Tema 5 – Redes de Neuronas Profundas

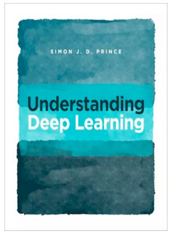
Aprendizaje Automático II - Grado en Inteligencia Artificial
Universidad Rey Juan Carlos

Iván Ramírez Díaz
ivan.ramirez@urjc.es

José Miguel Buenaposada Biencinto
josemiguel.buenaposada@urjc.es

Bibliografía

- Understanding Deep Learning. Capítulo 4.



- Deep Learning: CS 182 2021. Lecture 7.
  Sergey Levine. UC Berkeley.
  Curso en youtube

Redes de Neuronas Profundas

- Redes con más de una capa oculta
- La intuición sobre su funcionamiento es más complicada

5.1 Redes de Neuronas Profundas

- Componer dos redes
- Combinar las dos redes en una
- Hiperparámetros
- Notación para redes profundas
- Redes no profundas vs profundas

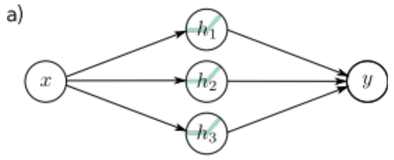
5.1 Redes de Neuronas Profundas

- Componer dos redes
- Combinar las dos redes en una
- Hiperparámetros
- Notación para redes profundas
- Redes no profundas vs profundas

Componer dos redes (shallow)

Red 1:

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x] \qquad y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$



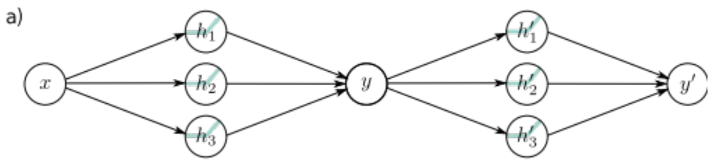
Componer dos redes (shallow)

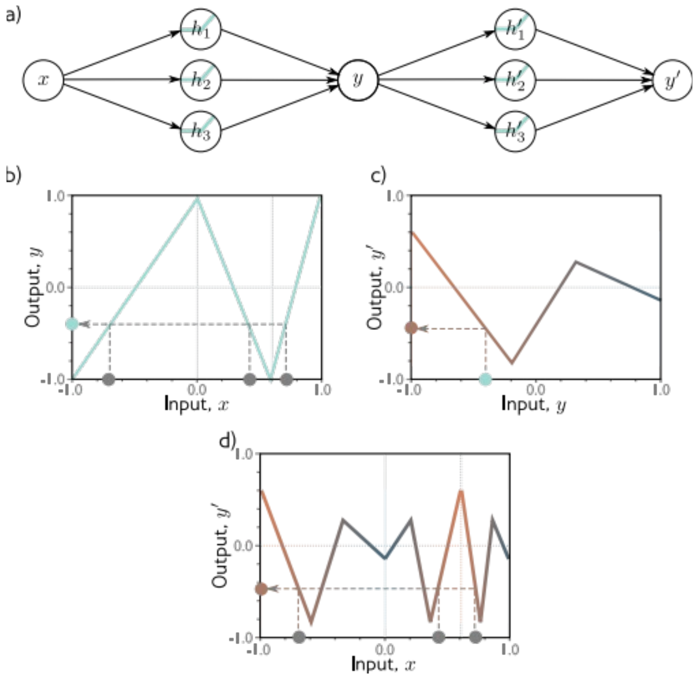
Red 1:

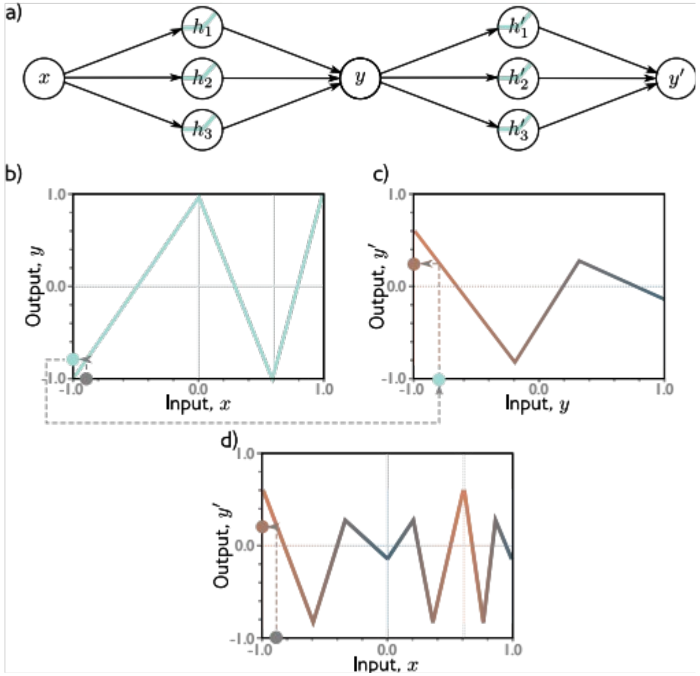
$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x] \qquad y = \phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$

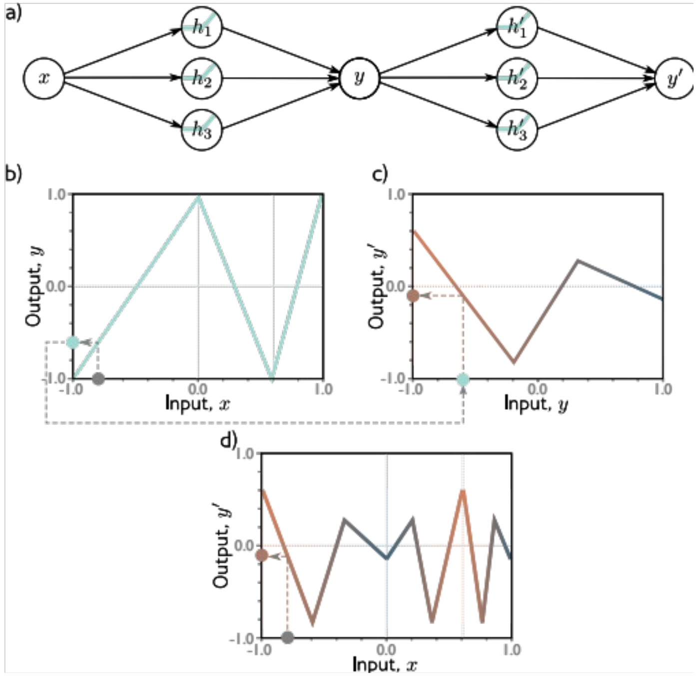
Red 2:

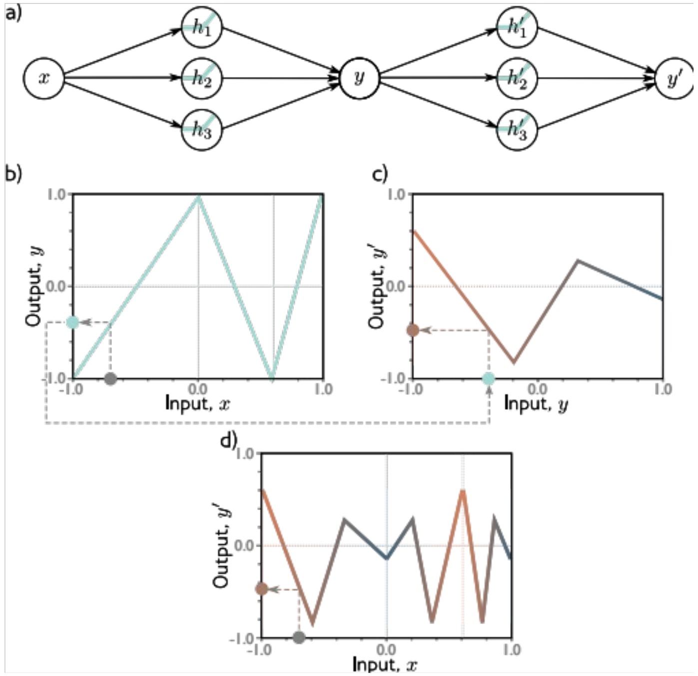
$$h'_1 = \mathrm{a}[\theta'_{10} + \theta'_{11}y]$$
$$h'_2 = \mathrm{a}[\theta'_{20} + \theta'_{21}y] \qquad y' = \phi'_0 + \phi'_1 h'_1 + \phi'_2 h'_2 + \phi'_3 h'_3$$
$$h'_3 = \mathrm{a}[\theta'_{30} + \theta'_{31}y]$$

