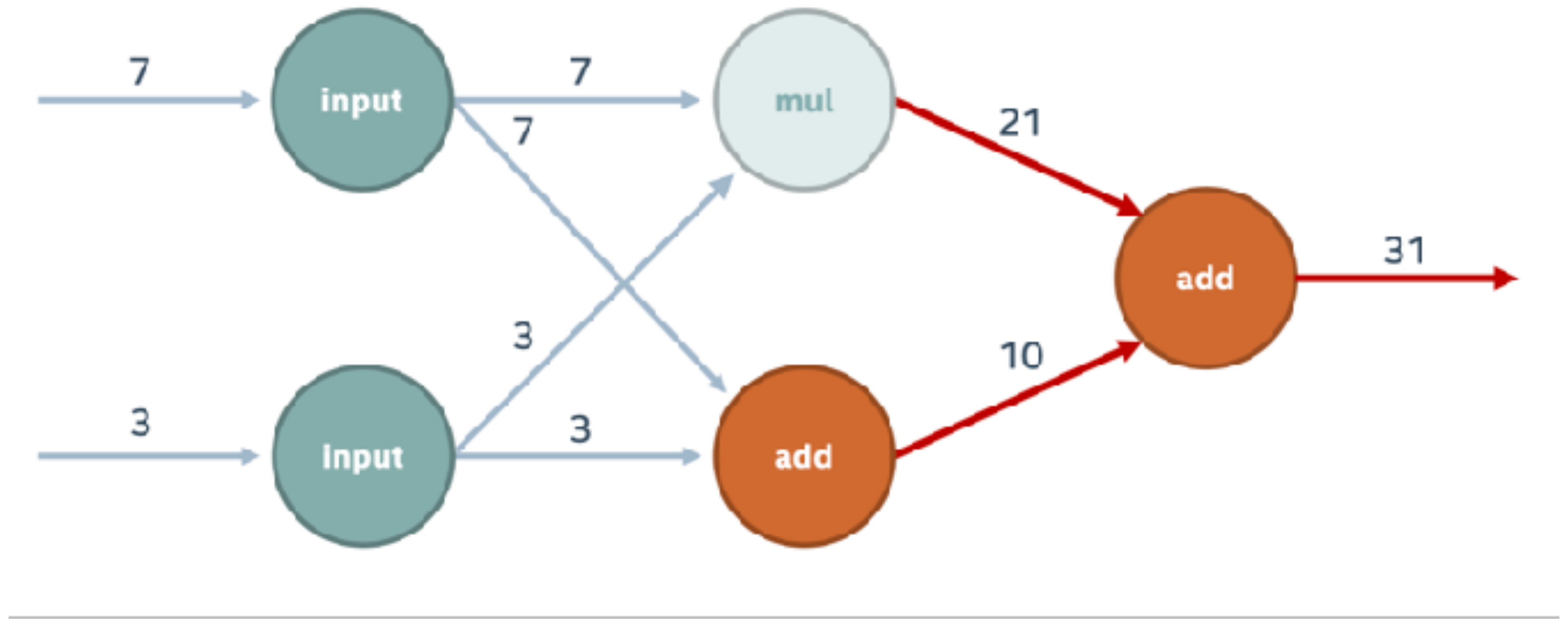


DATA FLOW

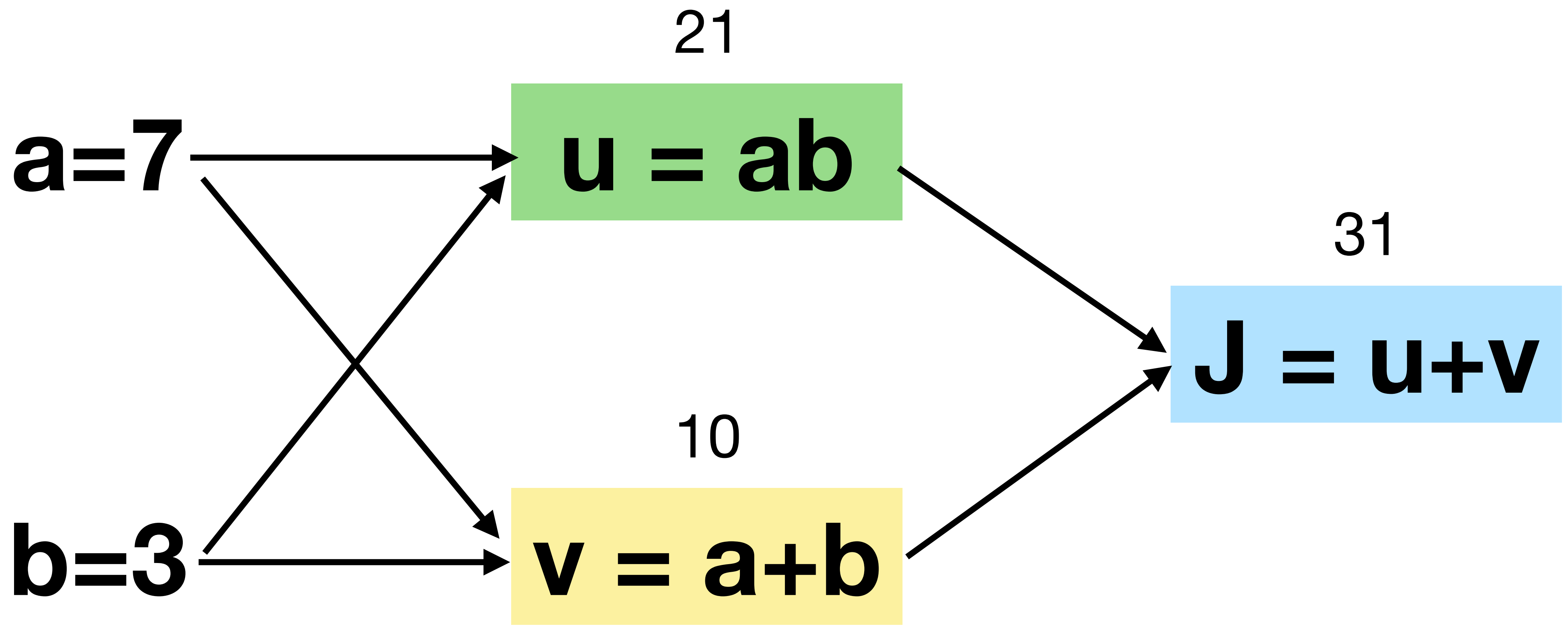


DATA FLOW



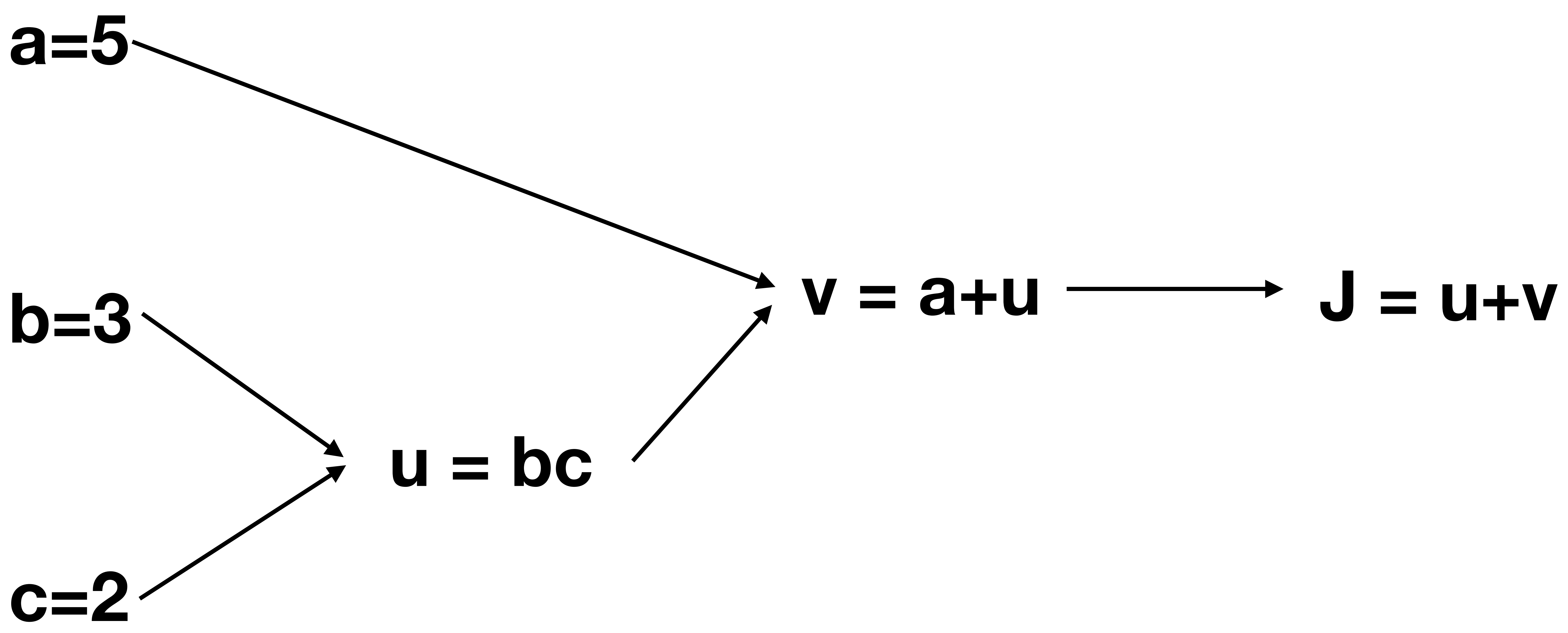


$$J(a,b) = \underbrace{\underbrace{ab}_{\mathbf{u}} + \underbrace{(a+