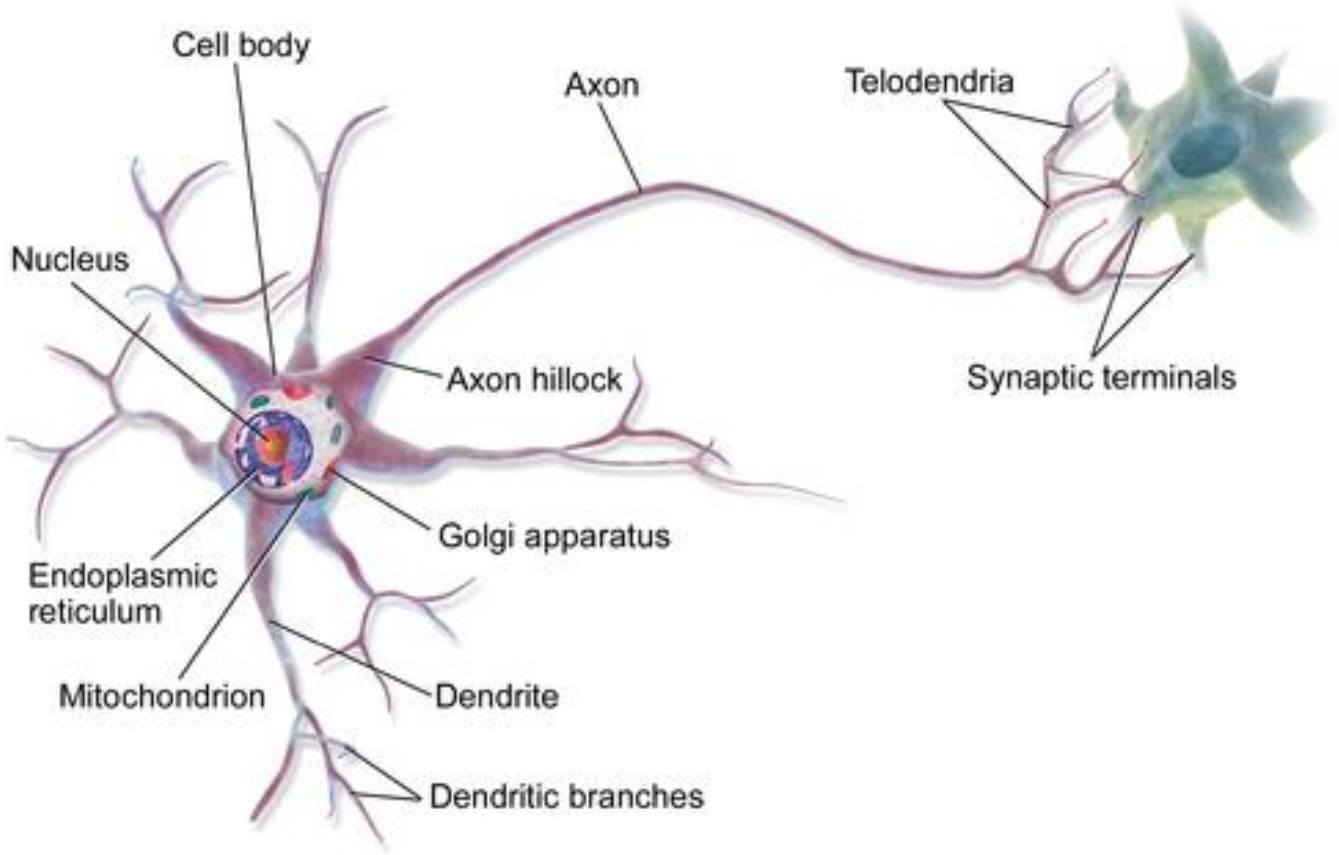
REDES NEURAIS ARTIFICIAIS

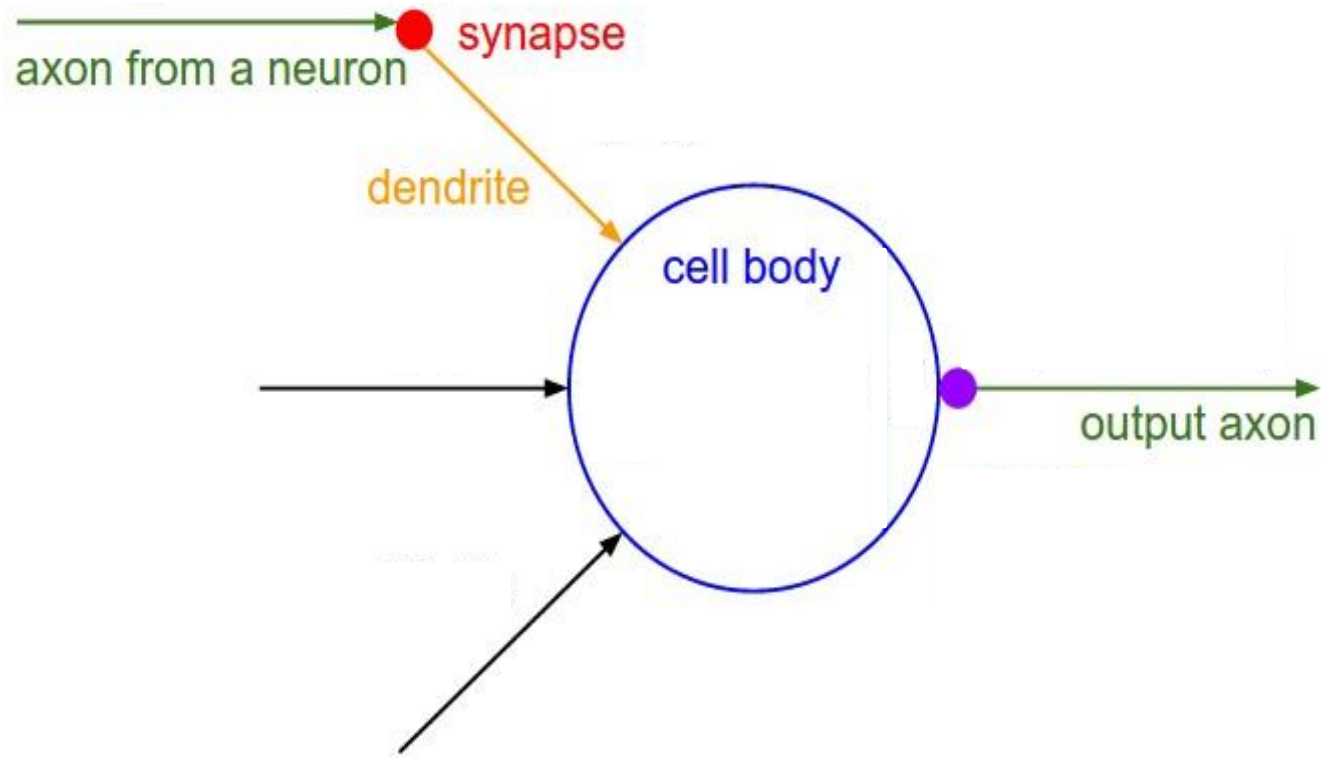
FUNÇÃO DE ATIVAÇÃO

Zenilton K. G. Patrocínio Jr
zenilton@pucminas.br

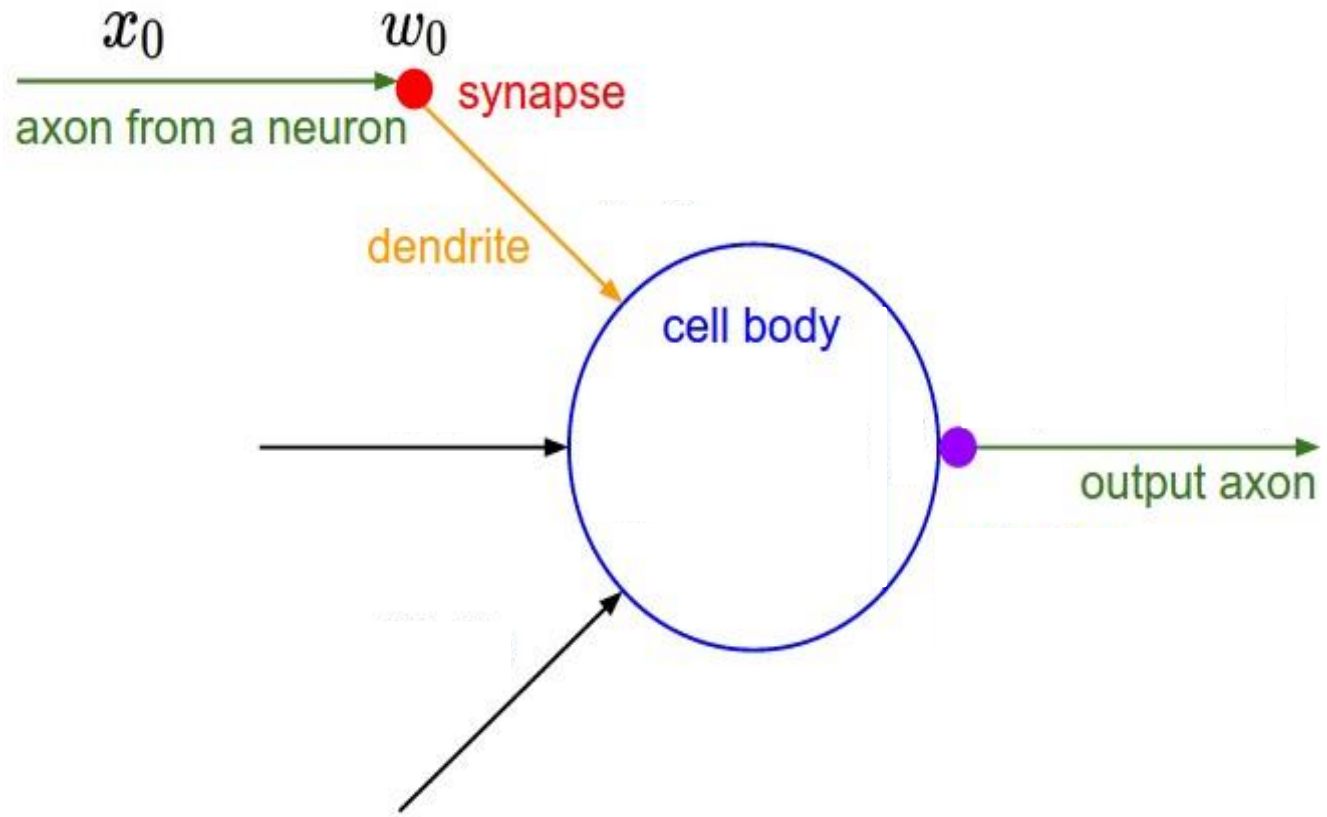
Funções de Ativação – Inspiração Biológica