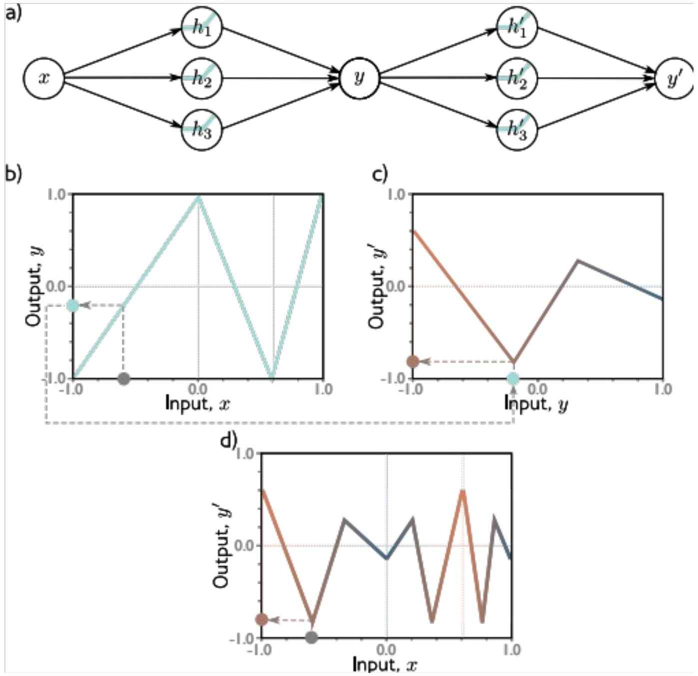


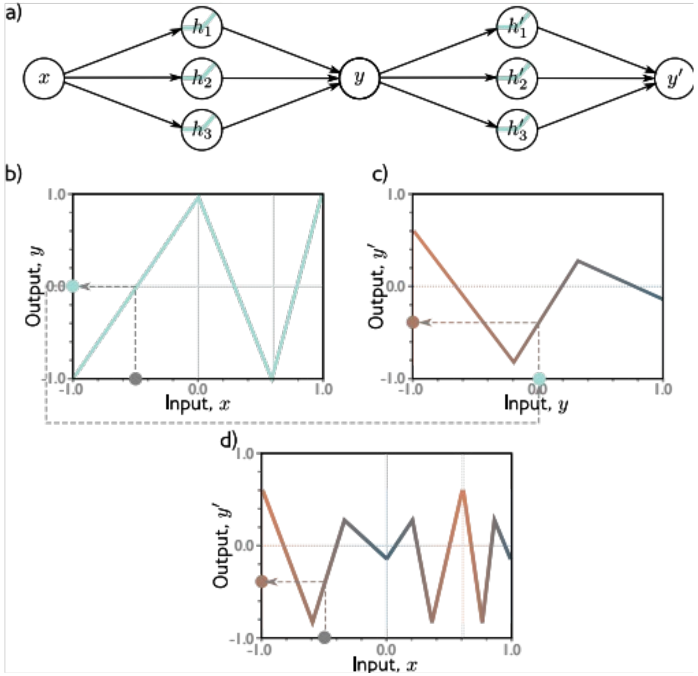


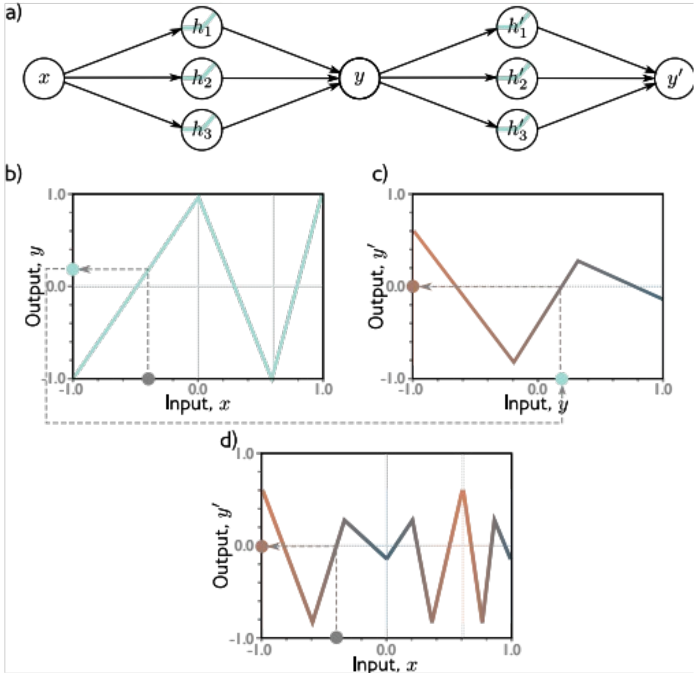

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x
c)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, y
d)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x

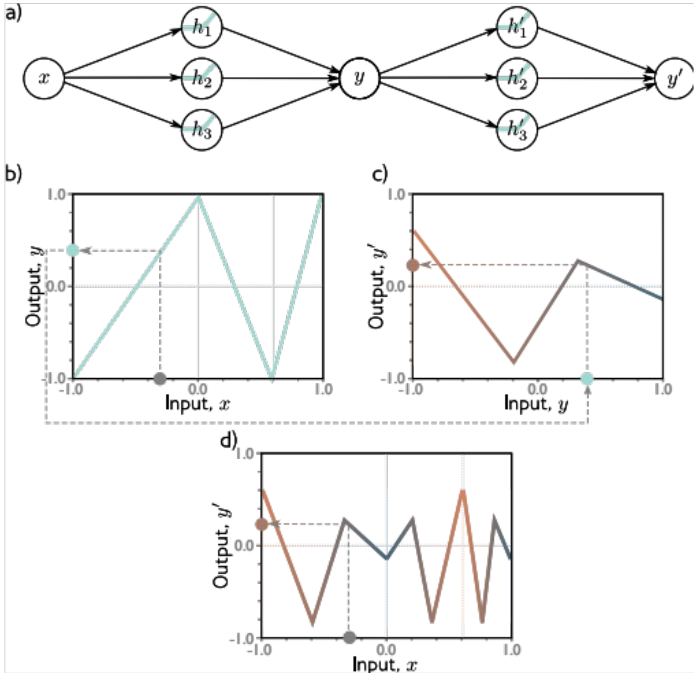

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x
c)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, y
d)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x

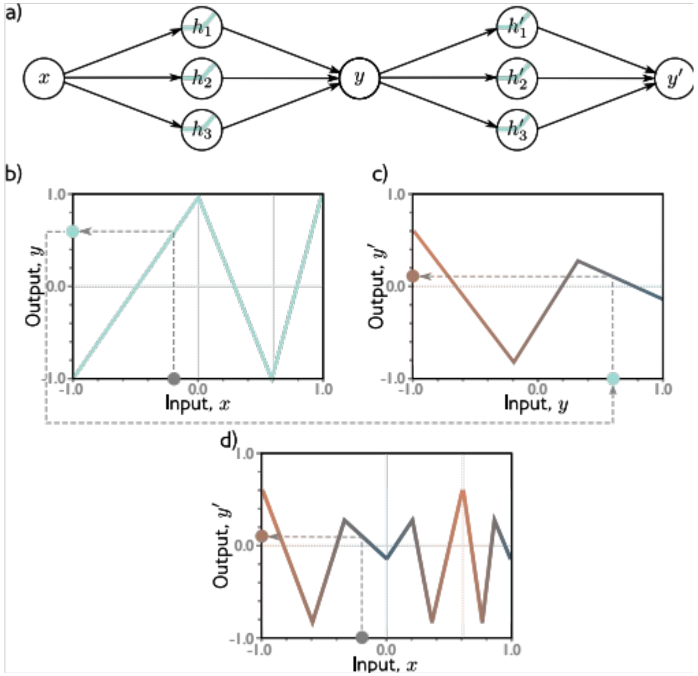

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x
c)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, y
d)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x

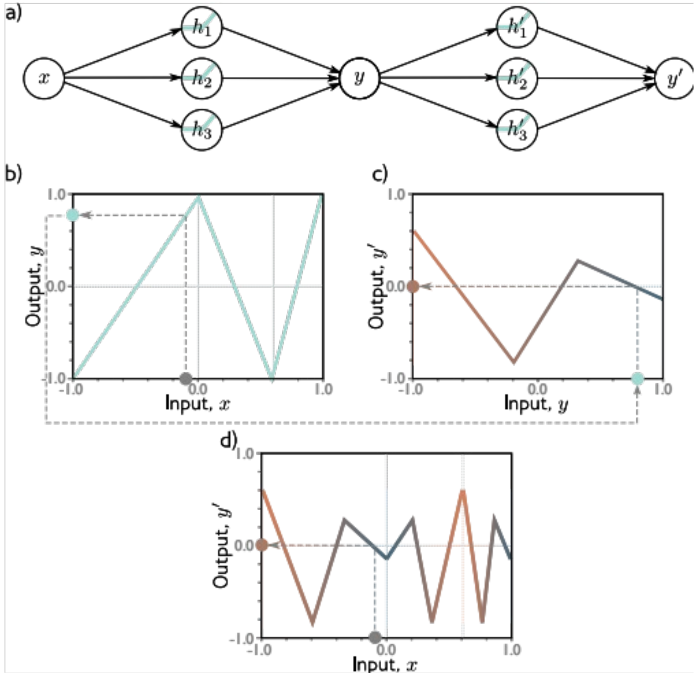

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x
c)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, y
d)
Output, y'
1.0
0.0
-1.0
-1.0
0.0
1.0
Input, x