b)}_{\mathbf{v}}}_{\mathbf{J}}$$



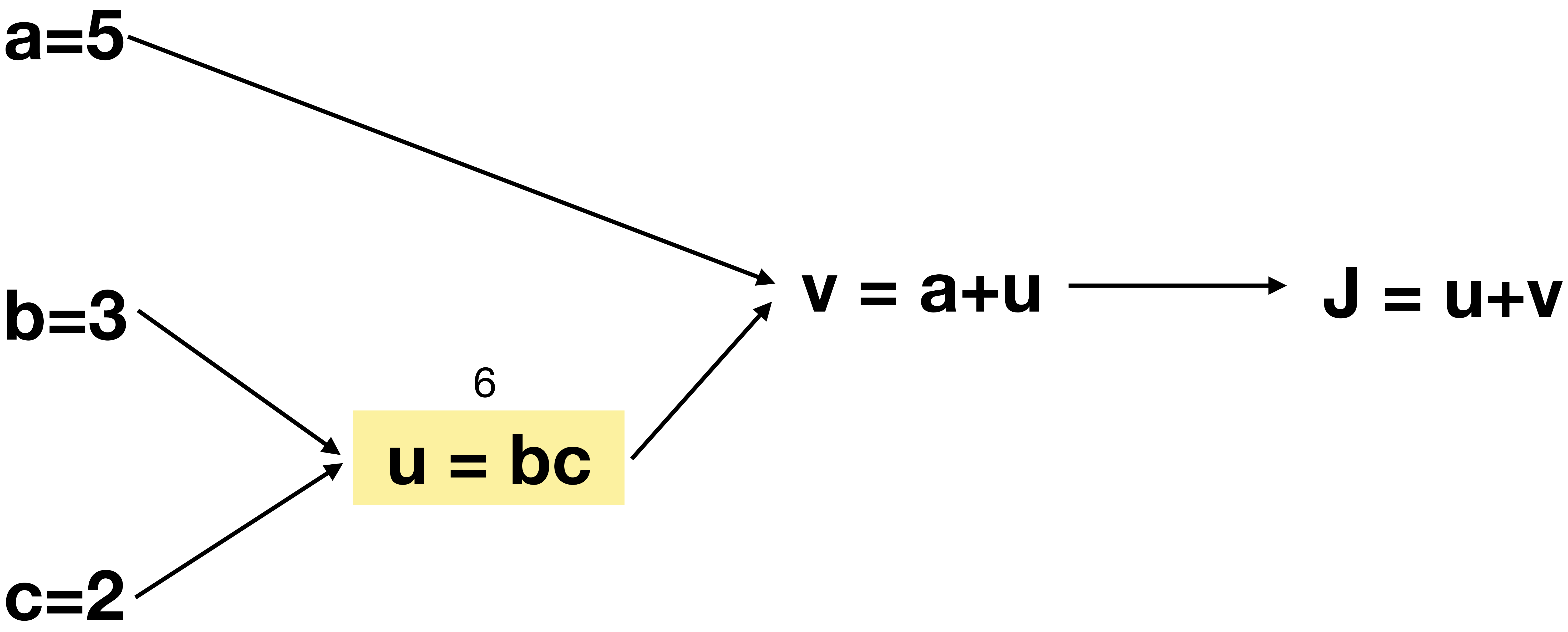
$$J(a,b,c) = 3(a + bc)$$

$$J(a,b,c) = 3(5 + 3 \times 2)$$



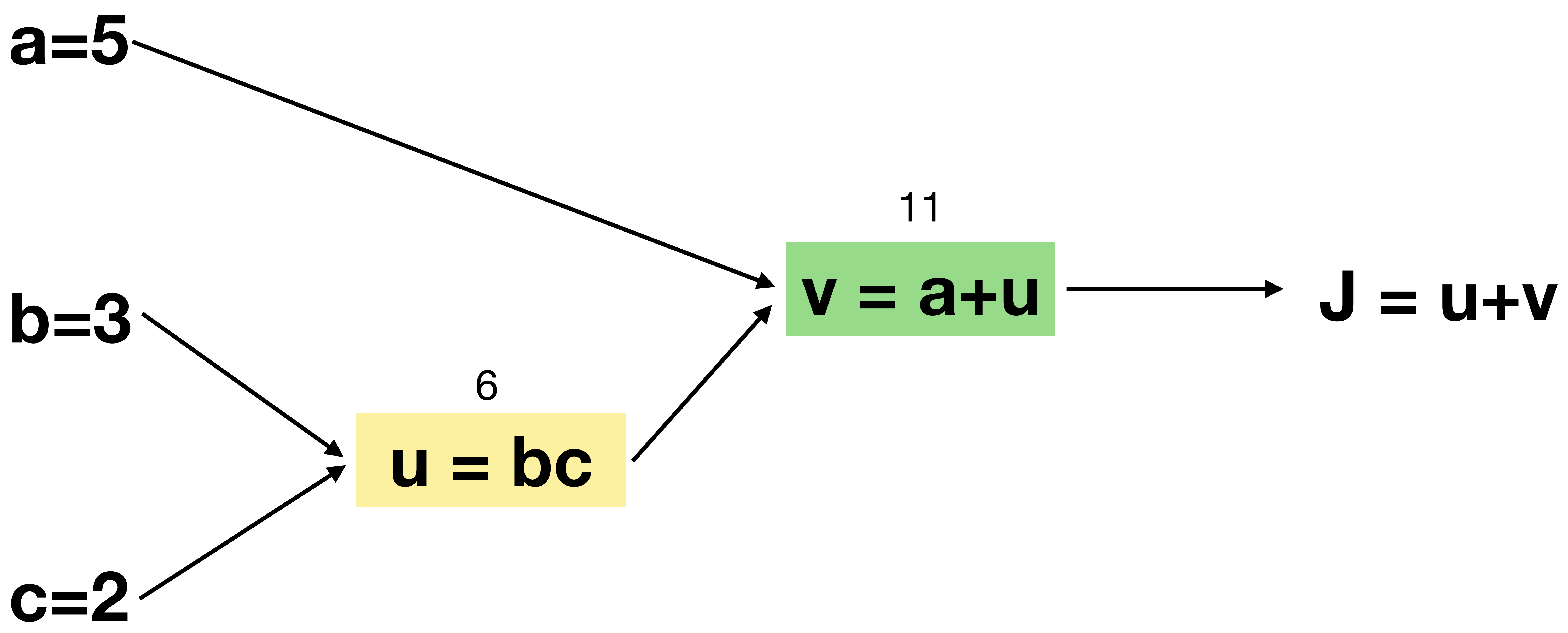
$$J(a,b,c) = 3(a + \underbrace{bc}_{u})$$