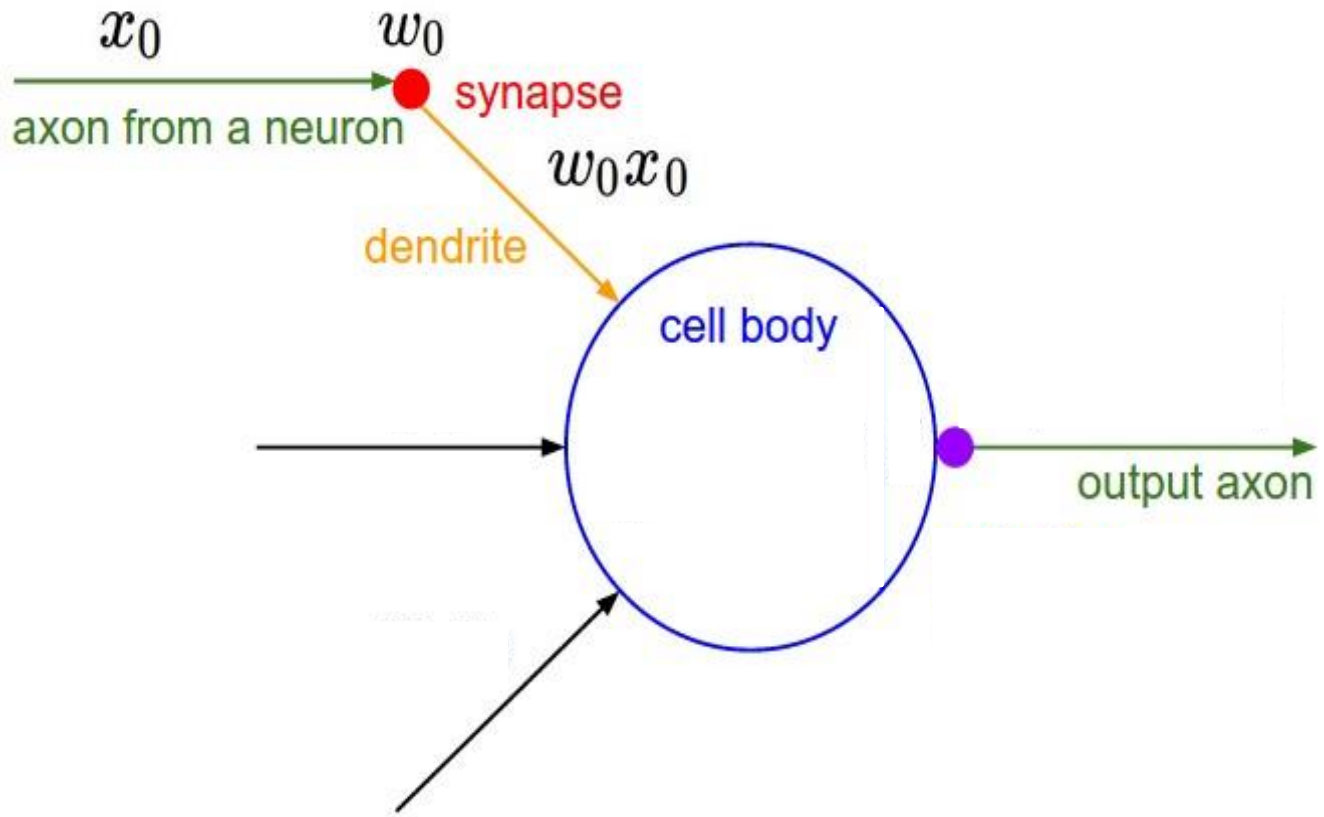
Funções de Ativação



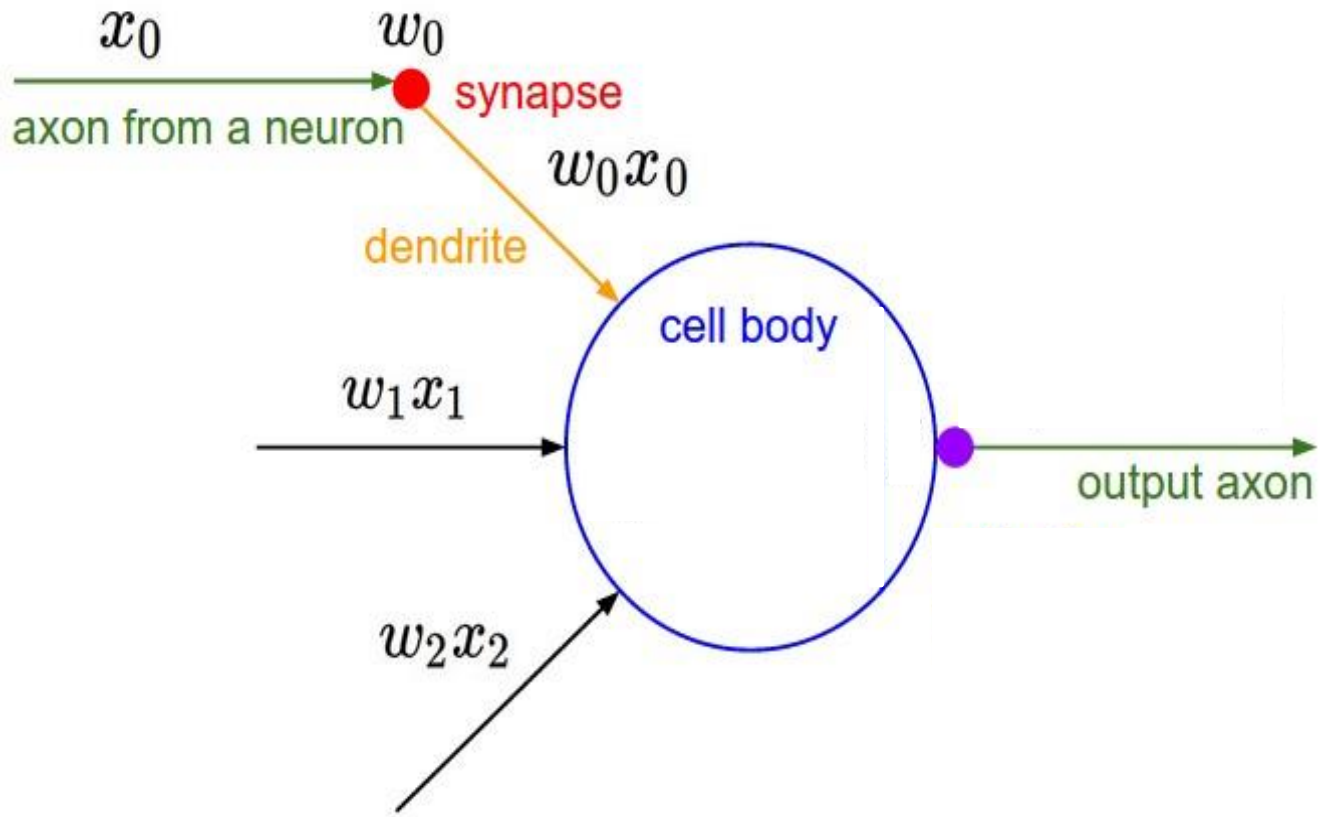
Funções de Ativação



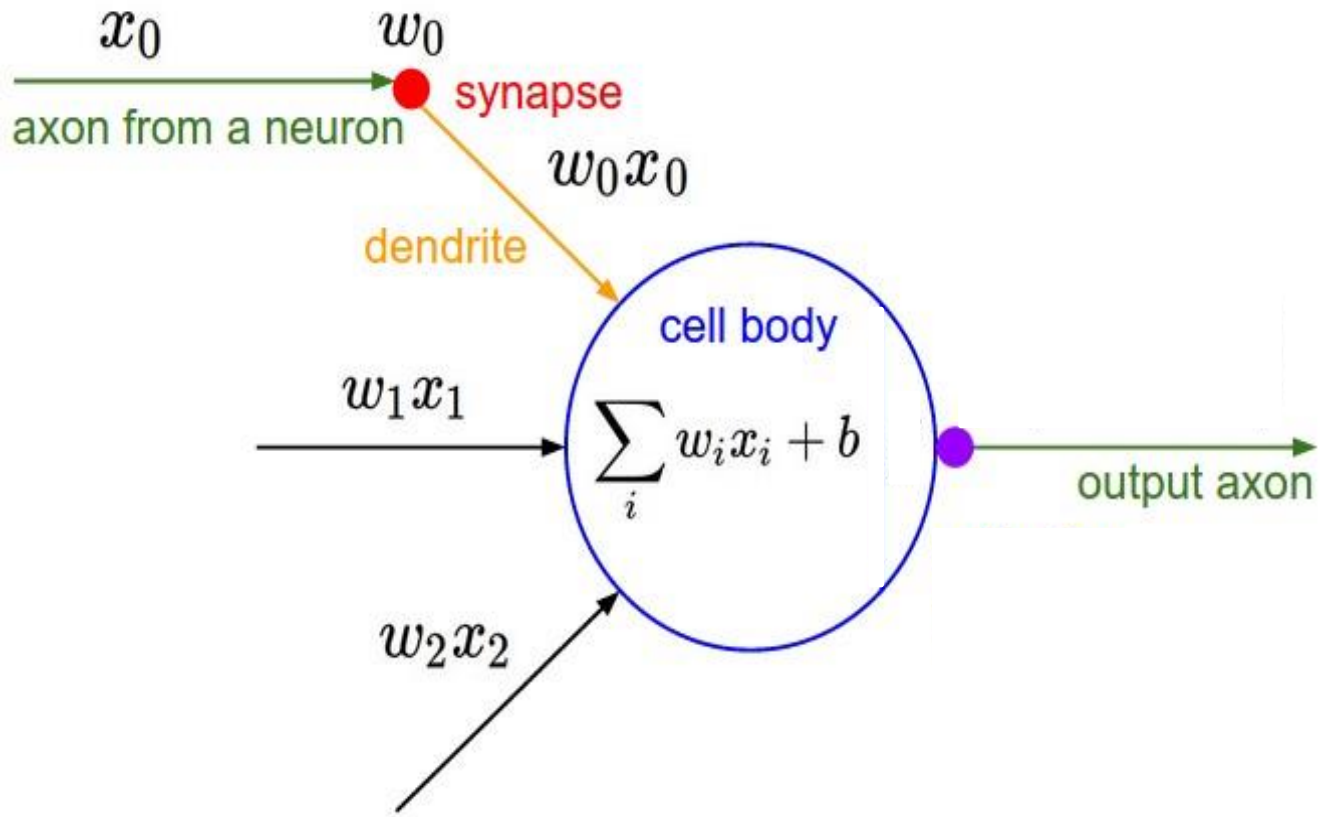
Funções de Ativação



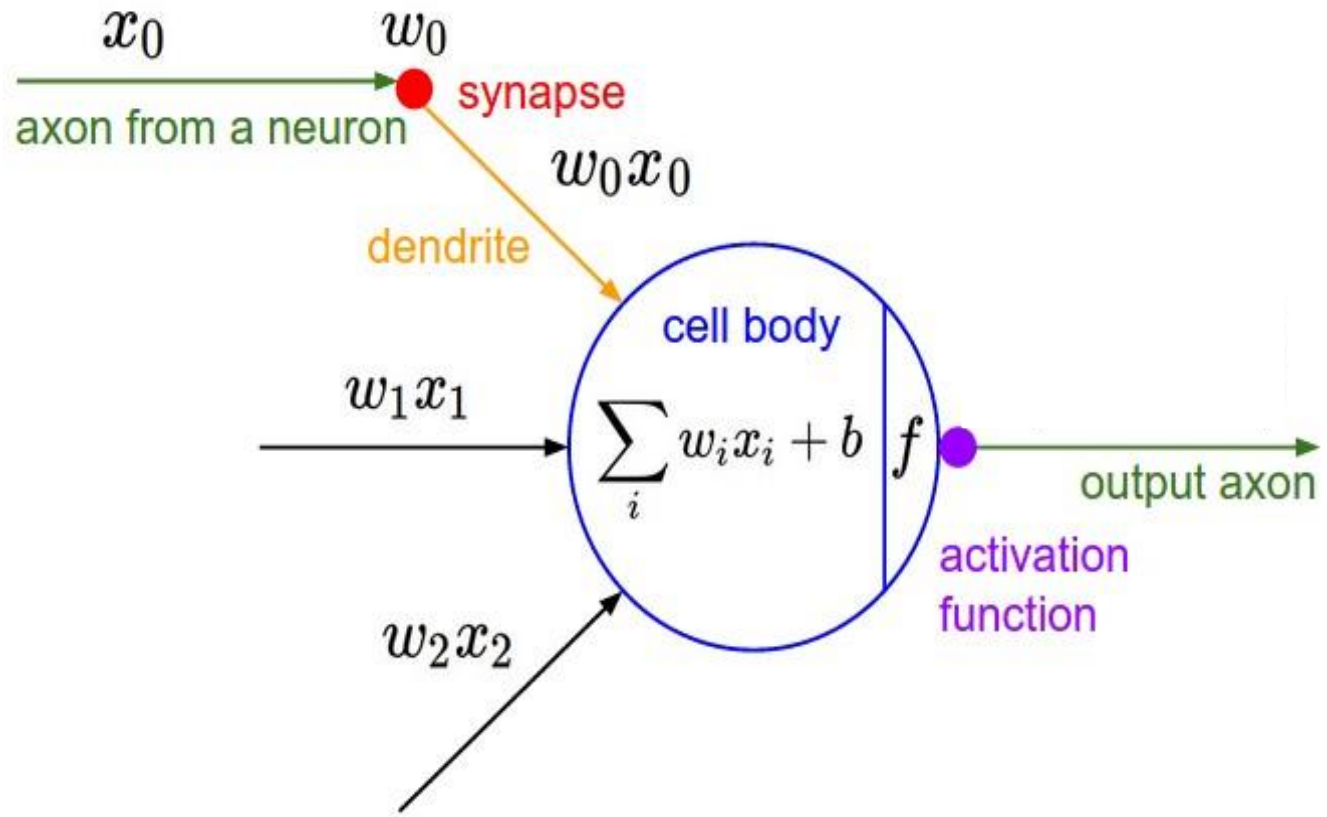
Funções de Ativação



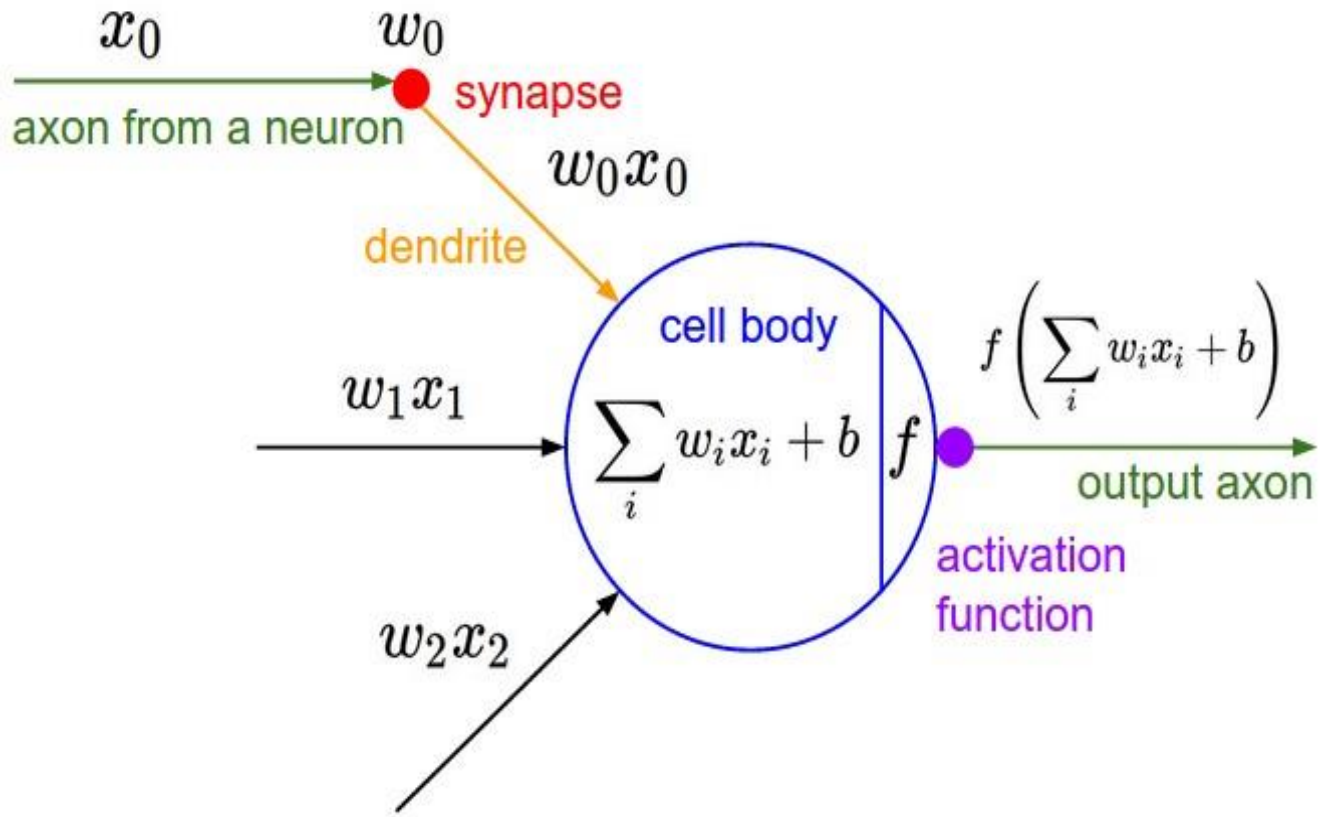
Funções de Ativação



Funções de Ativação



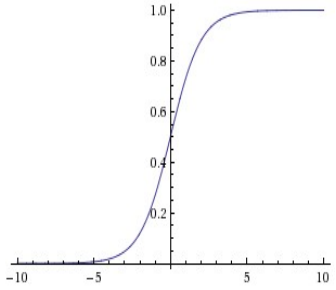
Funções de Ativação



Algumas Funções de Ativação

Sigmoid

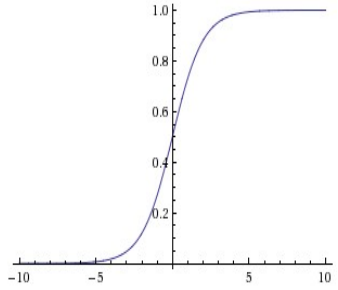
$$\sigma(x) = 1/(1 + e^{-x})$$



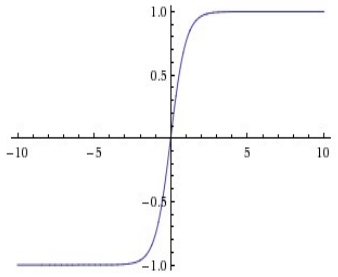
Algumas Funções de Ativação

Sigmoide