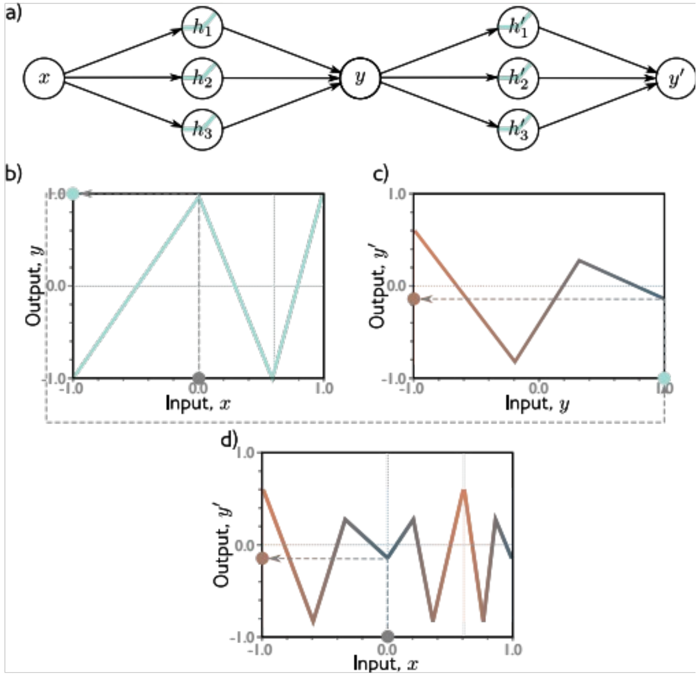

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x

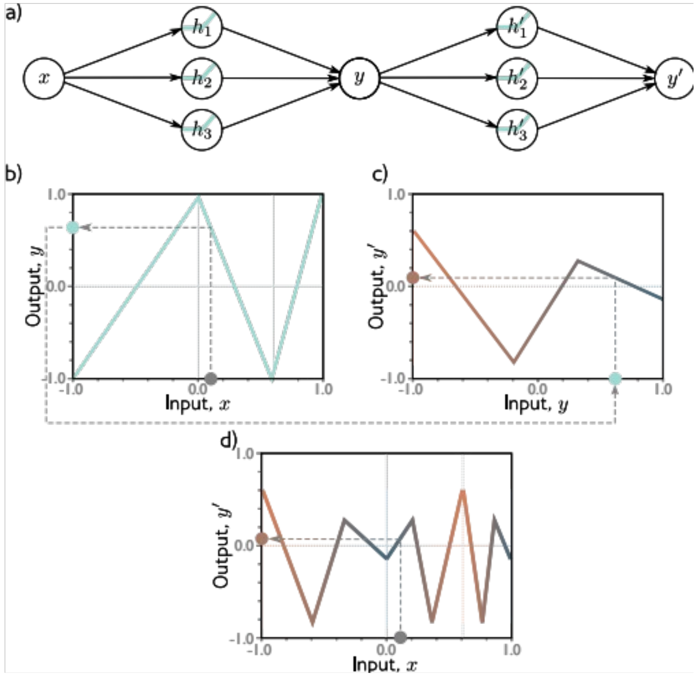

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x

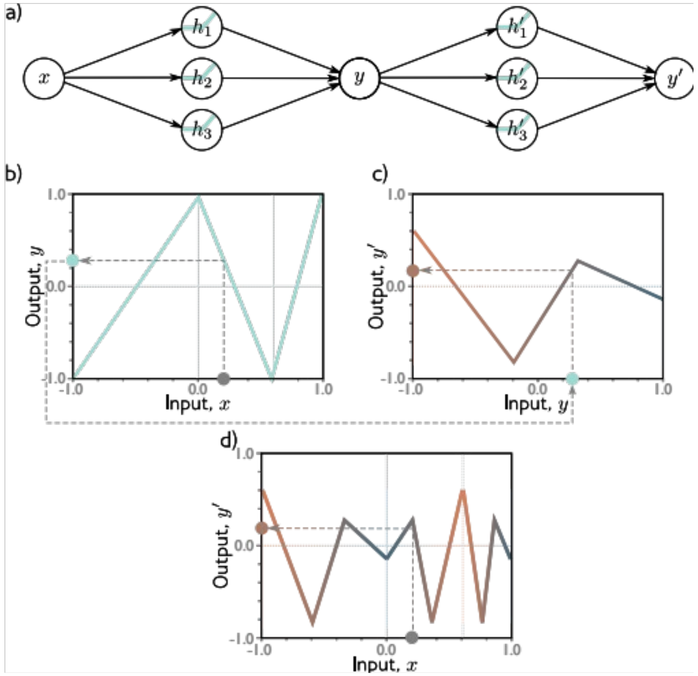

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x

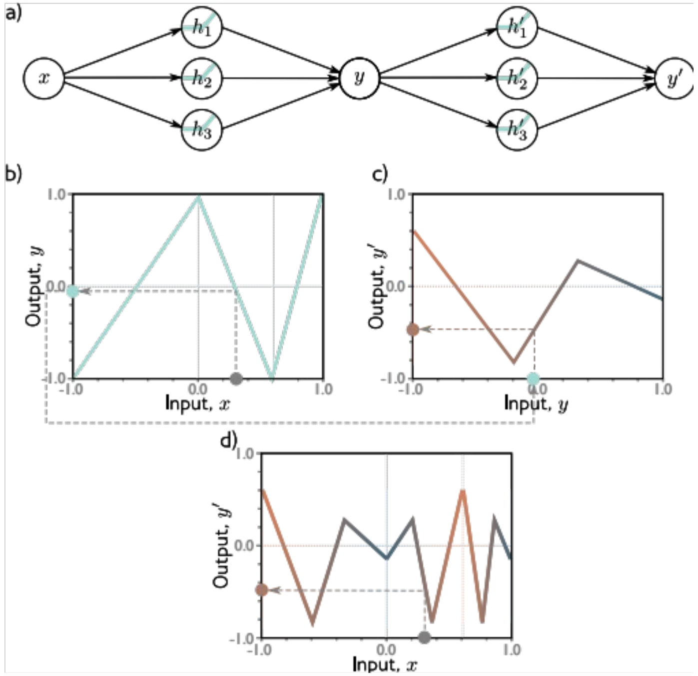

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x

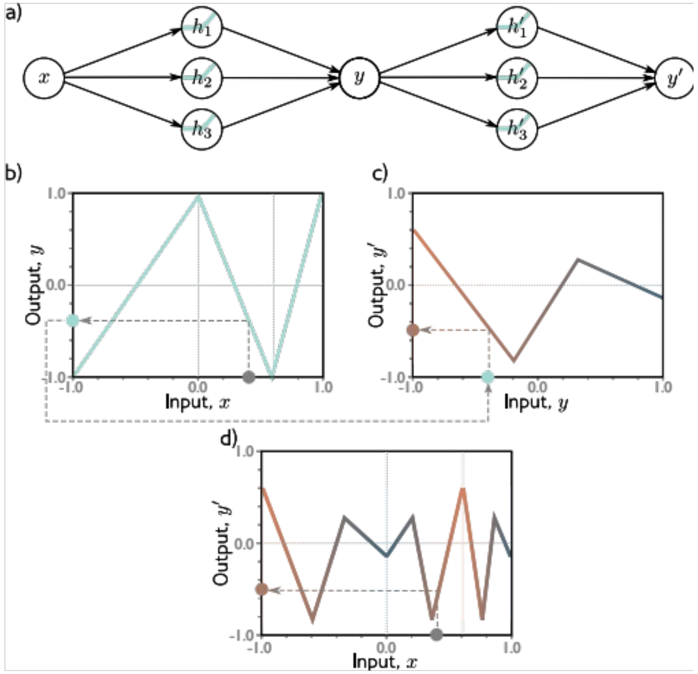

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0
c)
Output, y'
Input, y
1.0
0.0
-1.0
-1.0
0.0
1.0
d)
Output, y'
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0