$$J(a,b,c) = 3(5 + 3 \times 2)$$



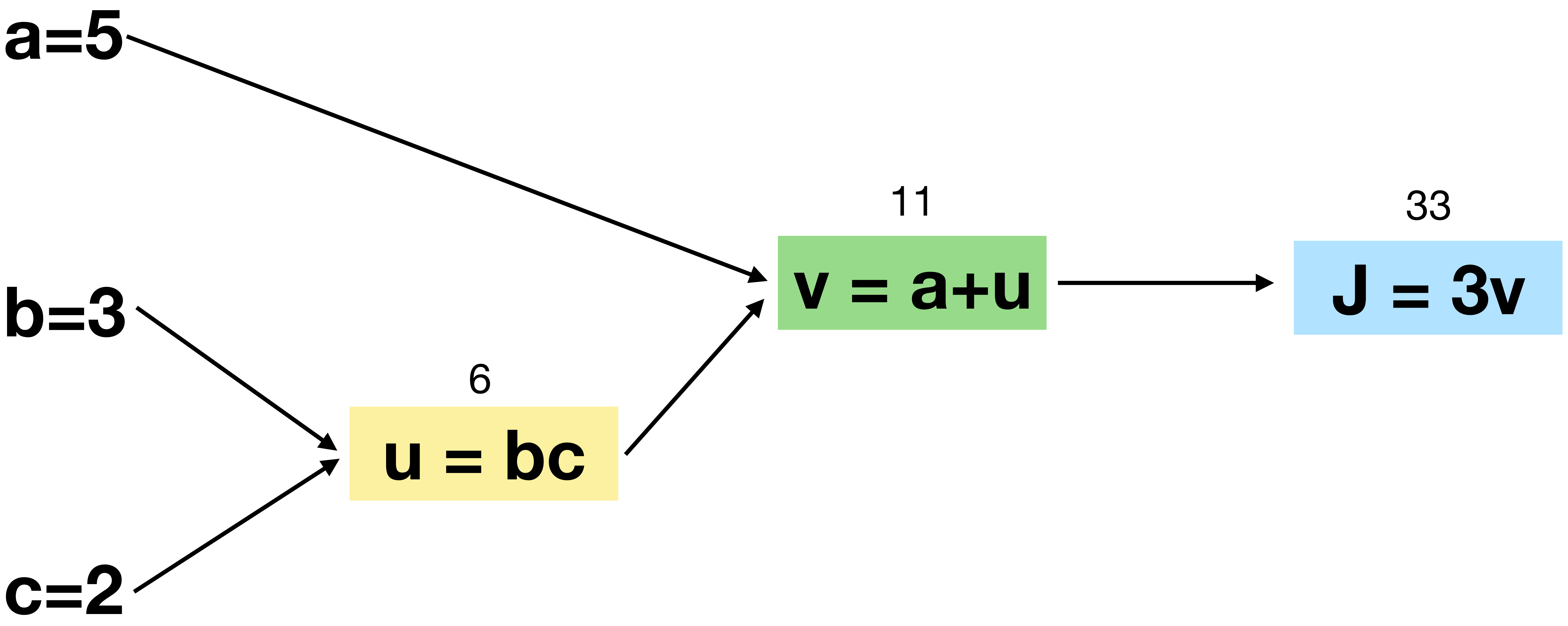
$$J(a,b,c) = 3(\underbrace{a + \underbrace{bc}_{u}}_{v})$$