$$\sigma(x) = 1/(1 + e^{-x})$$



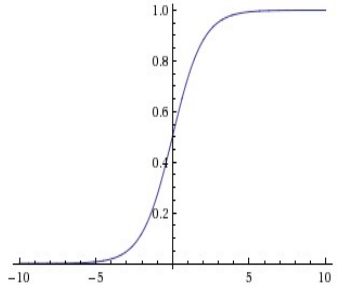
Tanh tanh(x)



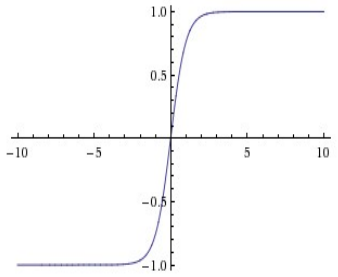
Algumas Funções de Ativação

Sigmoid

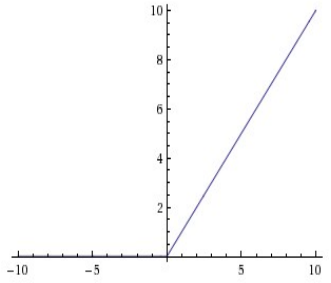
$$\sigma(x) = 1/(1 + e^{-x})$$



Tanh tanh(x)



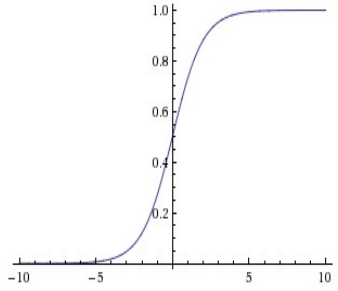
ReLU max(0,x)



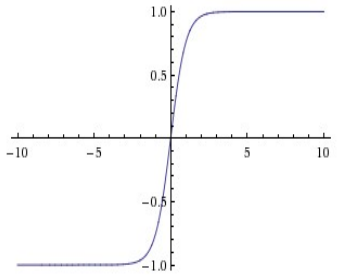
Algumas Funções de Ativação

Sigmoid

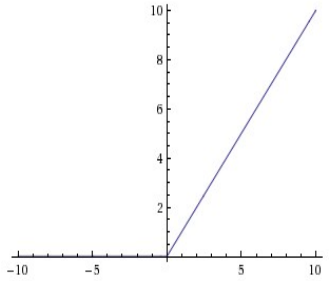
$$\sigma(x) = 1/(1 + e^{-x})$$



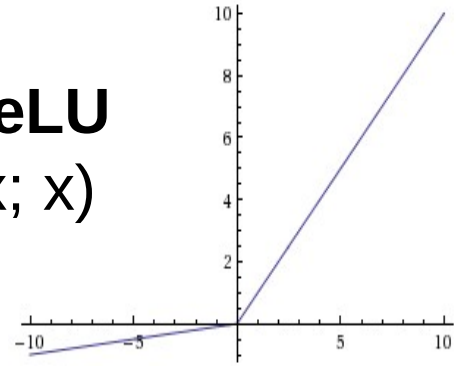
Tanh tanh(x)



ReLU max(0,x)



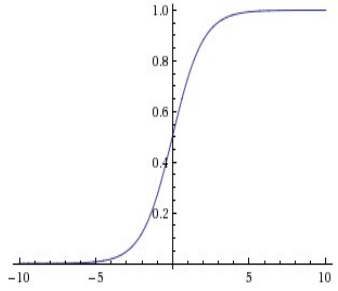
Leaky ReLU
max(0,1x; x)