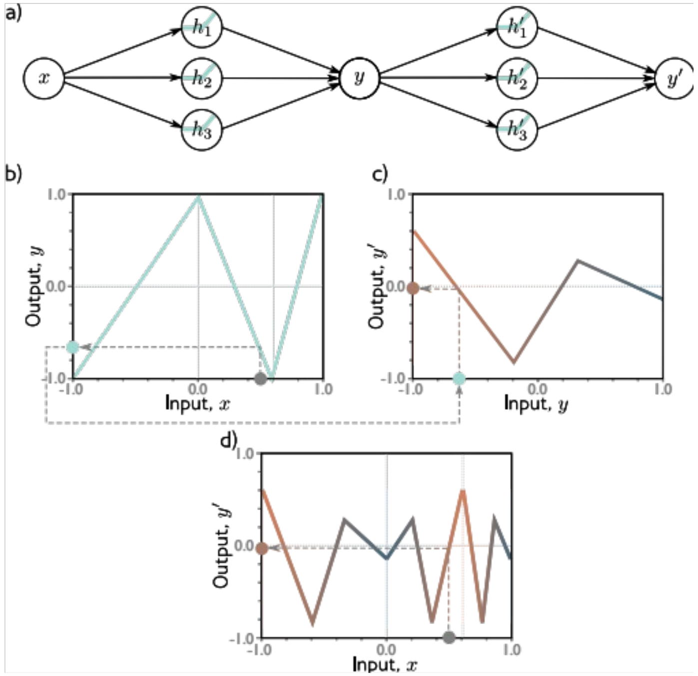

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0
c)
Output, y'
Input, y
1.0
0.0
-1.0
-1.0
0.0
1.0
d)
Output, y'
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0

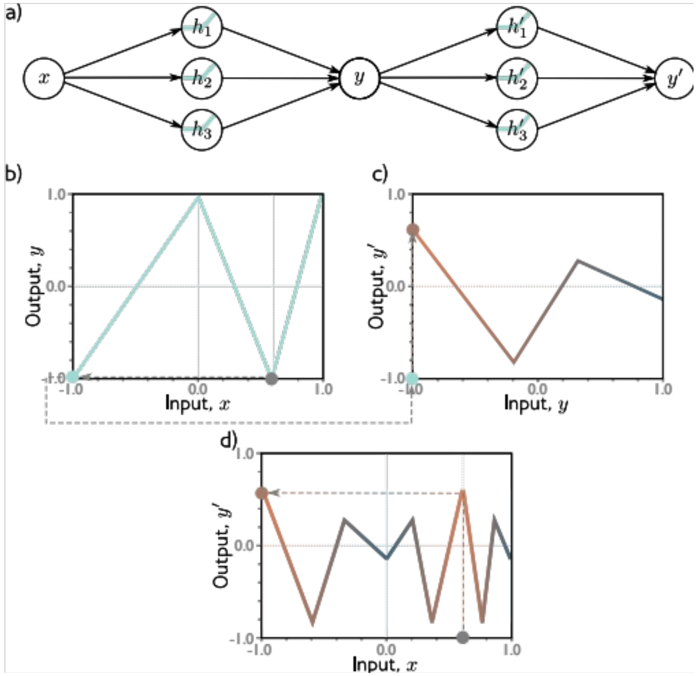

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0
c)
Output, y'
Input, y
1.0
0.0
-1.0
-1.0
0.0
1.0
d)
Output, y'
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0

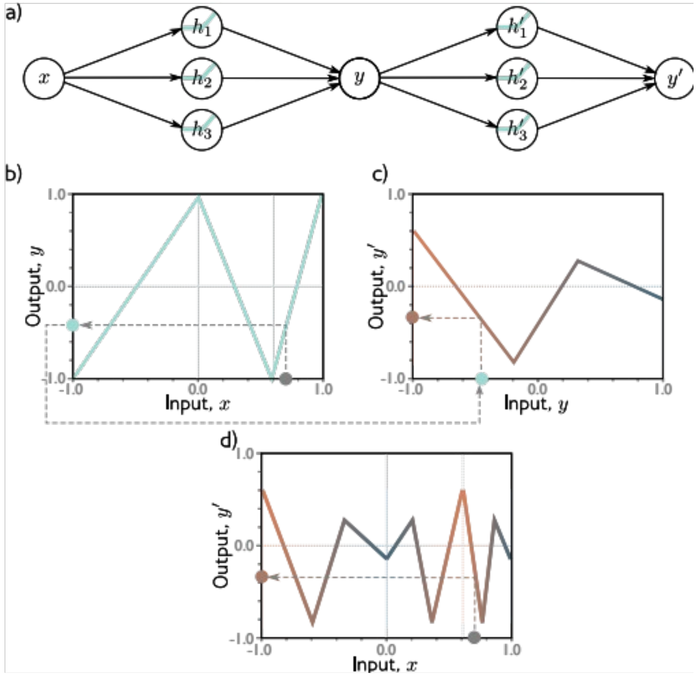

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0
c)
Output, y'
Input, y
1.0
0.0
-1.0
-1.0
0.0
1.0
d)
Output, y'
Input, x
1.0
0.0
-1.0
-1.0
0.0
1.0

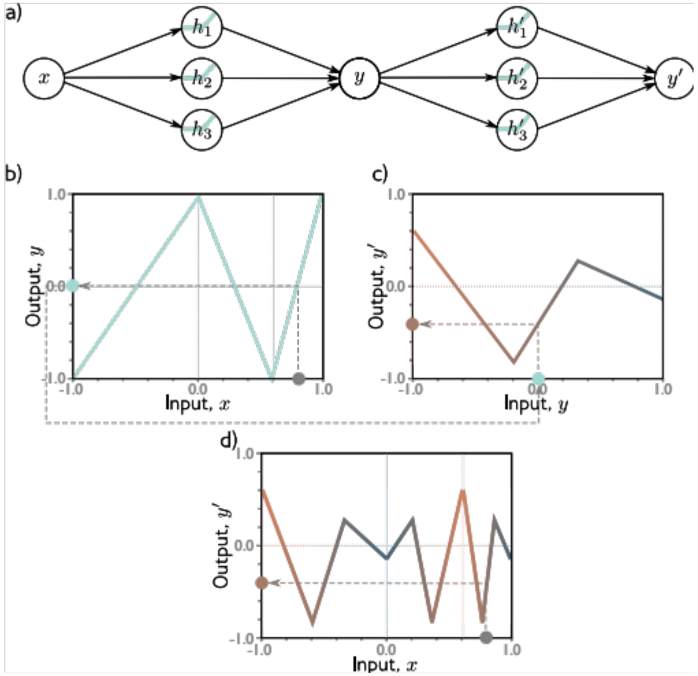

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x

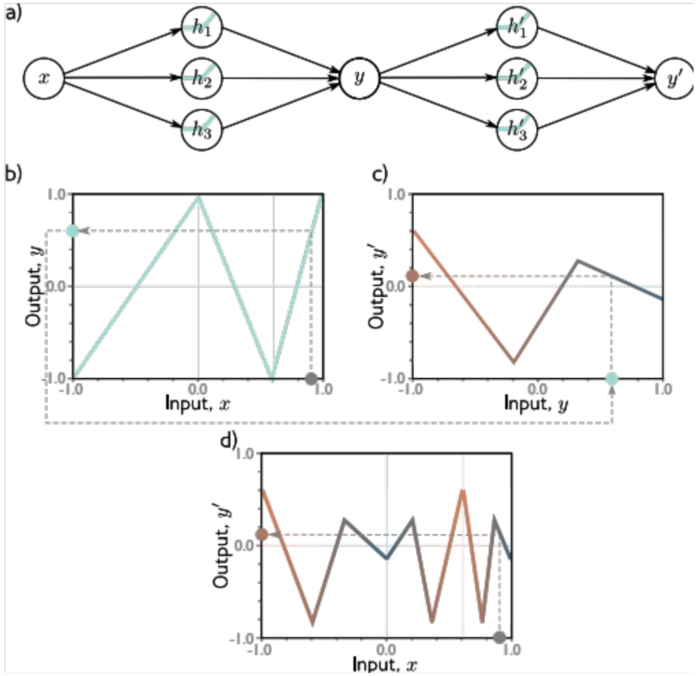

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x


a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x

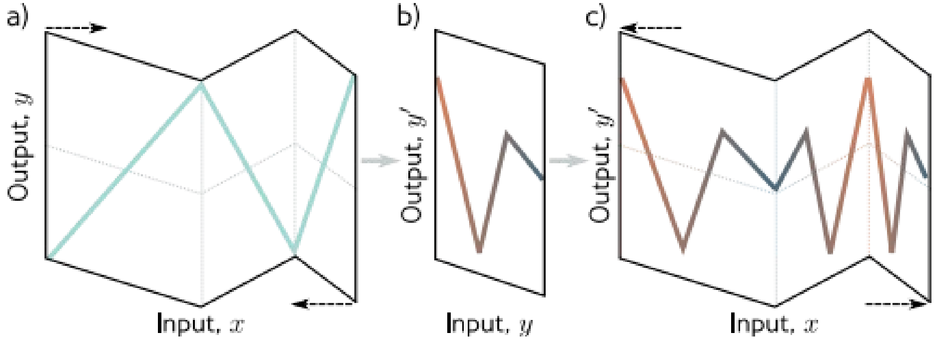

a)
x
h_1
h_2
h_3
y
h'_1
h'_2
h'_3
y'
b)
Output, y
Input, x
1.0
0.0
-1.0
c)
Output, y'
Input, y
d)
Output, y'
Input, x