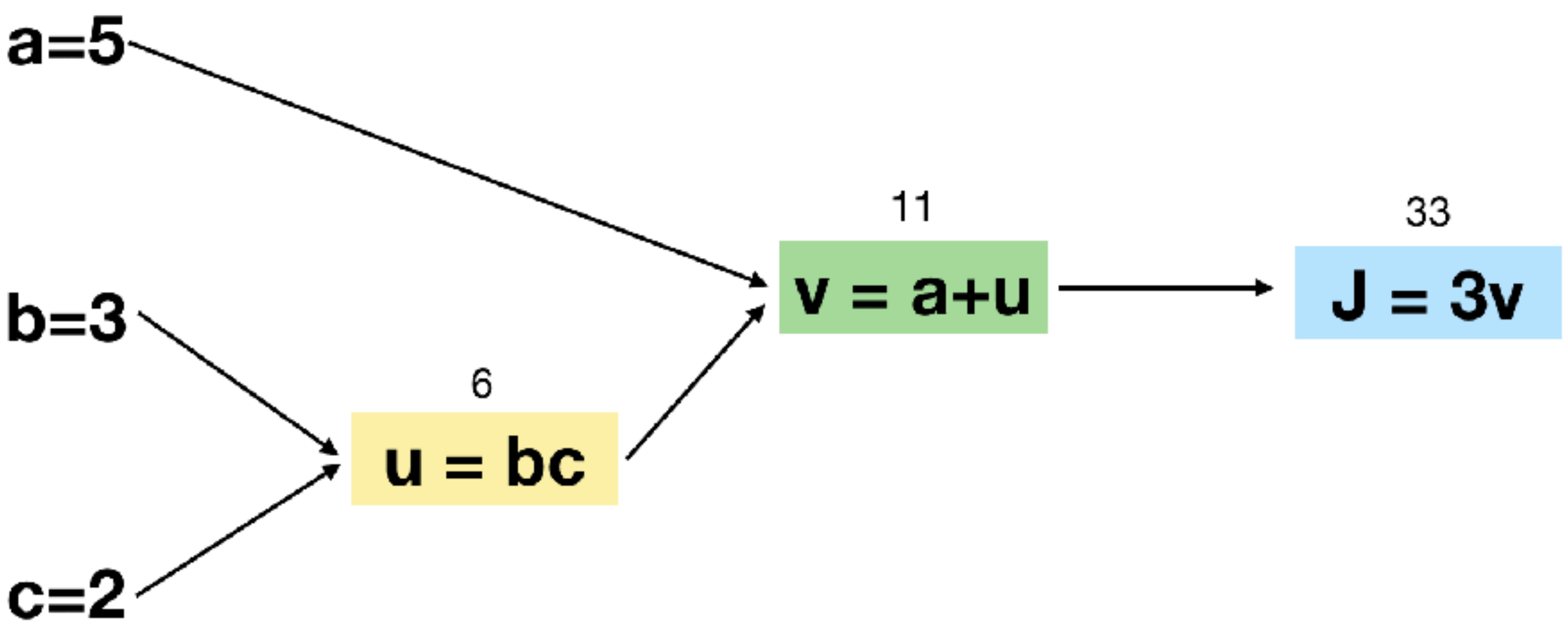
$$J(a,b,c) = 3(5 + 3 \times 2)$$



$$J(a, b, c) = \underbrace{3(\underbrace{a + \underbrace{bc}_{u}}_{v})}_{J}$$

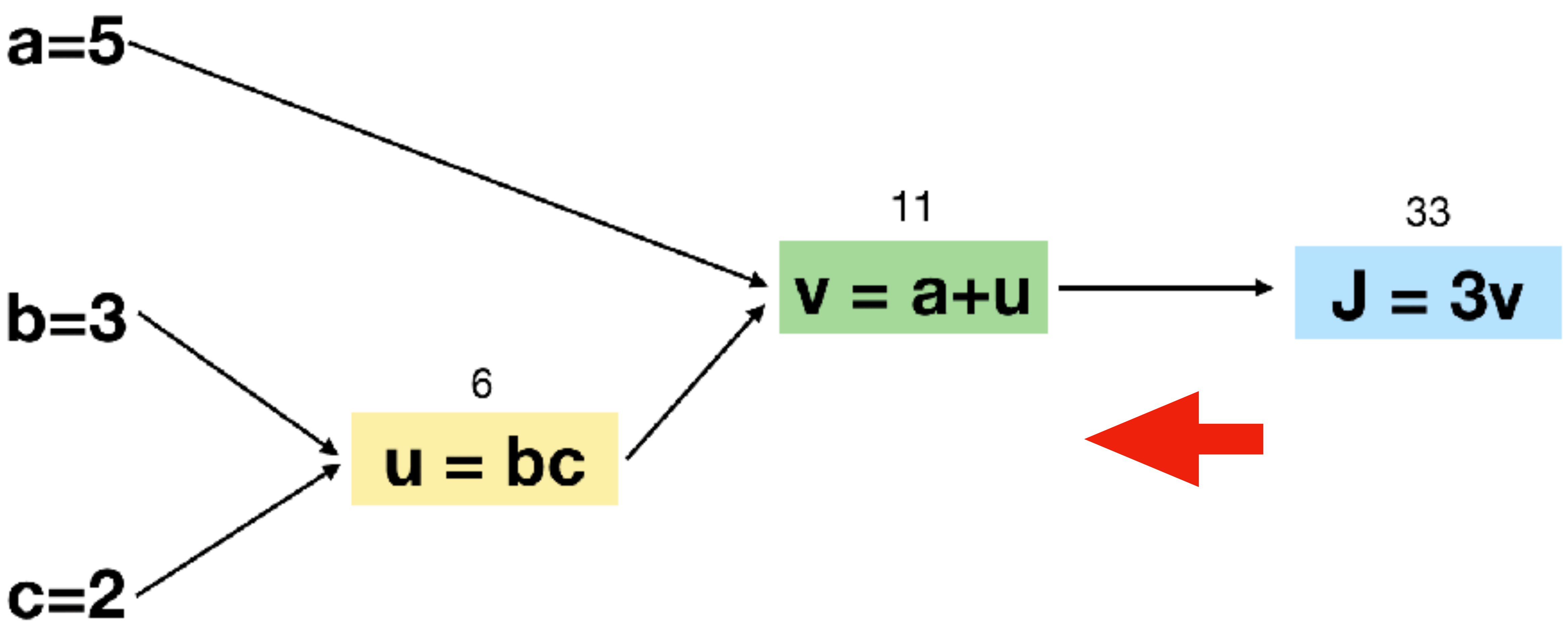
$$J(a, b, c) = 3(5 + 3 \times 2)$$





Derivada de *J* em relação a *u:*

$$J = 3v$$
$$v = 11$$
$$J = 33$$

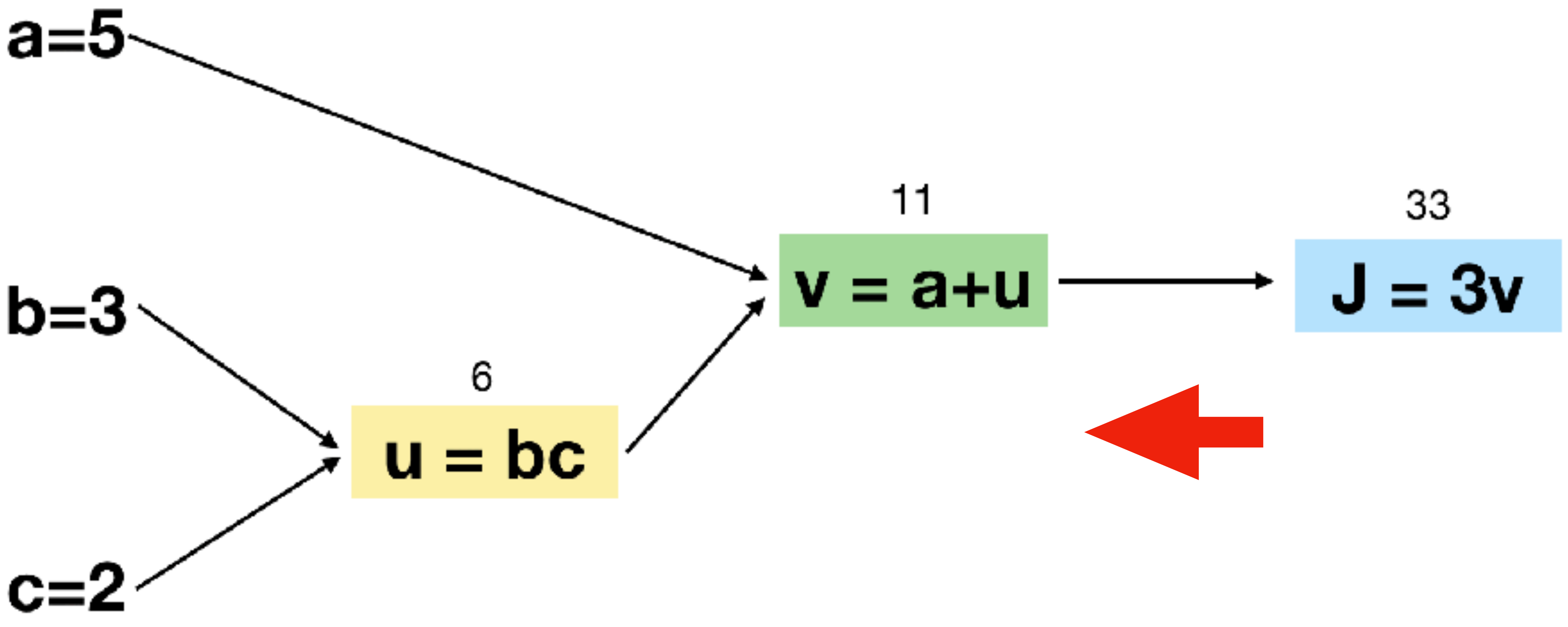


Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33$$