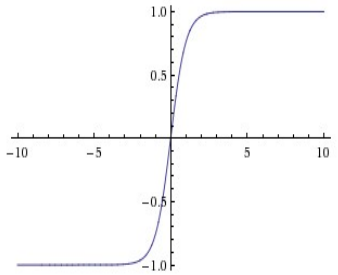
Algumas Funções de Ativação

Sigmoid

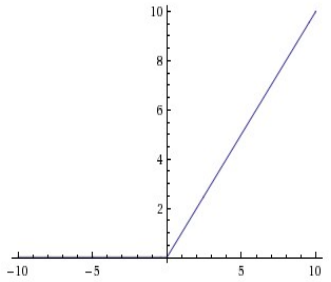
$$\sigma(x) = 1/(1 + e^{-x})$$



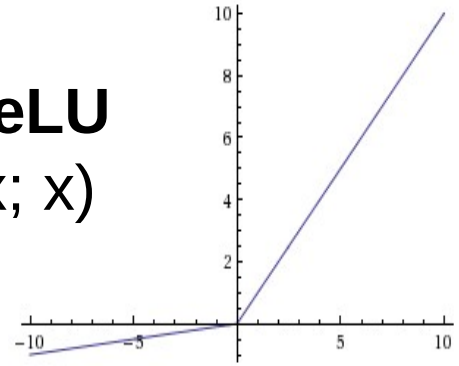
Tanh tanh(x)



ReLU max(0,x)

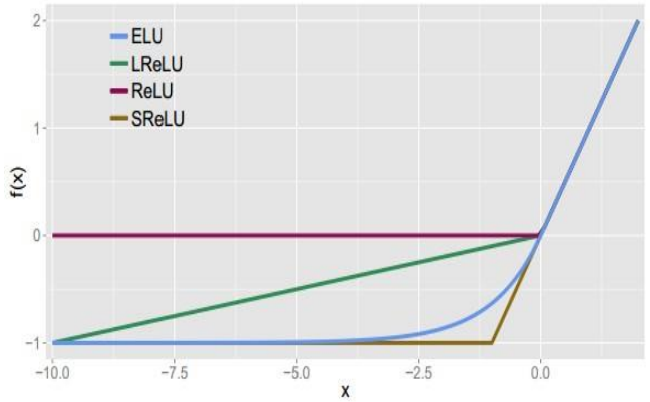


Leaky ReLU
max(0,1x; x)



ELU

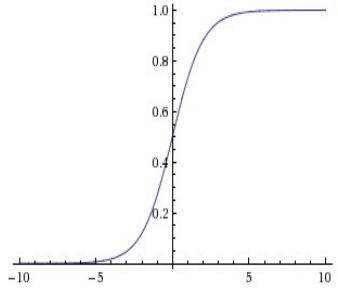
$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha\,(\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$



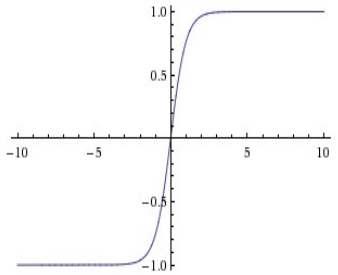
Algumas Funções de Ativação

Sigmoid

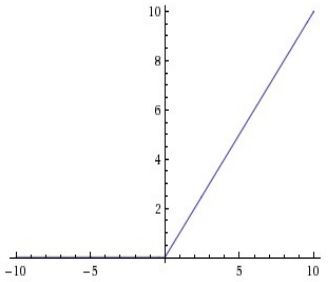
$$\sigma(x) = 1/(1 + e^{-x})$$



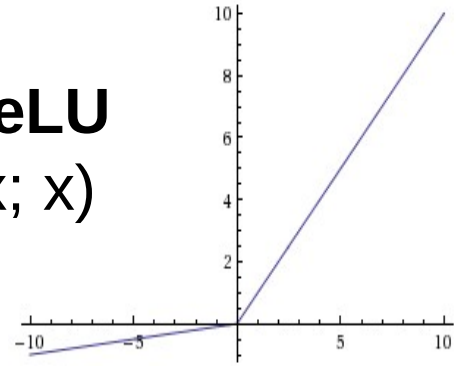
Tanh tanh(x)



ReLU max(0,x)

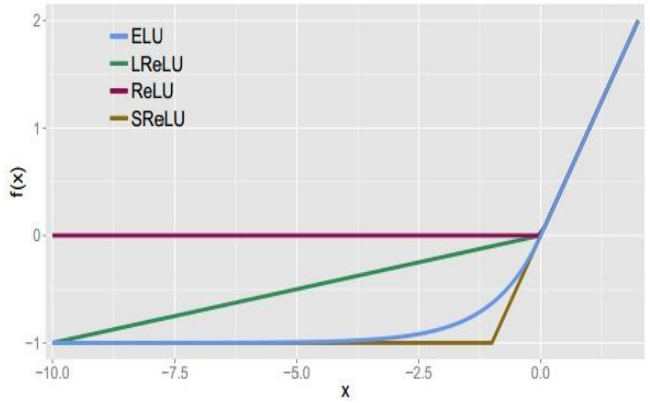


Leaky ReLU
max(0,1x; x)



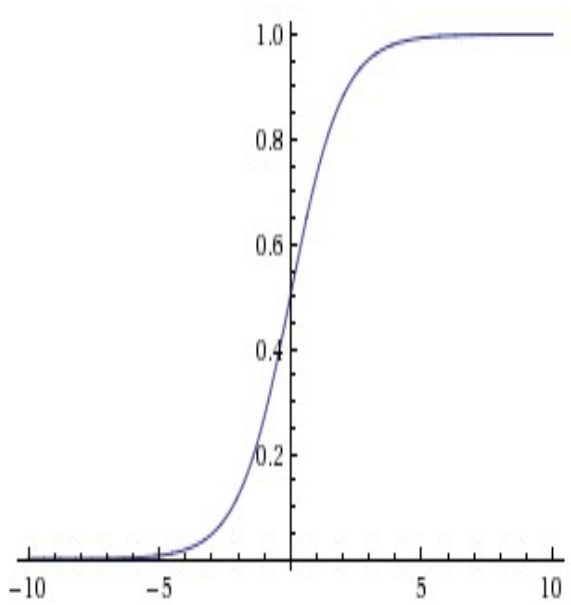
ELU

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha\,(\exp(x) - 1) & \text{if } x \leq 0 \end{cases}$$



Maxout $\max(w_1^T x + b_1, w_2^T x + b_2)$

Função de Ativação – Sigmoide

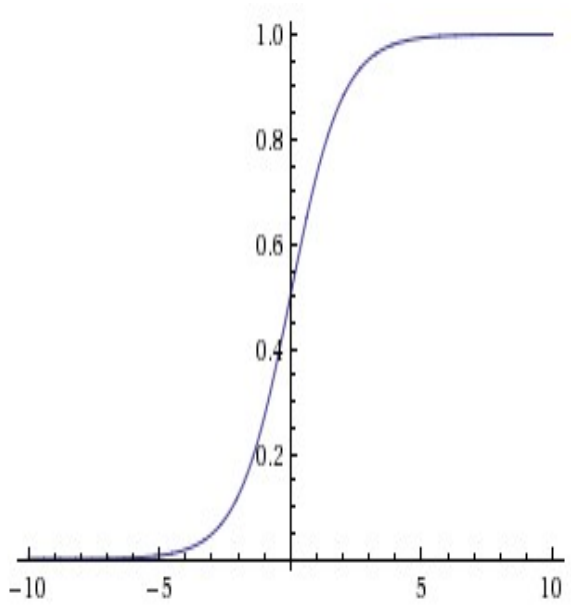


Função Sigmoide (logística)

$$\sigma(x) = 1/(1 + e^{-x})$$

- Historicamente popular, uma vez que tem uma boa interpretação como uma "taxa de disparo" de um neurônio saturado

Função de Ativação – Sigmoide

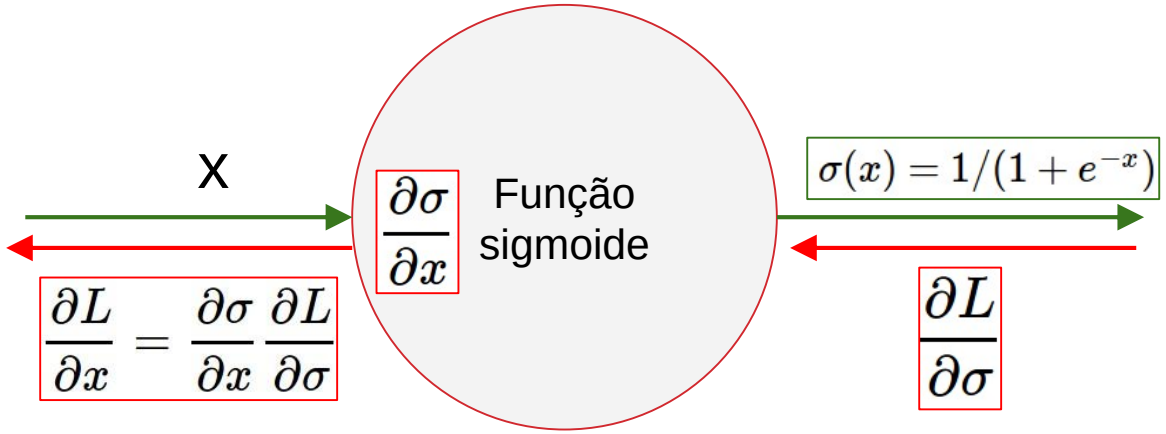


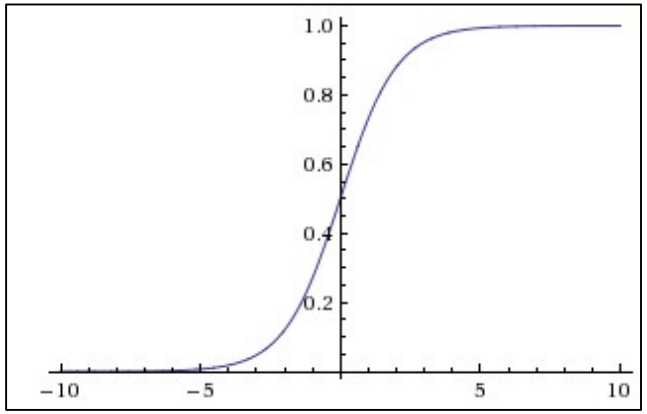
Função Sigmoide (logística)

$$\sigma(x) = 1/(1 + e^{-x})$$

- Historicamente popular, uma vez que tem uma boa interpretação como uma "taxa de disparo" de um neurônio saturado
- “Espreme” os valores para o intervalo [0,1] – pode “matar” (zerar) os gradientes

Função de Ativação – Sigmoide



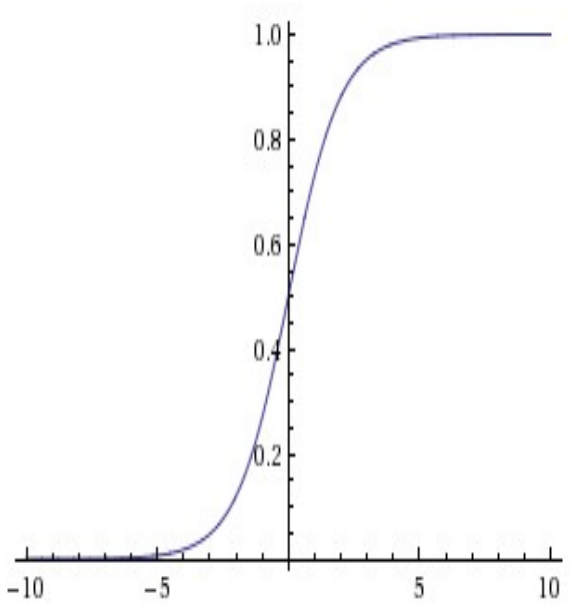


O que acontece quando x = -10?