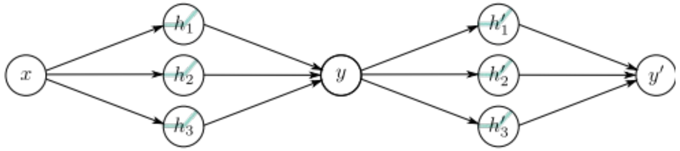




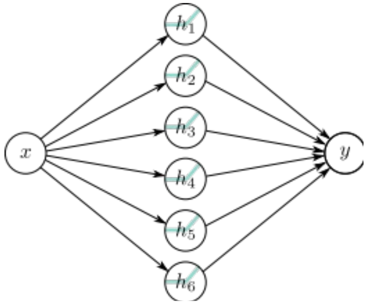
“Analogía del Plegado”



Comparación con una red de 6 unidades ocultas

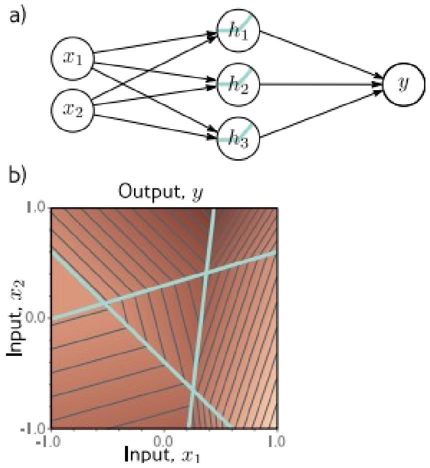


- 20 parámetros
- (al menos) 9 regiones

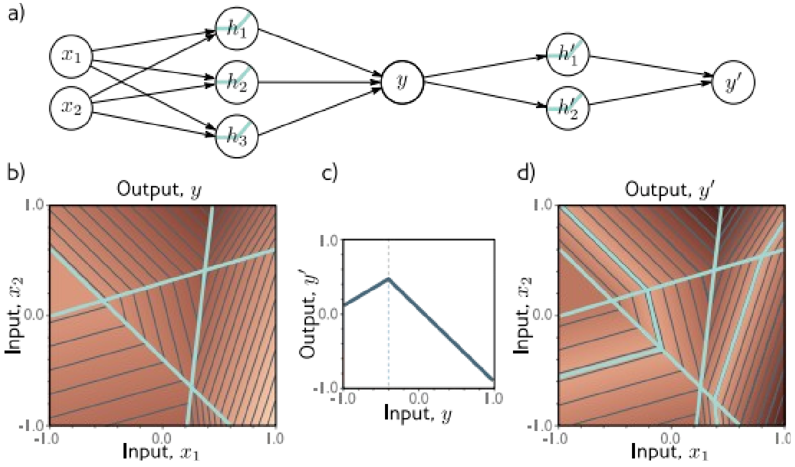


- 19 parámetros
- Máximo 7 regiones

Composición de redes en 2D



Composición de redes en 2D



5.1 Redes de Neuronas Profundas

- Componer dos redes
- Combinar las dos redes en una
- Hiperparámetros
- Notación para redes profundas
- Redes no profundas vs profundas

Componer dos redes (shallow) en una

Red 1:

$$
\begin{aligned}
h_1 &= \mathrm{a}[\theta_{10} + \theta_{11} x] \\
h_2 &= \mathrm{a}[\theta_{20} + \theta_{21} x] \\
h_3 &= \mathrm{a}[\theta_{30} + \theta_{31} x]
\end{aligned}
\qquad y = \boxed{\phi_0 + \phi_1 h_1 + \phi_2 h_2 + \phi_3 h_3}
$$

Red 2:

$$
\begin{aligned}
h'_1 &= \mathrm{a}[\theta'_{10} + \boxed{\theta'_{11} y}] \\
h'_2 &= \mathrm{a}[\theta'_{20} + \theta'_{21} y] \\
h'_3 &= \mathrm{a}[\theta'_{30} + \theta'_{31} y]
\end{aligned}
\qquad y' = \phi'_0 + \phi'_1 h'_1 + \phi'_2 h'_2 + \phi'_3 h'_3
$$

Las unidades ocultas de la segunda red en términos de la primera red:

$$
\begin{aligned}
h'_1 = &\quad \mathrm{a}[\theta'_{10} + \theta'_{11} y] &= &\quad \mathrm{a}[\theta'_{10} + \boxed{\theta'_{11}\phi_0 + \theta'_{11}\phi_1 h_1 + \theta'_{11}\phi_2 h_2 + \theta'_{11}\phi_3 h_3}] \\
h'_2 = &\quad \mathrm{a}[\theta'_{20} + \theta'_{21} y] &= &\quad \mathrm{a}[\theta'_{20} + \theta'_{21}\phi_0 + \theta'_{21}\phi_1 h_1 + \theta'_{21}\phi_2 h_2 + \theta'_{21}\phi_3 h_3] \\
h'_3 = &\quad \mathrm{a}[\theta'_{30} + \theta'_{31} y] &= &\quad \mathrm{a}[\theta'_{30} + \theta'_{31}\phi_0 + \theta'_{31}\phi_1 h_1 + \theta'_{31}\phi_2 h_2 + \theta'_{31}\phi_3 h_3]
\end{aligned}
$$

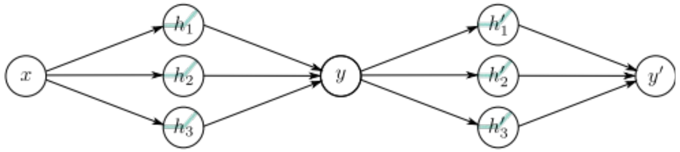
Componer dos redes (shallow) en una

$$
\begin{aligned}
h'_1 = &\quad \mathrm{a}[\theta'_{10} + \theta'_{11} y] &= &\quad \mathrm{a}[\theta'_{10} + \theta'_{11}\phi_0 + \theta'_{11}\phi_1 h_1 + \theta'_{11}\phi_2 h_2 + \theta'_{11}\phi_3 h_3] \\
h'_2 = &\quad \mathrm{a}[\theta'_{20} + \theta'_{21} y] &= &\quad \mathrm{a}[\theta'_{20} + \theta'_{21}\phi_0 + \theta'_{21}\phi_1 h_1 + \theta'_{21}\phi_2 h_2 + \theta'_{21}\phi_3 h_3] \\
h'_3 = &\quad \mathrm{a}[\theta'_{30} + \theta'_{31} y] &= &\quad \mathrm{a}[\boxed{\theta'_{30} + \theta'_{31}\phi_0} + \boxed{\theta'_{31}\phi_1} h_1 + \boxed{\theta'_{31}\phi_2} h_2 + \boxed{\theta'_{31}\phi_3} h_3]
\end{aligned}
$$

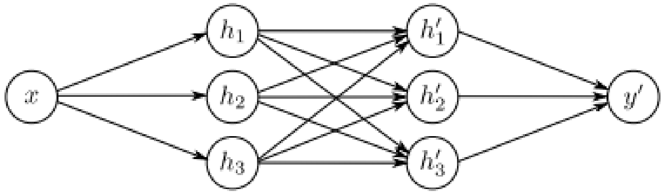
Renombrando parámetros:

$$
\begin{aligned}
h'_1 &= \mathrm{a}[\psi_{10} + \psi_{11} h_1 + \psi_{12} h_2 + \psi_{13} h_3] \\
h'_2 &= \mathrm{a}[\psi_{20} + \psi_{21} h_1 + \psi_{22} h_2 + \psi_{23} h_3] \\
h'_3 &= \mathrm{a}[\boxed{\psi_{30}} + \boxed{\psi_{31}} h_1 + \boxed{\psi_{32}} h_2 + \boxed{\psi_{33}} h_3]
\end{aligned}
$$

Componer dos redes (shallow) → una red de dos capas ocultas



La composición de dos redes con 1 salida y 1 entrada es un caso especial de una red con dos capas ocultas con ciertas restricciones en los pesos de la segunda capa oculta.