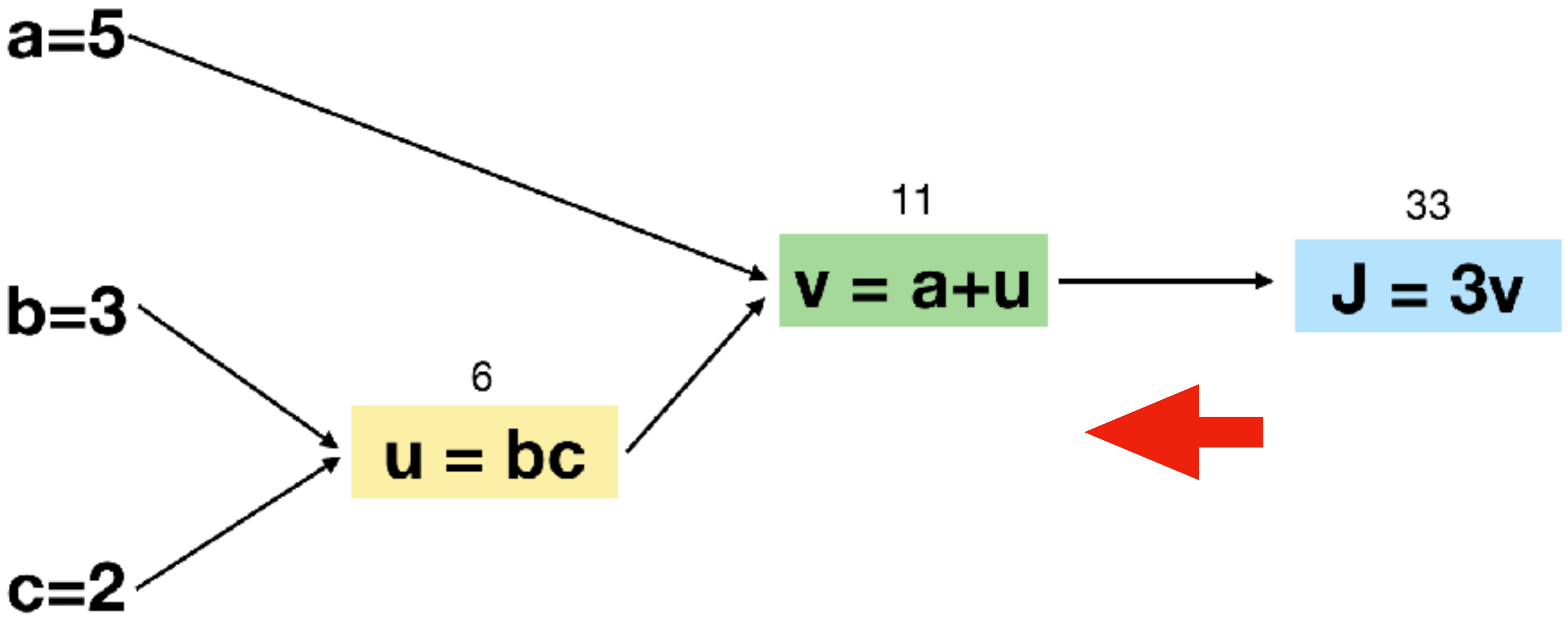


Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$



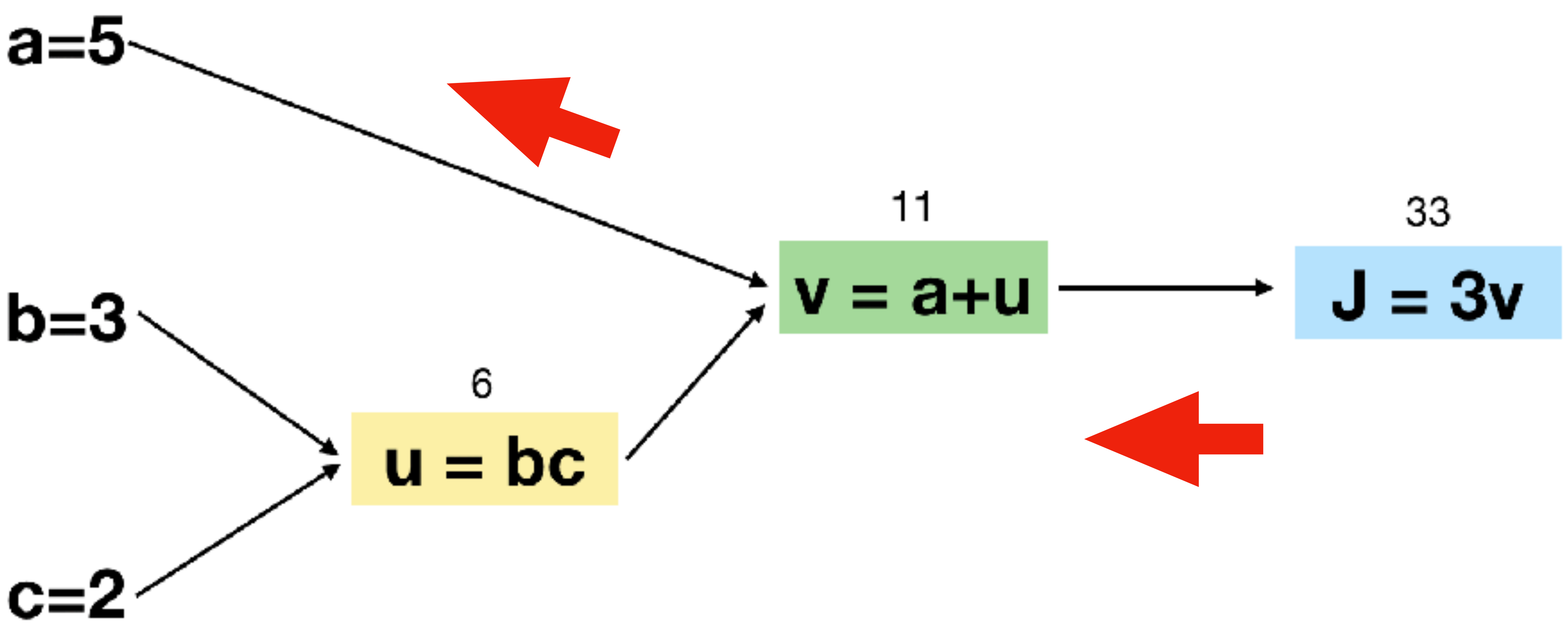
Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

$$\frac{\partial J}{\partial v} = 3$$



Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

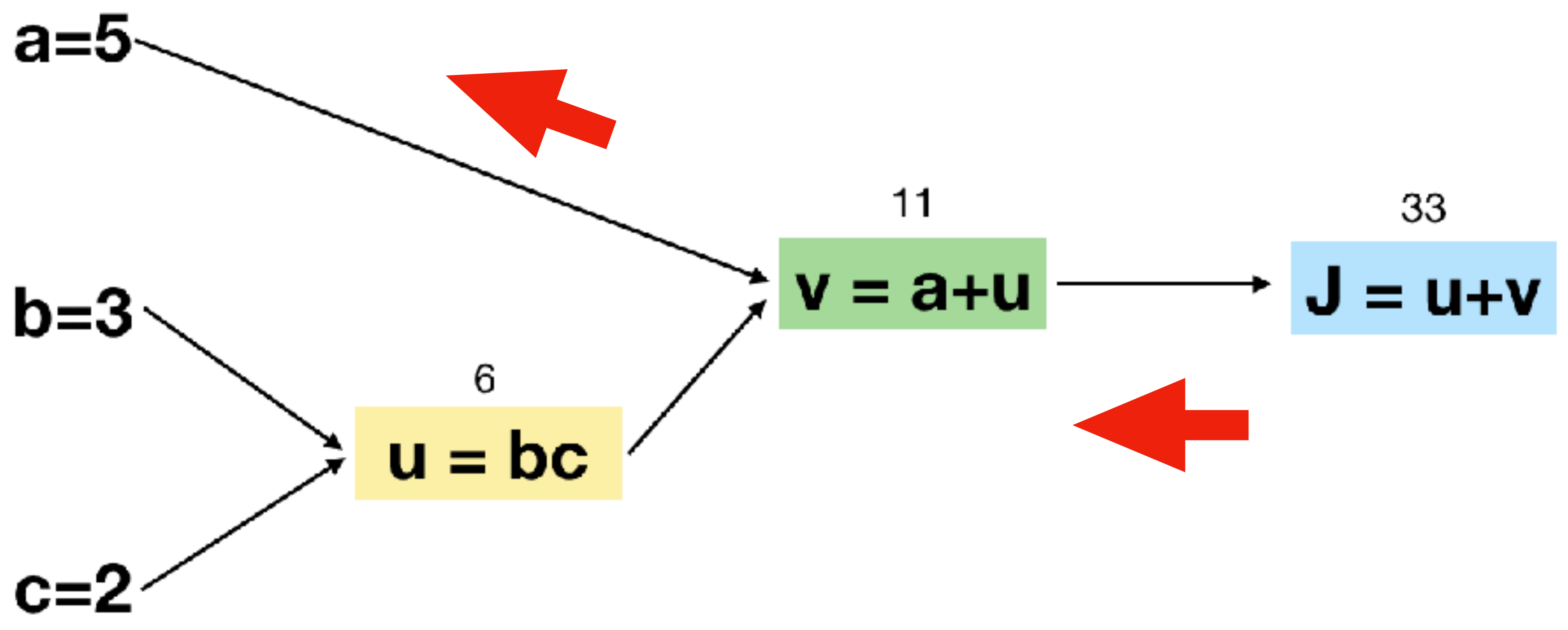
$$\frac{\partial J}{\partial v} = 3$$

Derivada de *J* em relação a *a*:

$$a = 5$$

$$v = 11$$

$$J = 33$$



Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

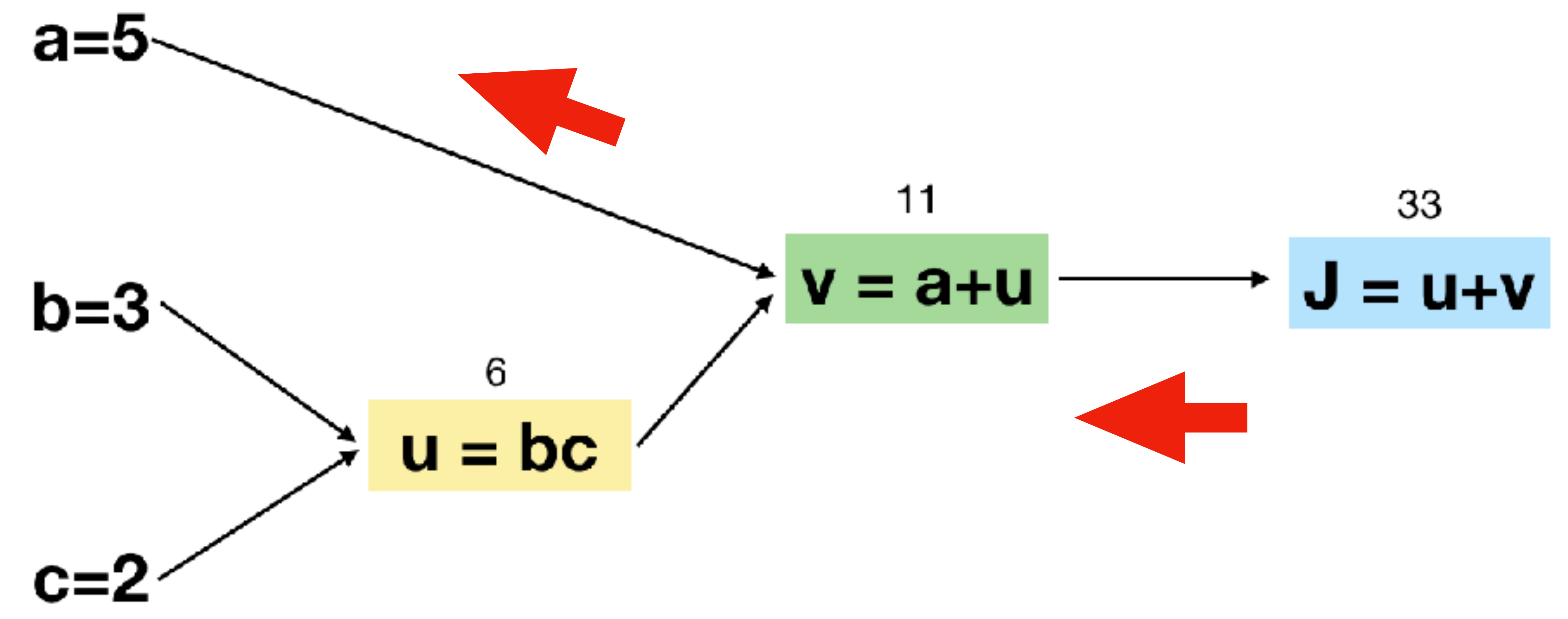
$$\frac{\partial J}{\partial v} = 3$$

Derivada de *J* em relação a *a*:

$$a = 5 \longrightarrow a = 5.001$$

$$v = 11$$

$$J = 33$$



Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

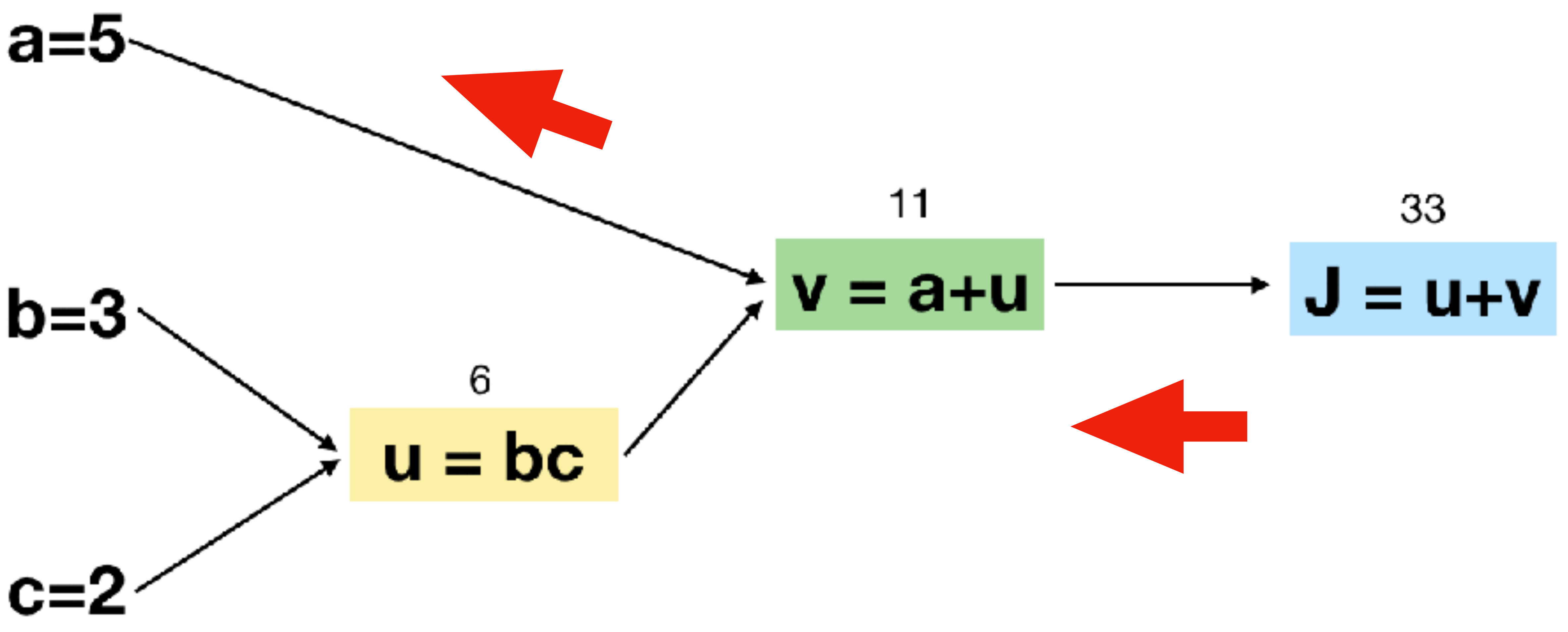
$$\frac{\partial J}{\partial v} = 3$$

Derivada de *J* em relação a *a*:

$$a = 5 \longrightarrow a = 5.001$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33$$



Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

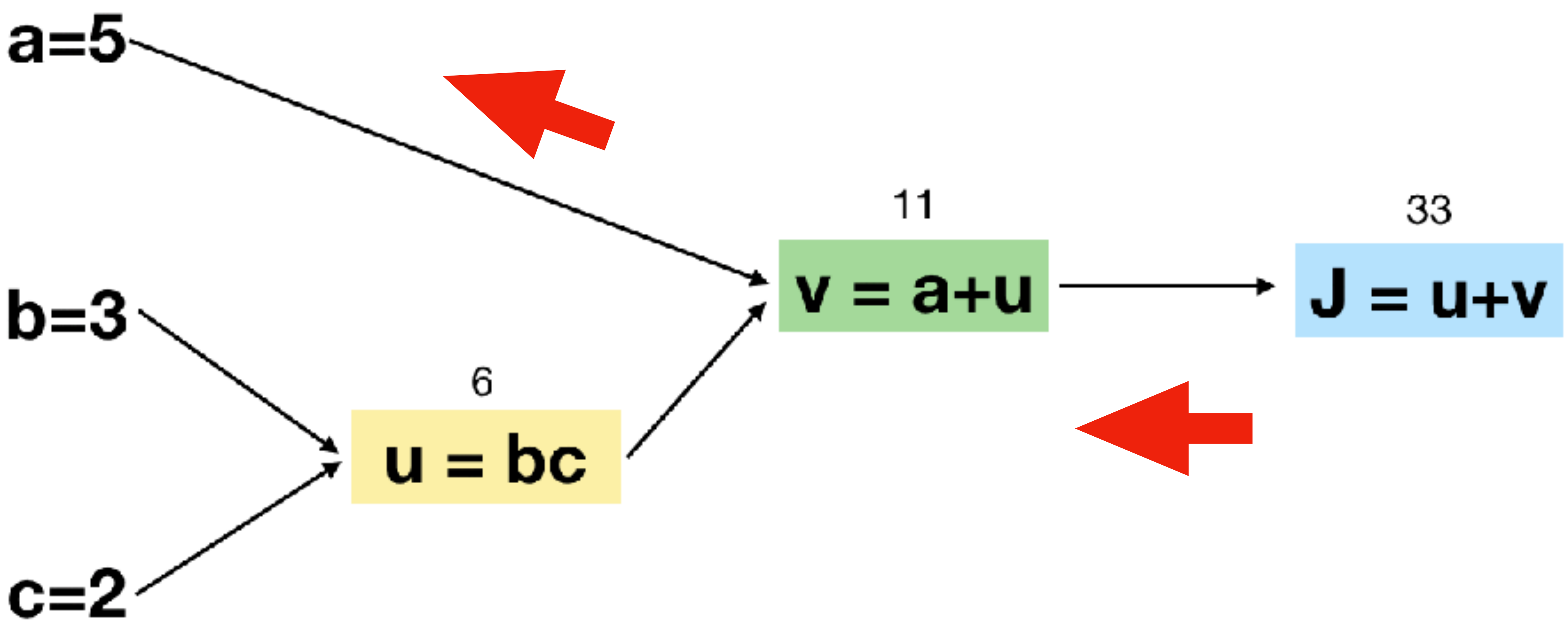
$$\frac{\partial J}{\partial v} = 3$$

Derivada de *J* em relação a *a*:

$$a = 5 \longrightarrow a = 5.001$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$



Derivada de *J* em relação a *u*:

$$J = 3v$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

$$\frac{\partial J}{\partial v} = 3$$

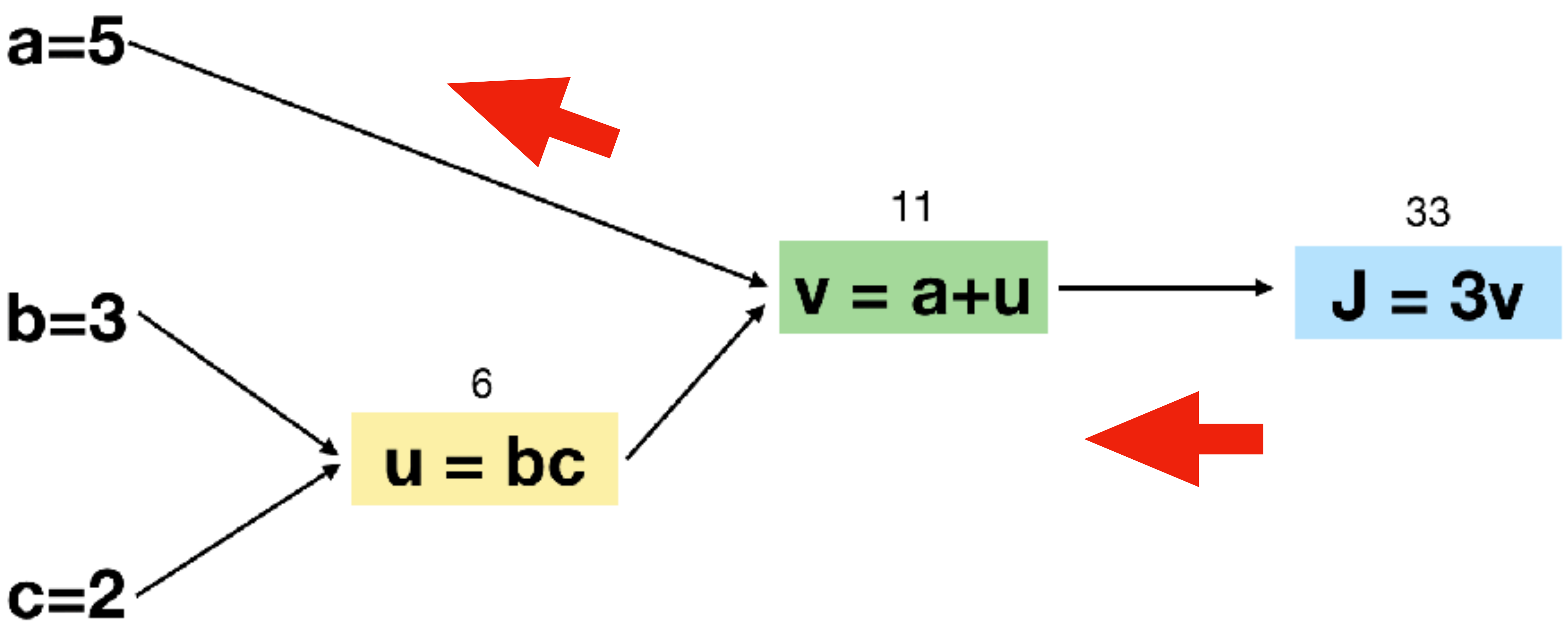
Derivada de *J* em relação a *a*:

$$a = 5 \longrightarrow a = 5.001$$

$$v = 11 \longrightarrow v = 11.001$$

$$J = 33 \longrightarrow J = 31.003$$

$$\frac{\partial J}{\partial a} = 3$$

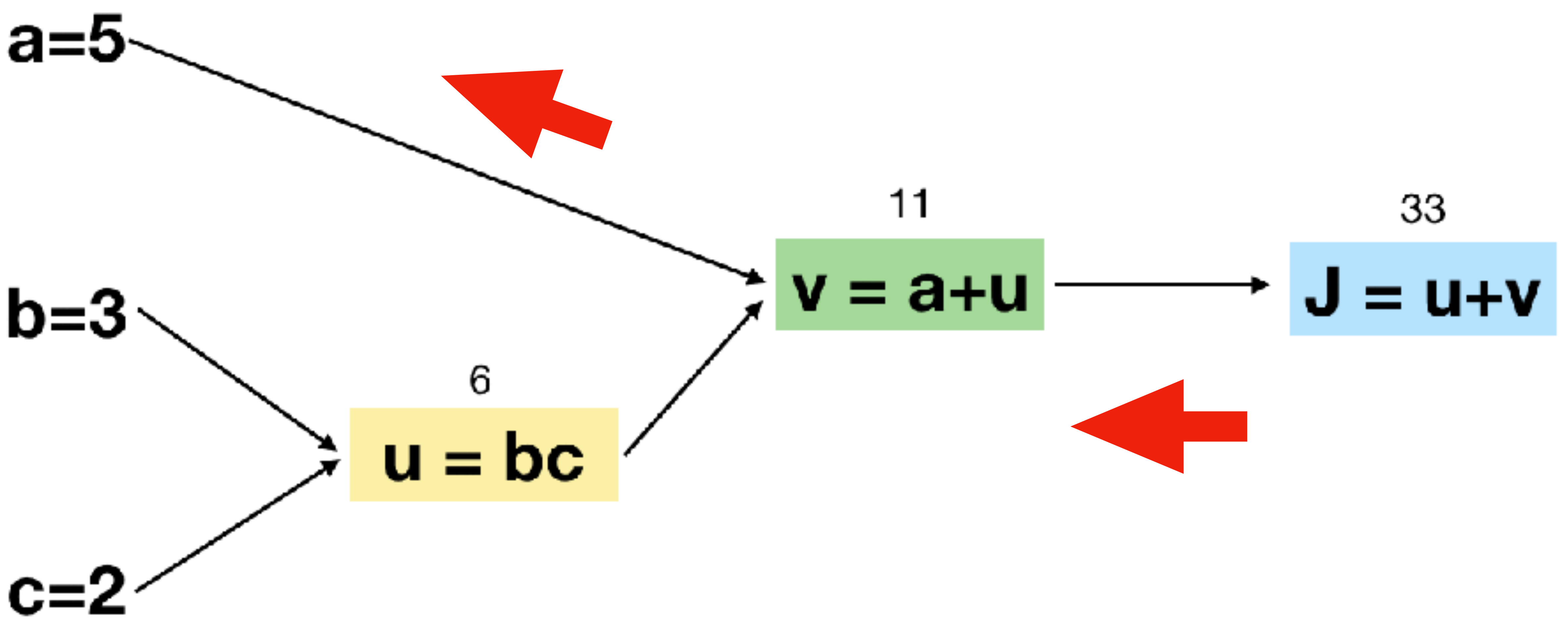


Derivada de *J* em relação a *u*:

$$J = 3v$$
$$v = 11 \longrightarrow v = 11.001 \qquad \frac{\partial J}{\partial v} = 3$$
$$J = 33 \longrightarrow J = 31.003$$

Derivada de *J* em relação a *a*:

$$a = 5 \longrightarrow a = 5.001$$
$$v = 11 \longrightarrow v = 11.001 \qquad \frac{\partial J}{\partial a} = 3 \qquad \frac{\partial J}{\partial a} = \frac{\partial J}{\partial v}\frac{\partial v}{\partial a} = 3$$
$$J = 33 \longrightarrow J = 31.003$$



Derivada de *J* em relação a *u*:

$$J = 3v$$
$$v = 11 \longrightarrow v = 11.001$$
$$J = 33 \longrightarrow J = 31.003$$

$$\frac{\partial J}{\partial v} = 3$$

Derivada de *J* em relação a *a*:

$$a = 5 \longrightarrow a = 5.001$$
$$v = 11 \longrightarrow v = 11.001$$
$$J = 33 \longrightarrow J = 31.003$$

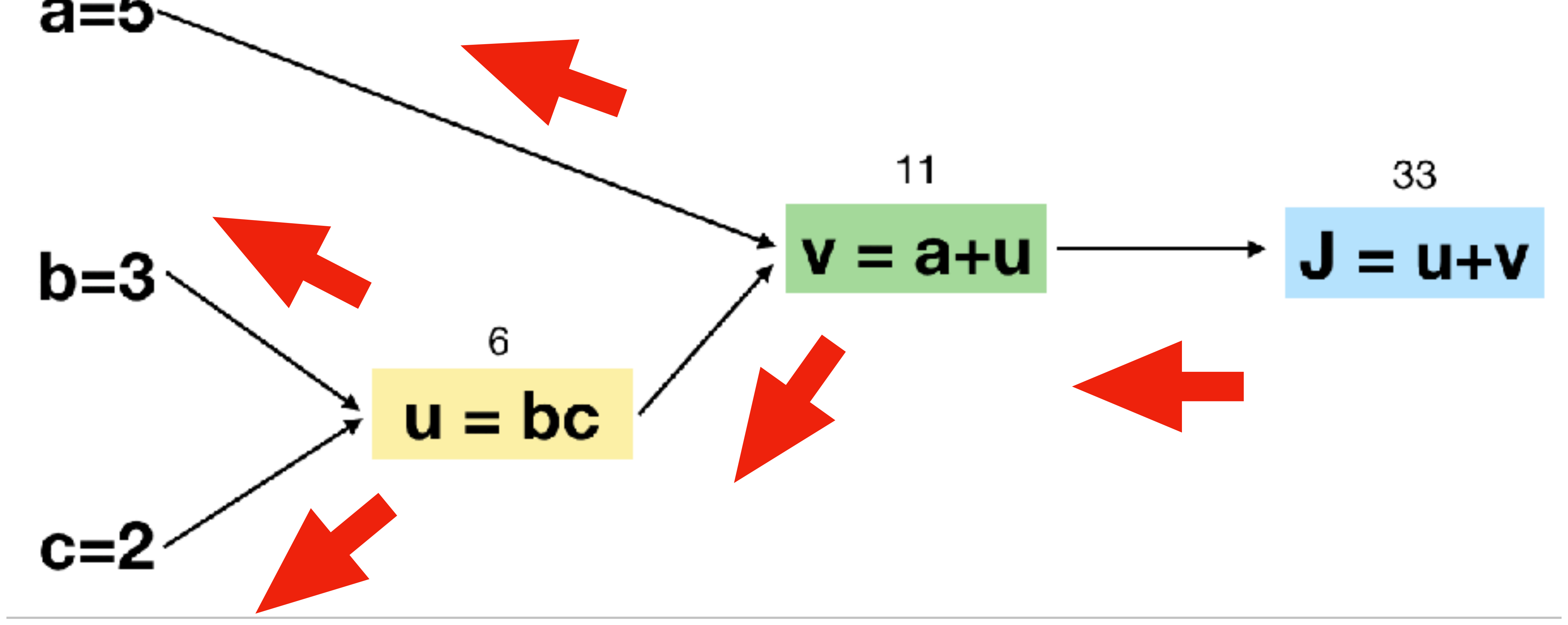
$$\frac{\partial J}{\partial a} = 3$$

$$\frac{\partial J}{\partial a} = \underset{3}{\frac{\partial J}{\partial v}}\,\underset{1}{\frac{\partial v}{\partial a}} = 3$$

$$\frac{\partial J}{\partial a}$$

$$\frac{\partial J}{\partial b}$$

$$\frac{\partial J}{\partial c}$$



$$\hat{y} = \sigma(w^T x + b) \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathscr{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})]$$

Queremos minimizar J(w, b)

- **Encontrar w e b que representem o menor custo**

$$\hat{y} = \sigma(w^T x + b) \qquad \sigma(z) = \frac{1}{1 + e^{-z}}$$

$$J(w, b) = \frac{1}{m}\sum_{i=1}^{m} \mathscr{L}(\hat{y}^{(i)}, y^{(i)}) = -\frac{1}{m}\sum_{i=1}^{m} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) log(1 - \hat{y}^{(i)})]$$

$$\frac{\partial J}{\partial w} = \frac{1}{m} X(A - Y)^T$$

$$\frac{\partial J}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} (a^{(i)} - y^{(i)})$$

Minimizar J(w, b):

$$w := w - \alpha \frac{\partial J(w,b)}{\partial w}$$

$$b := b - \alpha \frac{\partial J(w,b)}{\partial b}$$

$$\frac{\partial J}{\partial w} = \frac{1}{m} X(A - Y)^T$$

$$\frac{\partial J}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} (a^{(i)} - y^{(i)})$$