O que acontece quando x = 0?

O que acontece quando x = 10?

Função de Ativação – Sigmoide



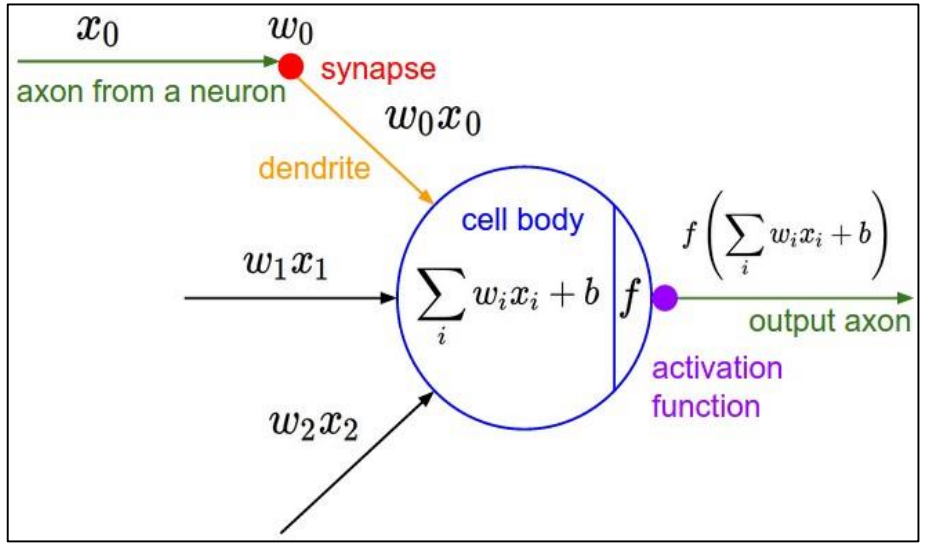
Função Sigmoide (logística)

$$\sigma(x) = 1/(1 + e^{-x})$$

- Historicamente popular, uma vez que tem uma boa interpretação como uma "taxa de disparo" de um neurônio saturado
- “Espreme” os valores para o intervalo [0,1] – pode “matar” (zerar) os gradientes
- Não é centrada em torno de zero

Função de Ativação – Sigmoide

Considere o que acontece quando a entrada de um neurônio (x) é sempre positiva:



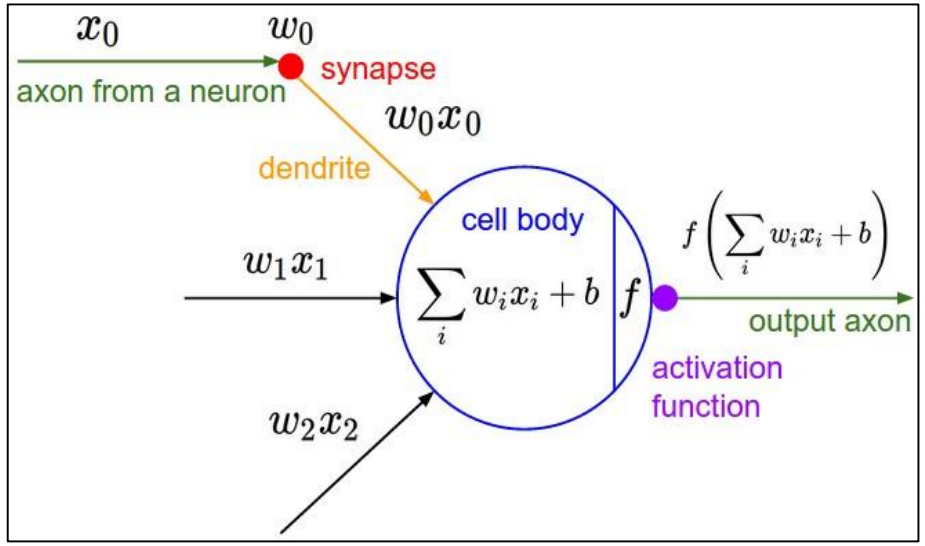
$$f\left(\sum_i w_i x_i + b\right)$$

$$\sigma(x) = 1/(1+e^{-x})$$

O que se pode dizer sobre os gradientes em relação a **w**?

Função de Ativação – Sigmoide

Considere o que acontece quando a entrada de um neurônio (x) é sempre positiva:



$$f\left(\sum_i w_i x_i + b\right)$$

$$\sigma(x) = 1/(1+e^{-x})$$

O que se pode dizer sobre os gradientes em relação a **w**?

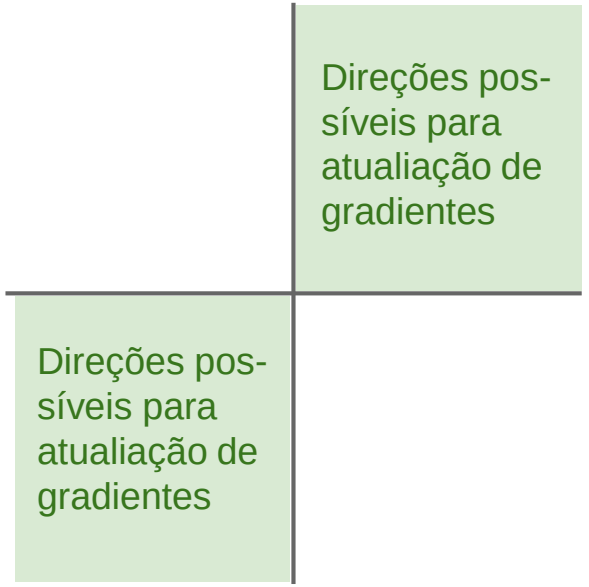
Sempre todos positivos ou negativos :(

Função de Ativação – Sigmoide

Considere o que acontece quando a entrada de um neurônio (x) é sempre positiva:

$$f\left(\sum_i w_i x_i + b\right)$$

$$\sigma(x) = 1/(1+e^{-x})$$



O que se pode dizer sobre os gradientes em relação a **w**?

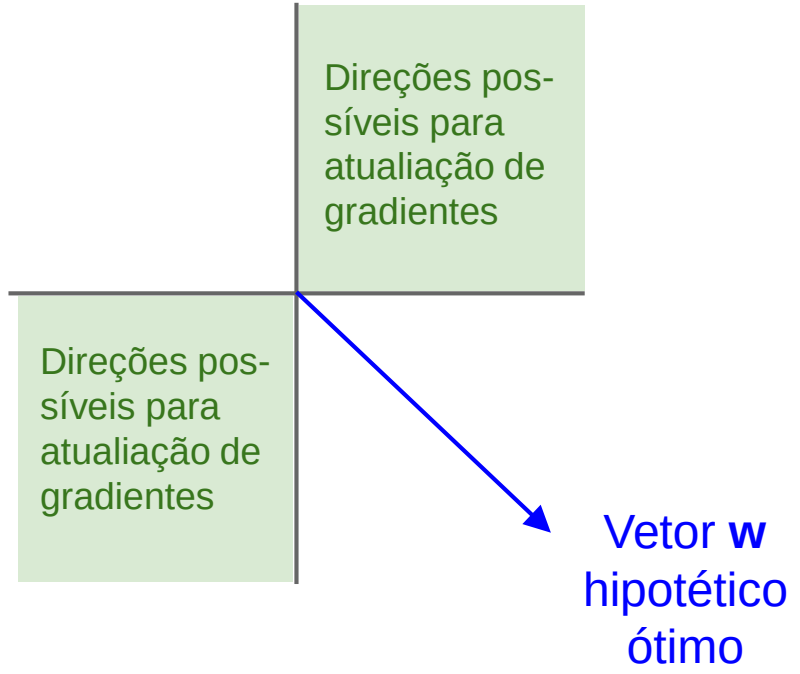
Sempre todos positivos ou negativos :(

Função de Ativação – Sigmoide

Considere o que acontece quando a entrada de um neurônio (x) é sempre positiva:

$$f\left(\sum_i w_i x_i + b\right)$$

$$\sigma(x) = 1/(1+e^{-x})$$



O que se pode dizer sobre os gradientes em relação a **w**?

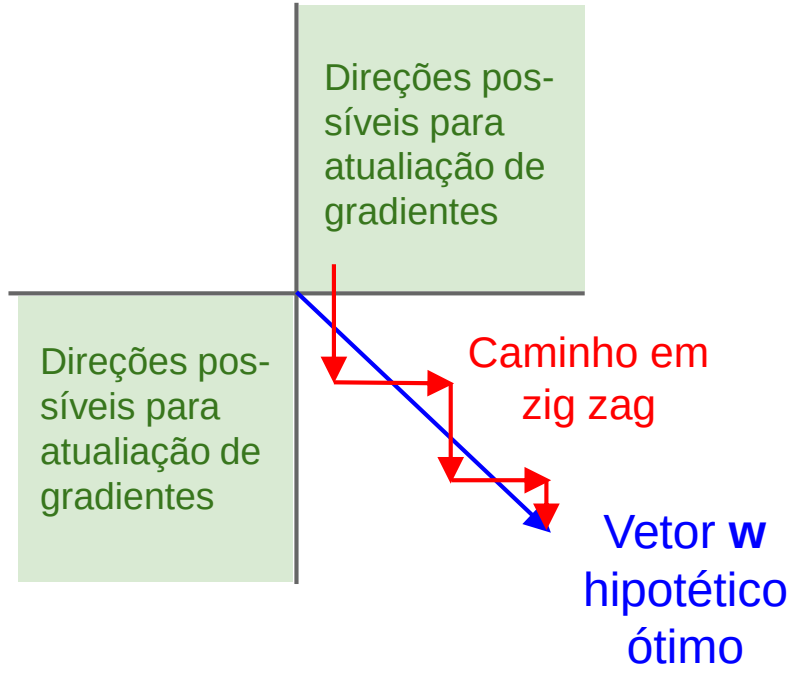
Sempre todos positivos ou negativos :(

Função de Ativação – Sigmoide

Considere o que acontece quando a entrada de um neurônio (x) é sempre positiva:

$$f\left(\sum_i w_i x_i + b\right)$$

$$\sigma(x) = 1/(1+e^{-x})$$



O que se pode dizer sobre os gradientes em relação a **w**?