La red con 2 capas ocultas es más flexible que la composición de dos redes (shallow)

Componer dos redes (shallow) en una

$$
\begin{aligned}
h'_1 = &\quad \mathrm{a}[\theta'_{10} + \theta'_{11} y] &= &\quad \mathrm{a}[\theta'_{10} + \theta'_{11}\phi_0 + \theta'_{11}\phi_1 h_1 + \theta'_{11}\phi_2 h_2 + \theta'_{11}\phi_3 h_3] \\
h'_2 = &\quad \mathrm{a}[\theta'_{20} + \theta'_{21} y] &= &\quad \mathrm{a}[\theta'_{20} + \theta'_{21}\phi_0 + \theta'_{21}\phi_1 h_1 + \theta'_{21}\phi_2 h_2 + \theta'_{21}\phi_3 h_3] \\
h'_3 = &\quad \mathrm{a}[\theta'_{30} + \theta'_{31} y] &= &\quad \mathrm{a}[\boxed{\theta'_{30} + \theta'_{31}\phi_0} + \boxed{\theta'_{31}\phi_1} h_1 + \boxed{\theta'_{31}\phi_2} h_2 + \boxed{\theta'_{31}\phi_3} h_3]
\end{aligned}
$$

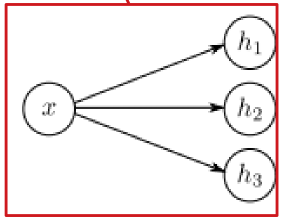
Restricciones sobre los pesos, *ψ*, de la segunda capa oculta de una red con 2 capas ocultas equivalente (***ψ = θ·Φ***, *producto escalar).*

Renombrando parámetros:

$$
\begin{aligned}
h'_1 &= \mathrm{a}[\psi_{10} + \psi_{11} h_1 + \psi_{12} h_2 + \psi_{13} h_3] \\
h'_2 &= \mathrm{a}[\psi_{20} + \psi_{21} h_1 + \psi_{22} h_2 + \psi_{23} h_3] \\
h'_3 &= \mathrm{a}[\boxed{\psi_{30}} + \boxed{\psi_{31}} h_1 + \boxed{\psi_{32}} h_2 + \boxed{\psi_{33}} h_3]
\end{aligned}
$$

Una red de dos capas ocultas

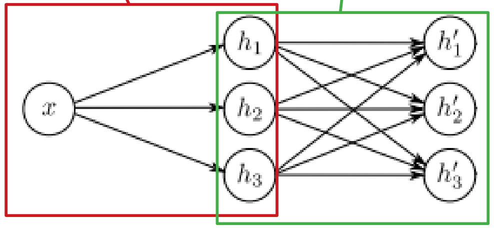
$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$



Una red de dos capas ocultas

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$

$$h'_1 = \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3]$$
$$h'_2 = \mathrm{a}[\psi_{20} + \psi_{21}h_2 + \psi_{22}h_2 + \psi_{23}h_3]$$
$$h'_3 = \mathrm{a}[\psi_{30} + \psi_{31}h_2 + \psi_{32}h_2 + \psi_{33}h_3]$$

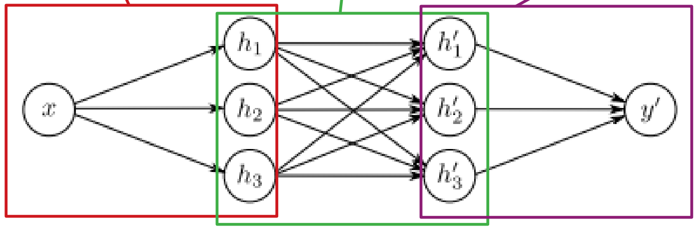


Una red de dos capas ocultas

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$

$$h'_1 = \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3]$$
$$h'_2 = \mathrm{a}[\psi_{20} + \psi_{21}h_2 + \psi_{22}h_2 + \psi_{23}h_3]$$
$$h'_3 = \mathrm{a}[\psi_{30} + \psi_{31}h_2 + \psi_{32}h_2 + \psi_{33}h_3]$$

$$y' = \phi'_0 + \phi'_1 h'_1 + \phi'_2 h'_2 + \phi'_3 h'_3$$



Una red de dos capas ocultas

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x]$$

$$h'_1 = \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3]$$
$$h'_2 = \mathrm{a}[\psi_{20} + \psi_{21}h_2 + \psi_{22}h_2 + \psi_{23}h_3]$$
$$h'_3 = \mathrm{a}[\psi_{30} + \psi_{31}h_2 + \psi_{32}h_2 + \psi_{33}h_3]$$

$$y' = \phi'_0 + \phi'_1 h'_1 + \phi'_2 h'_2 + \phi'_3 h'_3$$

Y en una ecuación tenemos:

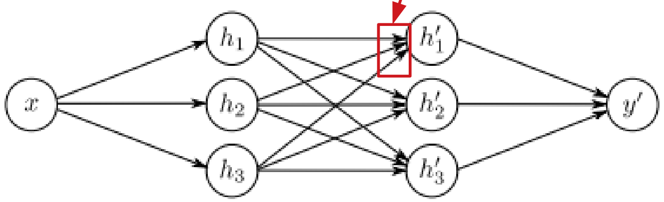
$$\begin{aligned} y' = \phi'_0 &+ \phi'_1 \mathrm{a}\left[\psi_{10} + \psi_{11}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{12}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{13}\mathrm{a}[\theta_{30} + \theta_{31}x]\right] \\ &+ \phi'_2 \mathrm{a}[\psi_{20} + \psi_{21}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{22}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{23}\mathrm{a}[\theta_{30} + \theta_{31}x]] \\ &+ \phi'_3 \mathrm{a}[\psi_{30} + \psi_{31}\mathrm{a}[\theta_{10} + \theta_{11}x] + \psi_{32}\mathrm{a}[\theta_{20} + \theta_{21}x] + \psi_{33}\mathrm{a}[\theta_{30} + \theta_{31}x]] \end{aligned}$$

Esta ecuación relaciona valores de x con valores de y’

Redes como composición de funciones

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x] \qquad h'_1 = \mathrm{a}[\boxed{\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3}]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x] \qquad h'_2 = \mathrm{a}[\psi_{20} + \psi_{21}h_2 + \psi_{22}h_2 + \psi_{23}h_3]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x] \qquad h'_3 = \mathrm{a}[\psi_{30} + \psi_{31}h_2 + \psi_{32}h_2 + \psi_{33}h_3]$$

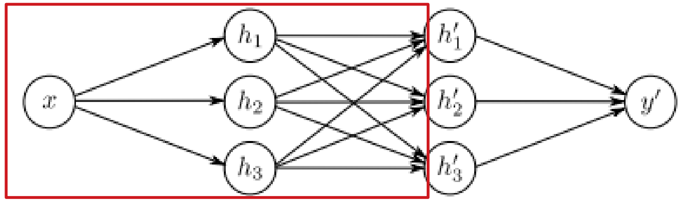
- Consideremos las preactivaciones en la segunda capa oculta



Redes como composición de funciones

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x] \qquad h'_1 = \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x] \qquad h'_2 = \mathrm{a}[\psi_{20} + \psi_{21}h_2 + \psi_{22}h_2 + \psi_{23}h_3]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x] \qquad h'_3 = \mathrm{a}[\psi_{30} + \psi_{31}h_2 + \psi_{32}h_2 + \psi_{33}h_3]$$

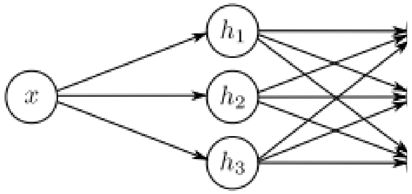
- Consideremos las preactivaciones en la segunda capa oculta
- En ese punto de la red, tenemos una red de una capa oculta con 3 salidas:



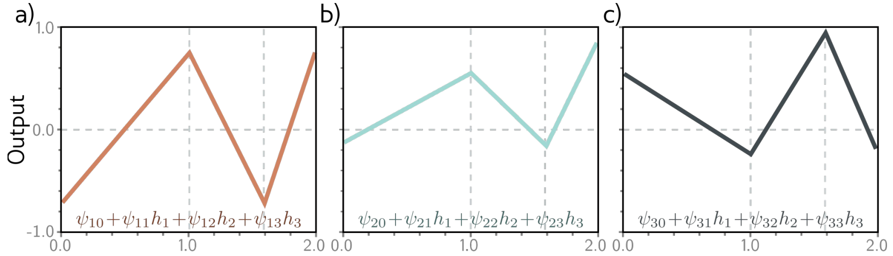
Redes como composición de funciones

$$h_1 = \mathrm{a}[\theta_{10} + \theta_{11}x] \qquad h'_1 = \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3]$$
$$h_2 = \mathrm{a}[\theta_{20} + \theta_{21}x] \qquad h'_2 = \mathrm{a}[\psi_{20} + \psi_{21}h_2 + \psi_{22}h_2 + \psi_{23}h_3]$$
$$h_3 = \mathrm{a}[\theta_{30} + \theta_{31}x] \qquad h'_3 = \mathrm{a}[\psi_{30} + \psi_{31}h_2 + \psi_{32}h_2 + \psi_{33}h_3]$$