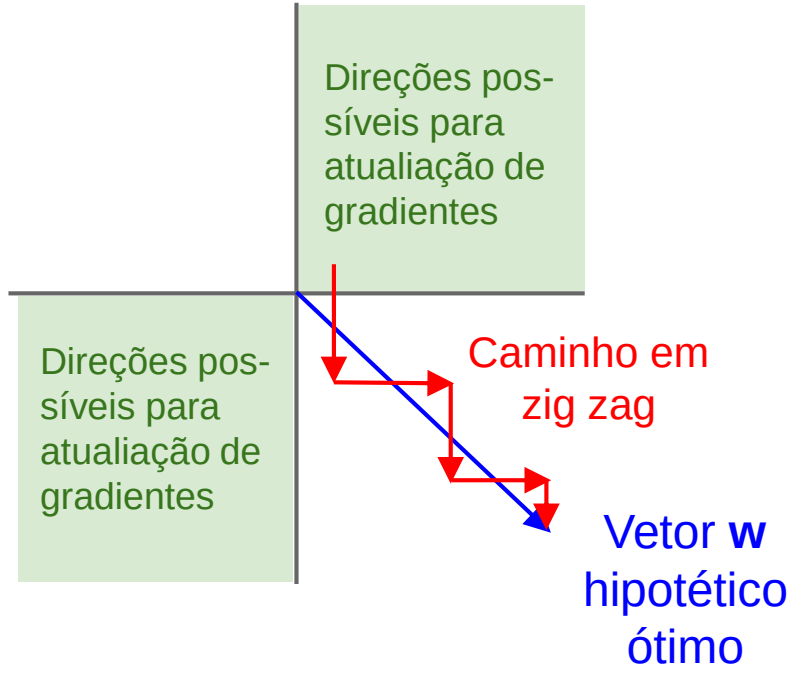
Sempre todos positivos ou negativos :(

Função de Ativação – Sigmoide

Considere o que acontece quando a entrada de um neurônio (x) é sempre positiva:

$$f\left(\sum_i w_i x_i + b\right)$$

$$\sigma(x) = 1/(1+e^{-x})$$

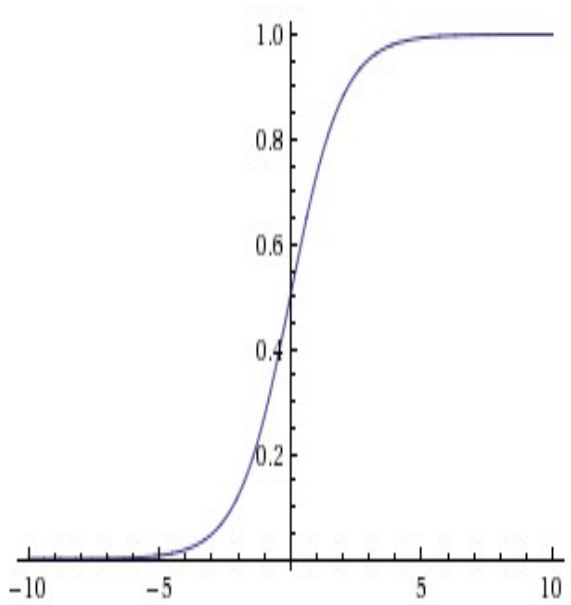


O que se pode dizer sobre os gradientes em relação a **w**?

Sempre todos positivos ou negativos :(

É também por isso que se deseja dados com média zero!

Função de Ativação – Sigmoide

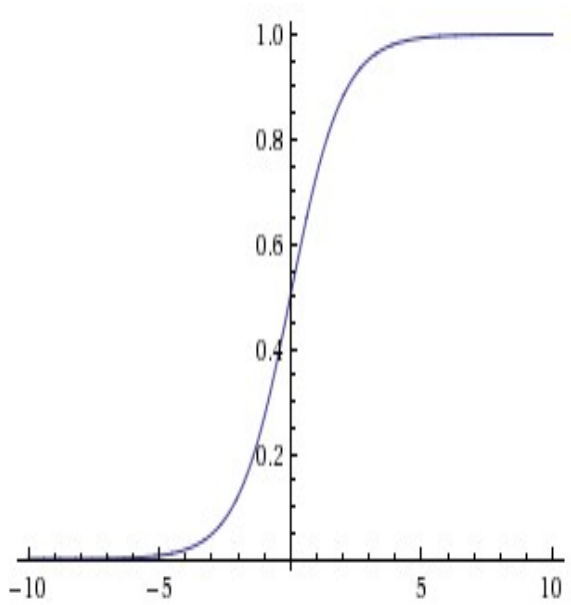


Função Sigmoide (logística)

$$\sigma(x) = 1/(1 + e^{-x})$$

- Historicamente popular, uma vez que tem uma boa interpretação como uma "taxa de disparo" de um neurônio saturado
- “Espreme” os valores para o intervalo [0,1] – pode “matar” (zerar) os gradientes
- Não é centrada em torno de zero
- O uso de exp() é um pouco “caro”
- Não é adequada para tratamento de imagens (substituída por ReLU)

Função de Ativação – Sigmoide

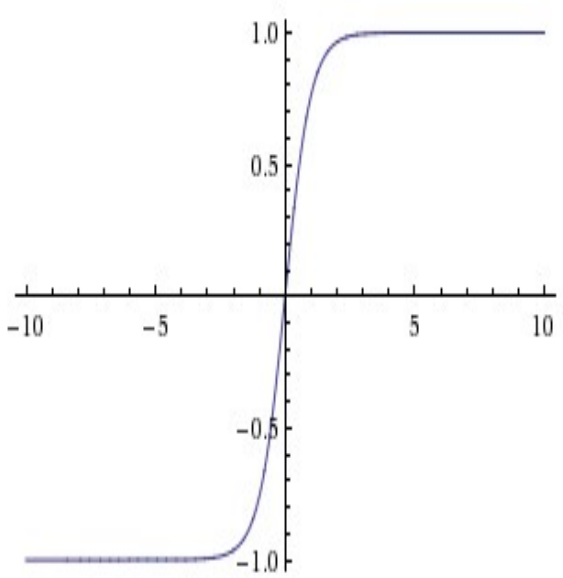


Função Sigmoide (logística)

$$\sigma(x) = 1/(1 + e^{-x})$$

- Historicamente popular, uma vez que tem uma boa interpretação como uma "taxa de disparo" de um neurônio saturado
- “Espreme” os valores para o intervalo [0,1] – pode “matar” (zerar) os gradientes
- Não é centrada em torno de zero
- O uso de exp() é um pouco “caro”
- Não é adequada para tratamento de imagens (substituída por ReLU)
- É um elemento chave em redes LSTM – “controle de sinais”
- Ideal para aprendizado de funções “lógicas” – pois produz resultado no intervalo [0, 1]

Função de Ativação – Tanh

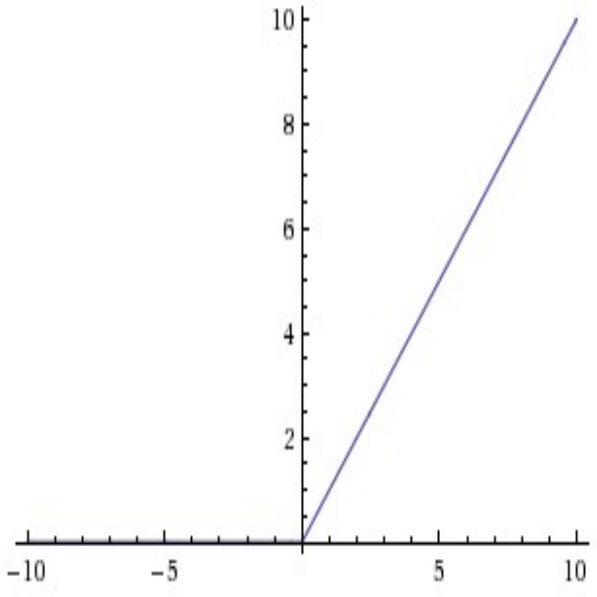


Tanh(x)

[LeCun et al., 1991]

- “Espreme” os valores para o intervalo [-1,1]
- É centrada em zero (que é bom)
- Ainda “mata” os gradientes quando saturada :(
- Também é usada em redes LSTM para valores limitados e com sinal
- Não é “boa” para funções binárias

Função de Ativação – ReLU

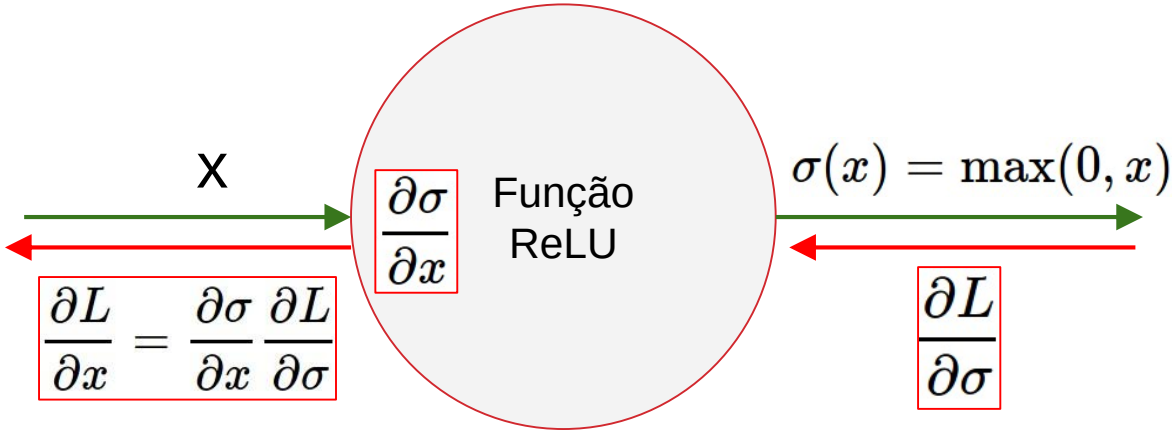


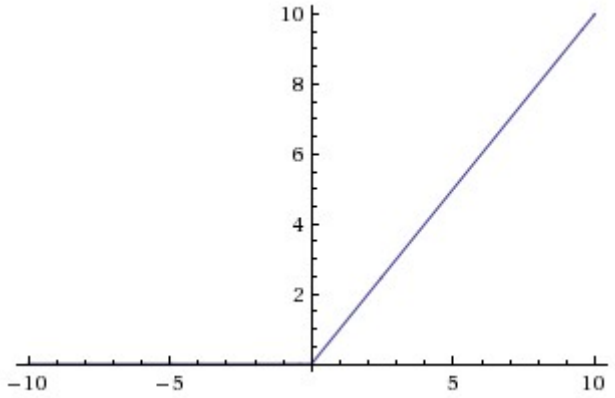
ReLU
(*Rectified Linear Unit*)

[Krizhevsky et al., 2012]

- Computa **f(x) = max(0,x)**
- Não fica saturada na região positiva
- É muito eficiente computacionalmente
- Converge mais rapidamente que sigmoide/tanh sobre imagens (≈ 6x)
- Sua saída não é centrada em zero
- Não é adequada para funções lógicas
- Não é usada no controle de redes recorrentes
- Apresenta uma inconveniência : qual o gradiente quando x < 0?

Função de Ativação – ReLU



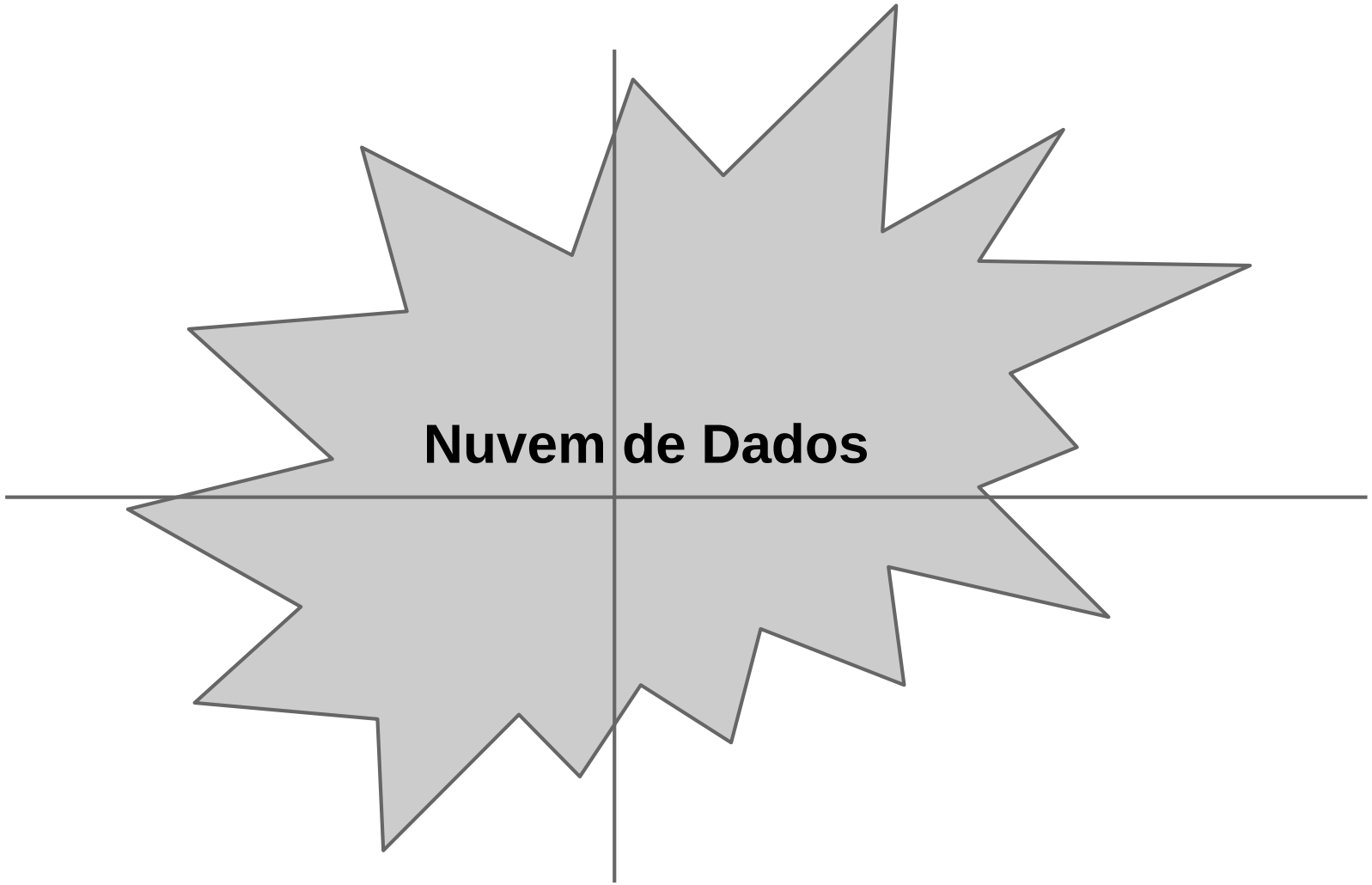


O que acontece quando x = -10?

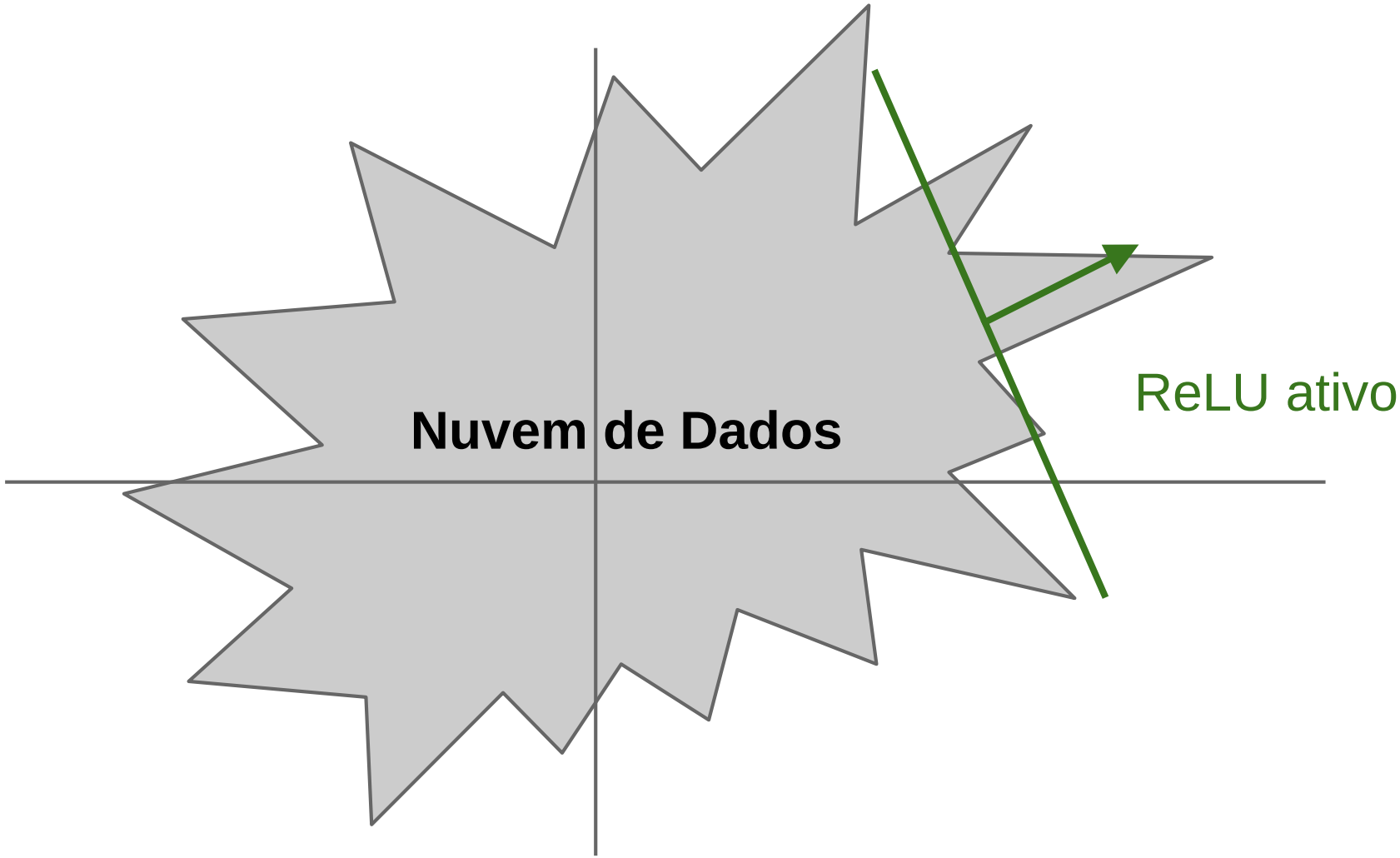
O que acontece quando x = 0?

O que acontece quando x = 10?