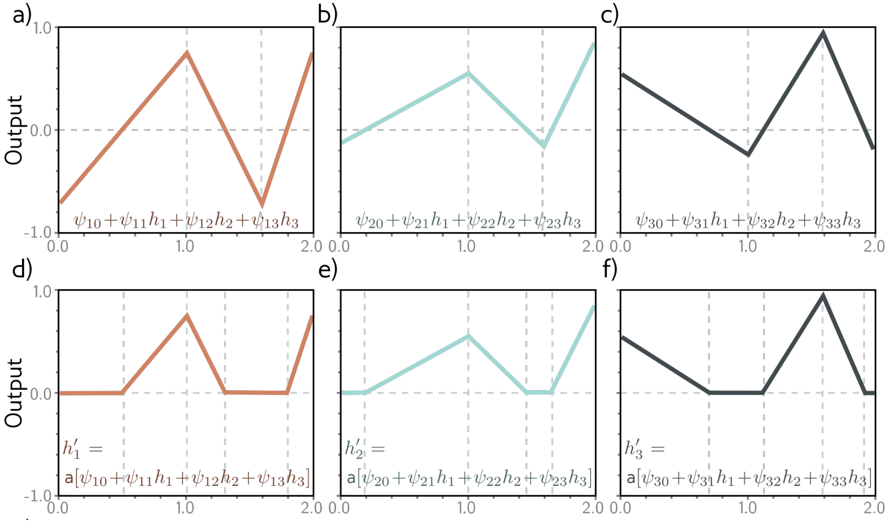
- Consideremos las preactivaciones en la segunda capa oculta
- En ese punto de la red, tenemos una red de una capa oculta con 3 salidas:



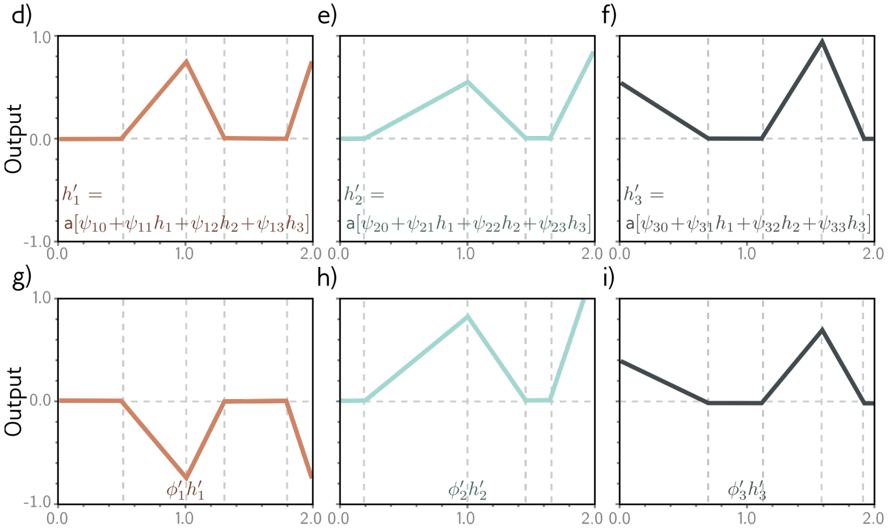
Redes como composición de funciones



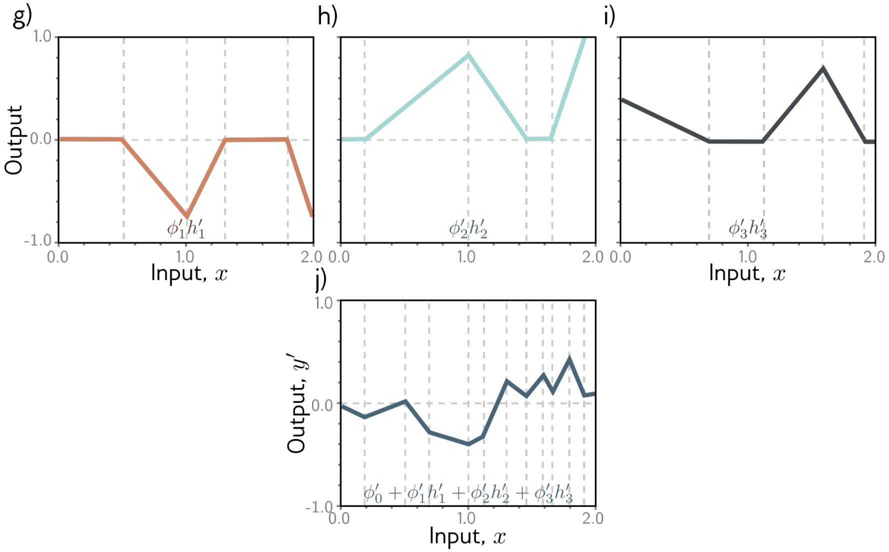
Redes como composición de funciones



Redes como composición de funciones



Redes como composición de funciones



5.1 Redes de Neuronas Profundas

- Componer dos redes
- Combinar las dos redes en una
- Hiperparámetros
- Notación para redes profundas
- Redes no profundas vs profundas

Hiperparámetros de la red

- K capas = profundidad de la red
- Unidades ocultas por capa = anchura de la red

- Estos son los hiperparámetros – elegidos antes de entrenar
- Se puede reentrenar con diferentes hiperparámetros – optimización de hiperparámetros o búsqueda de hiperparámetros

5.1 Redes de Neuronas Profundas

- Componer dos redes
- Combinar las dos redes en una
- Hiperparámetros
- Notación para redes profundas
- Redes no profundas vs profundas

Función de activación

$$\begin{aligned} h_1 &= \mathrm{a}[\theta_{10} + \theta_{11}x] \\ h_2 &= \mathrm{a}[\theta_{20} + \theta_{21}x] \\ h_3 &= \mathrm{a}[\theta_{30} + \theta_{31}x] \end{aligned} \quad \longrightarrow \quad \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \mathbf{a}\left[\begin{bmatrix} \theta_{10} \\ \theta_{20} \\ \theta_{30} \end{bmatrix} + \begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{31} \end{bmatrix} x\right]$$

$$\begin{aligned} h'_1 &= \mathrm{a}[\psi_{10} + \psi_{11}h_1 + \psi_{12}h_2 + \psi_{13}h_3 \\ h'_2 &= \mathrm{a}[\psi_{20} + \psi_{21}h_1 + \psi_{22}h_2 + \psi_{23}h_3 \\ h'_3 &= \mathrm{a}[\psi_{30} + \psi_{31}h_1 + \psi_{32}h_2 + \psi_{33}h_3 \end{aligned} \quad \longrightarrow \quad \begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \end{bmatrix} = \mathbf{a}\left[\begin{bmatrix} \psi_{10} \\ \psi_{20} \\ \psi_{30} \end{bmatrix} + \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}\right]$$

$$y' = \phi'_0 + \phi'_1 h'_1 + \phi'_2 h'_2 + \phi'_3 h'_3 \quad \longrightarrow \quad y' = \phi'_0 + \begin{bmatrix} \phi'_1 & \phi'_2 & \phi'_3 \end{bmatrix} \begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \end{bmatrix}$$

Vectores y matrices en las capas lineales

$$\begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix} = \mathbf{a}\left[\begin{bmatrix} \theta_{10} \\ \theta_{20} \\ \theta_{30} \end{bmatrix} + \begin{bmatrix} \theta_{11} \\ \theta_{21} \\ \theta_{31} \end{bmatrix} x\right] \quad \longrightarrow \quad \mathbf{h} = \mathbf{a}\left[\boldsymbol{\theta}_0 + \boldsymbol{\theta} x\right]$$

$$\begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \end{bmatrix} = \mathbf{a}\left[\begin{bmatrix} \psi_{10} \\ \psi_{20} \\ \psi_{30} \end{bmatrix} + \begin{bmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}\right] \quad \longrightarrow \quad \mathbf{h}' = \mathbf{a}\left[\boldsymbol{\psi}_0 + \boldsymbol{\Psi}\mathbf{h}\right]$$

$$y' = \phi'_0 + \begin{bmatrix} \phi'_1 & \phi'_2 & \phi'_3 \end{bmatrix} \begin{bmatrix} h'_1 \\ h'_2 \\ h'_3 \end{bmatrix} \quad \longrightarrow \quad y = \boldsymbol{\phi}'_0 + \boldsymbol{\phi}'\mathbf{h}'$$

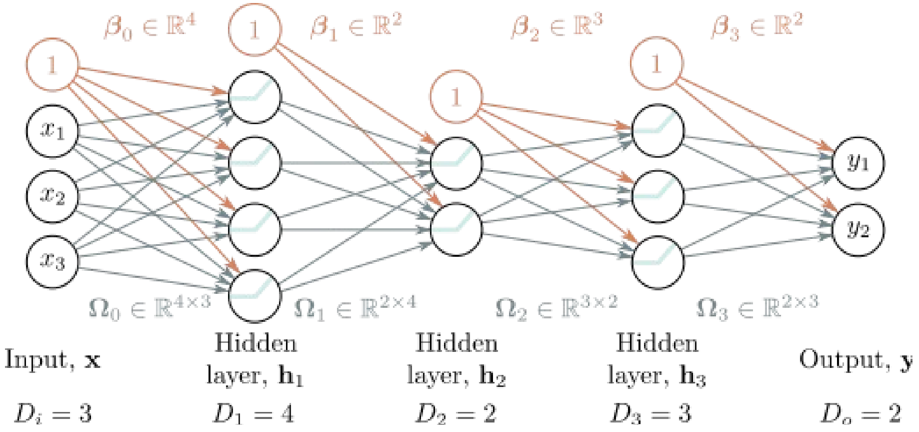
Estandarización de nombres de parámetros

$$\mathbf{h} = \mathbf{a}\left[\boldsymbol{\theta}_0 + \boldsymbol{\theta}x\right] \longrightarrow \mathbf{h}_1 = \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}]$$

$$\mathbf{h}' = \mathbf{a}\left[\boldsymbol{\psi}_0 + \boldsymbol{\Psi}\mathbf{h}\right] \longrightarrow \mathbf{h}_2 = \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1]$$

$$y = \boldsymbol{\phi}'_0 + \boldsymbol{\phi}'\mathbf{h}' \longrightarrow \mathbf{y} = \boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2$$

Ejemplo

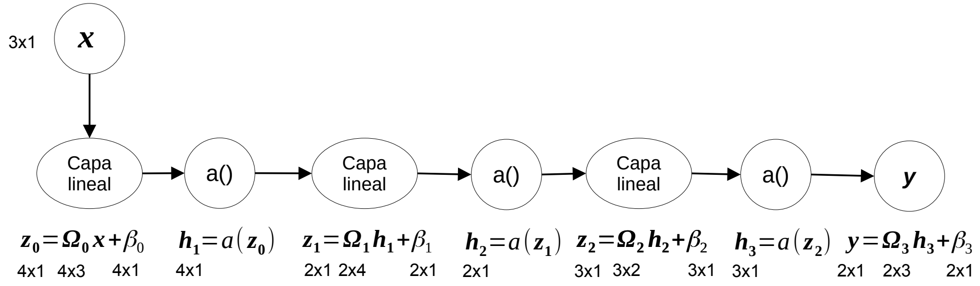


Ecuaciones generales para redes

$$\begin{aligned}
\mathbf{h}_1 &= \mathbf{a}[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}]\\
\mathbf{h}_2 &= \mathbf{a}[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{h}_1]\\
\mathbf{h}_3 &= \mathbf{a}[\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{h}_2]\\
&\vdots\\
\mathbf{h}_K &= \mathbf{a}[\boldsymbol{\beta}_{K-1} + \boldsymbol{\Omega}_{K-1}\mathbf{h}_{K-1}]\\
\mathbf{y} &= \boldsymbol{\beta}_K + \boldsymbol{\Omega}_K\mathbf{h}_K,
\end{aligned}$$

$$\mathbf{y} = \boldsymbol{\beta}_K + \boldsymbol{\Omega}_K\mathbf{a}\left[\boldsymbol{\beta}_{K-1} + \boldsymbol{\Omega}_{K-1}\mathbf{a}\left[\ldots\boldsymbol{\beta}_2 + \boldsymbol{\Omega}_2\mathbf{a}\left[\boldsymbol{\beta}_1 + \boldsymbol{\Omega}_1\mathbf{a}\left[\boldsymbol{\beta}_0 + \boldsymbol{\Omega}_0\mathbf{x}\right]\right]\ldots\right]\right]$$

Ejemplo



5.1 Redes de Neuronas Profundas

- Componer dos redes
- Combinar las dos redes en una
- Hiperparámetros
- Notación para redes profundas
- Redes no profundas vs profundas

Redes no profundas vs profundas

- Los mejores resultados se obtienen con redes profundas con muchas capas:
  - 50 a 1000 capas para la mayoría de aplicaciones.

Los mejores resultados en:
- Visión por Computadora
- Procesamiento del Lenguaje Natural
- Redes de Grafos
- Modelos Generativos
- Aprendizaje por Refuerzo

Todos utilizan redes profundas ¿por qué?

Redes no profundas vs profundas

1) ¿Habilidad de aproximar funciones diferentes?

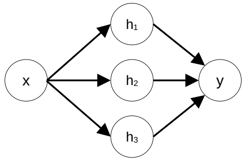
Ambos tipos de redes obedecen el teorema de aproximación universal

Argumento: Una capa es suficiente, y para redes profundas se puede hacer que el resto de capas calculen la función identidad

---

Redes no profundas vs profundas

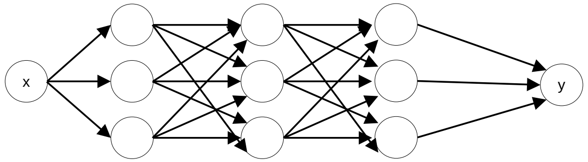
2) Número de regiones lineales por parámetro



Para una red no profunda con Dunidades en la capa oculta, 1 de entrada y 1 de salida:

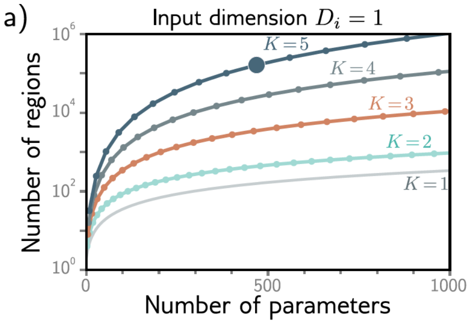
n.º de regiones máximo $R = D+1$ (puntos de articulación + 1)
n.º de parámetros $P = D·1 + D + D·1 + 1 = 3·D+1$

Para una red profunda con, K capas ocultas, D unidades en cada capa oculta, 1 de entrada y 1 de salida:



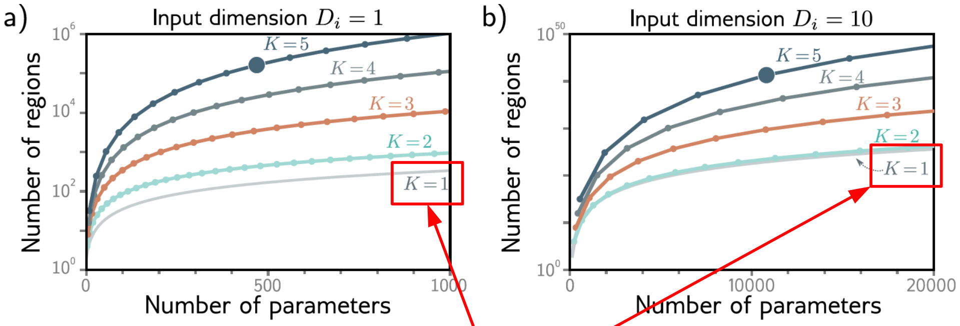
$R = (D+1)^K$
$P = 3·D+1 + (K-1)·(D·D+D)$

Número de regiones lineales por parámetro



5 capas ocultas
10 unidades ocultas por capa
471 parámetros
161501 regiones lineales

Número de regiones lineales por parámetro



5 capas ocultas
10 unidades ocultas por capa
471 parámetros
161501 regiones lineales

Red no profunda
(shallow)

5 capas ocultas
50 unidades ocultas por capa
10801 parámetros
Más de 10^{40} regiones lineales

Redes no profundas vs profundas

2) Número de regiones lineales por parámetro

- Las redes profundas crean muchas más regiones por parámetro
- Pero existen dependencias entre las regiones:
  - Pensad en el ejemplo del plegado (dependencias complejas y simetrías)
  - ¿Quizá existen simetrías similares en funciones del mundo real? (y esta característica de las redes profundas es una ventaja)

Redes no profundas vs profundas

3) Mayor eficiencia con el aumento de las capas (profundidad)

- Hay algunas funciones que requieren una red no profunda con un número exponencialmente mayor de unidades ocultas que una red profunda para lograr una aproximación equivalente.
- Esto se conoce como eficiencia de profundidad de las redes profundas.
- ¿Pero las funciones del mundo real que queremos aproximar tienen esa propiedad? Es algo desconocido

Redes no profundas vs profundas

4) Redes estructuradas muy grandes

- Las imágenes como entrada pueden tener 1M píxel
- No es práctico utilizar redes completamente conectadas
- Mejor tener pesos compartidos localmente, y compartir en toda la imagen
- Esto nos lleva a las redes convolucionales
- Integran gradualmente la información de toda la imagen: necesita múltiples capas

---

Redes no profundas vs profundas

5) Entrenamiento y generalización

- Entrenar con redes profundas parece ser más fácil hasta 20 capas
- Más allá se necesitan varios trucos para entrenar redes más profundas, por lo que (en forma básica), el ajuste se vuelve más difícil.
- La generalización es buena en redes profundas. ¿Por qué?

Redes no profundas vs profundas

5) Entrenamiento y generalización

Figure 20.2 MNIST-1D training. Four fully connected networks were fit to 4000 MNIST-1D examples with random labels using full batch gradient descent, He initialization, no momentum or regularization, and learning rate 0.0025. Models with 1,2,3,4 layers had 298, 100, 75, and 63 hidden units per layer and 15208, 15210, 15235, and 15139 parameters, respectively. All models train successfully, but deeper models require fewer epochs.