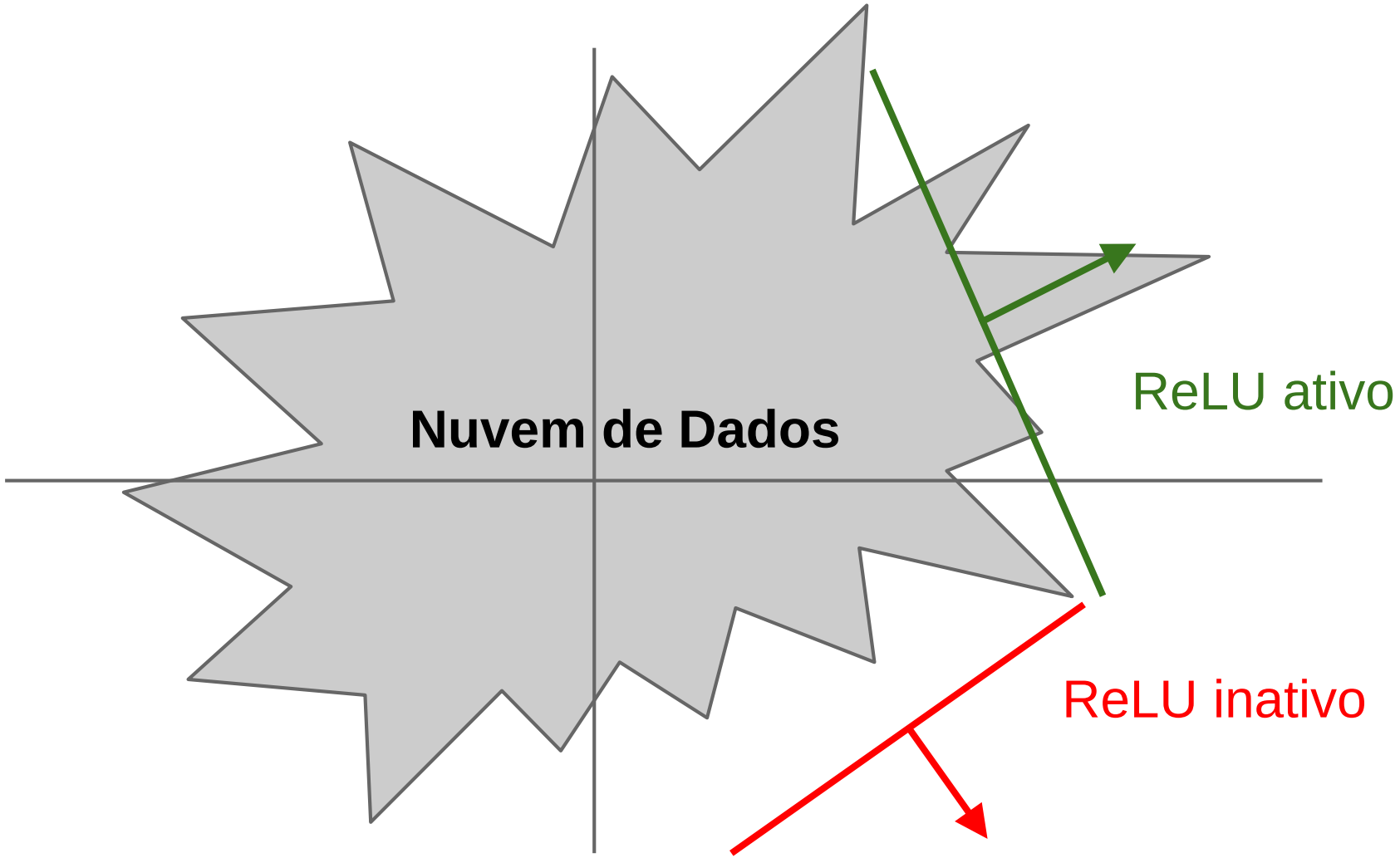
Função de Ativação – ReLU



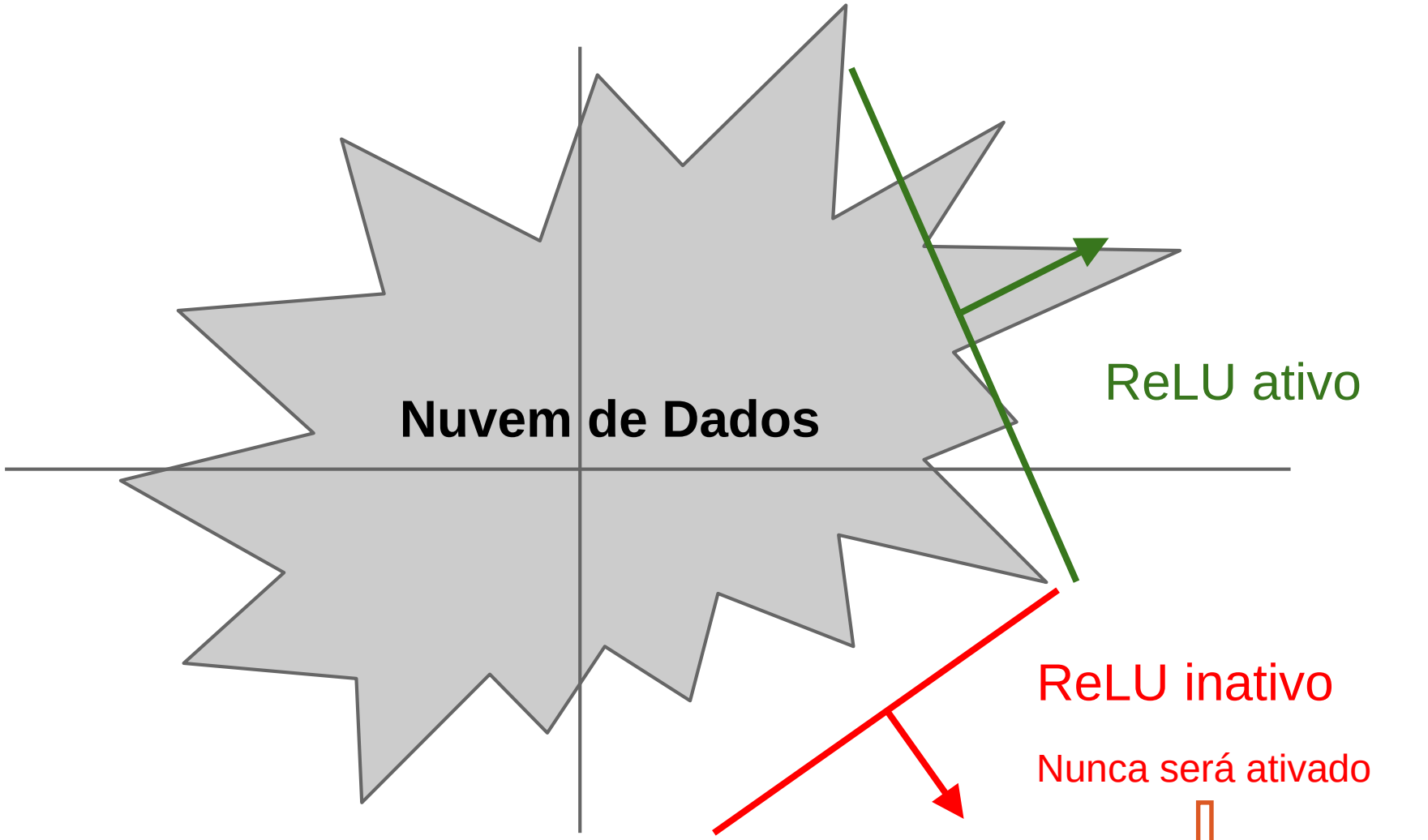
Função de Ativação – ReLU



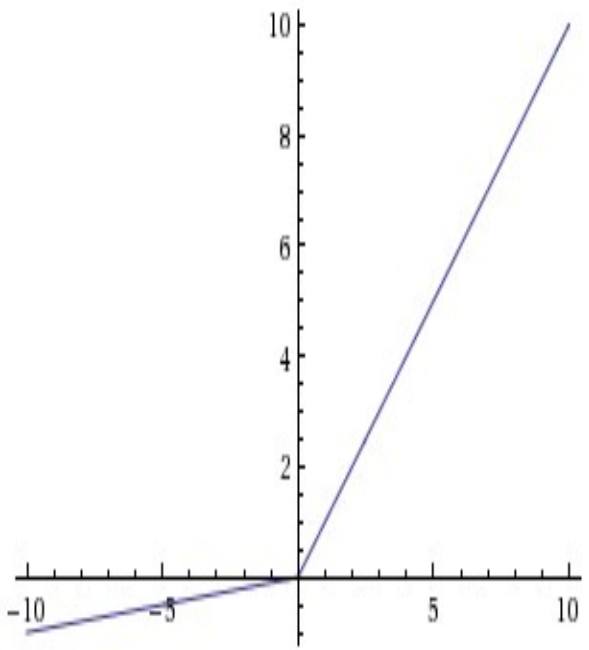
Função de Ativação – ReLU



Função de Ativação – ReLU



Função de Ativação – Leaky ReLU



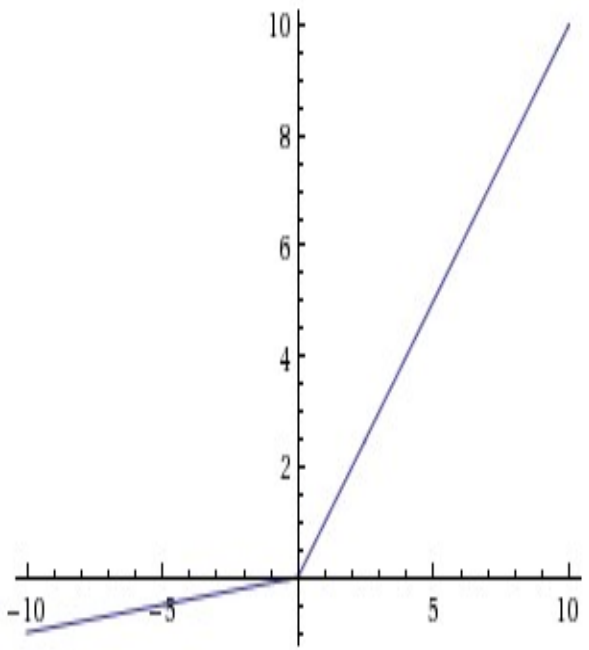
Leaky ReLU

$$f(x) = \max(0.01x, x)$$

[Mass et al., 2013]
[He et al., 2015]

- Não satura nunca
- É computacionalmente eficiente
- Converge mais rapidamente que sigmoide/tanh sobre imagens (≈ 6x)
- **Nunca "mata" os gradientes**

Função de Ativação – Leaky ReLU



- Não satura nunca
- É computacionalmente eficiente
- Converge mais rapidamente que sigmoide/tanh sobre imagens (≈ 6x)
- **Nunca "mata" os gradientes**

Leaky ReLU

$$f(x) = \max(0.01x, x)$$

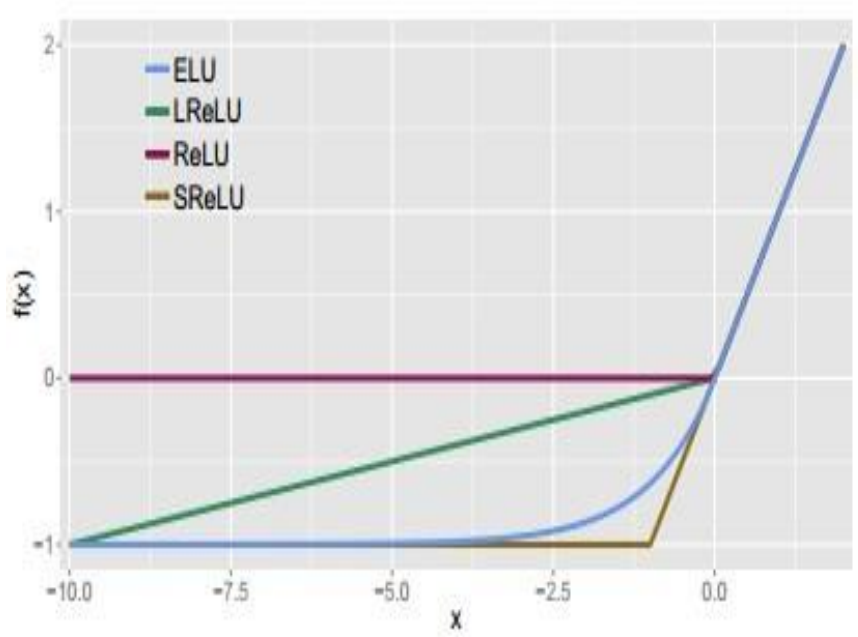
[Mass et al., 2013]
[He et al., 2015]

PReLU (*Parametric Rectifier Linear Unit*)

$$f(x) = \max(\alpha x, x)$$

BackProp sobre α (parâmetro)

Função de Ativação – ELU



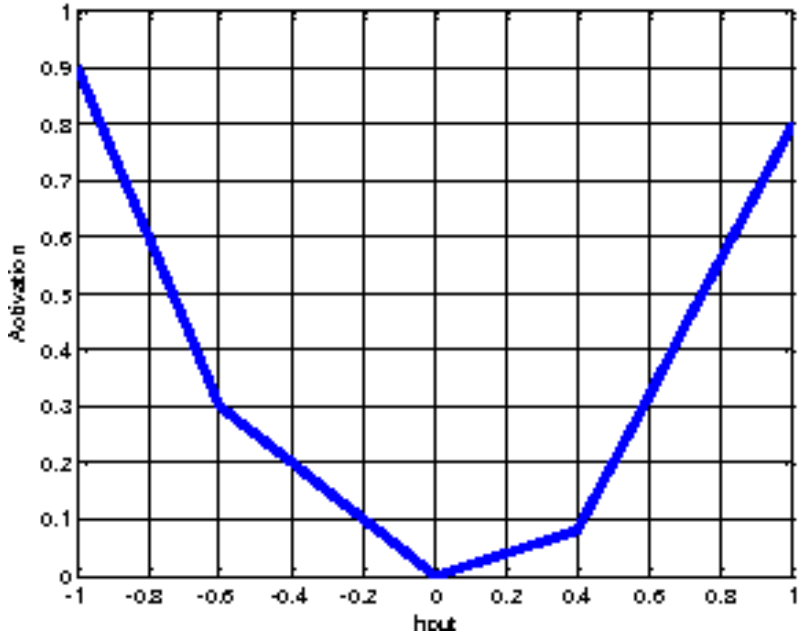
ELU
(*Exponential Linear Units*)

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha\left(\exp(x) - 1\right) & \text{if } x \leq 0 \end{cases}$$

[Clevert et al., 2015]

- Apresenta todos os benefícios de ReLU
- Não “mata” os gradientes
- Produz saídas com médias próximas de zero
- Necessita de exp() para seu cálculo

Função de Ativação – Maxout



$$\max(w_1^T x + b_1, w_2^T x + b_2)$$

[Goodfellow et al., 2013]

- Não possui a forma básica de produto interno seguido por não-linearidade
- Generaliza ReLU e Leaky ReLU
- Apresenta um regime linear! Nunca satura! Nunca “mata” os gradientes!

Problema: aumenta o número de parâmetros por neurônio :(