

Universidad Nacional de Rosario
Facultad de Ciencias Exactas, Ingeniería y Agrimensura
Escuela de Ciencias Exactas y Naturales
Departamento de Matemática

Tesina de Licenciatura en Matemática

Aproximación de funciones por redes neuronales: un enfoque constructivo

Autor: Rolando Cardozo

Director: Ariel Lombardi

Índice general

1. Introducción **1**
1.1. Aspectos matemáticos de Redes Neuronales profundas 1
1.2. Expresividad de funciones en normas clásicas 3
1.3. El curso (o *la maldición*) de la dimensión 5
1.4. Bosquejo de la técnica utilizada 6

2. Redes Neuronales **9**
2.1. Definiciones básicas 9
2.2. Operaciones con redes neuronales 11
2.2.1. Concatenación de redes 11
2.2.2. Paralelización de redes 14
2.3. Redes neuronales estándar 15

3. Aproximación de funciones en normas de Sobolev **20**
3.1. Cotas superiores 20
3.1.1. Algunos casos particulares: aproximación de x^2 y xy 22
3.1.2. Algunos lemas técnicos 33
3.1.3. Demostración del Teorema principal 53
3.2. Acotación inferior de complejidad 54

A. Definiciones y resultados necesarios **56**
A.1. Espacios de interpolación: 56
A.2. Espacios de Sobolev 57
A.2.1. Espacio de Sobolev Fraccionarios 58
A.2.2. Polinomio de Taylor promediado. 60

jkjklk

Capítulo 1

Introducción

Las redes neuronales profundas se conviertieron en la actualidad en una herramienta fundamental para tratar con problemas que involucran grandes volúmenes de datos, así están reemplazando a técnicas clásicas en diversas aplicaciones, como clasificación de imágenes, reconocimiento de voz y procesamiento de lenguaje natural.

A pesar de su éxito, aún no se ha encontrado una explicación matemática completa de su desempeño. Comprender teóricamente estas técnicas es clave para diseñar arquitecturas más eficientes y garantizar su seguridad en aplicaciones críticas, como pueden ser, la conducción autónoma y el diagnóstico médico. Un aspecto fundamental a explicar es el *error de generalización*, que mide la capacidad de una red, entrenada con cierto conjunto de datos, para predecir correctamente datos que no pertenecen a dicho conjunto. Su estudio se divide en un problema estadístico, dependiente de las muestras, y un problema teórico de aproximación, independiente de ellas.

El conocimiento sobre la capacidad de aproximación de las redes neuronales es especialmente relevante en la resolución de ecuaciones diferenciales parciales (PDEs) mediante aprendizaje profundo, donde se han logrado avances significativos. Para obtener estimaciones de error adecuadas, es crucial analizar la fidelidad de la aproximación en espacios de Lebesgue y normas tipo Sobolev.

Este trabajo estudia la capacidad de las redes neuronales profundas para *expresar*, aproximar, funciones con regularidad Sobolev, incluyendo espacios fraccionarios, analizando detalladamente las características necesarias para asegurar una tolerancia prefijada de error independiente de la función que se desea aproximar.

La Tesis recorre parte del trabajo [6], completando los detalles de algunas demostraciones.

1.1. Aspectos matemáticos de Redes Neuronales profundas

Consideraremos redes con *arquitectura feedforward* que implementan una función como una sucesión de transformaciones afines compuestas con una función no lineal, llamada *función de activación*. La longitud L de la sucesión es la cantidad de *capas* de la red. En el caso de redes neuronales *standard* (que definiremos formalmente más adelante) la *entrada* (o *input*) de una capa es la *salida* (u *output*) de la anterior. Podemos denotar por $N_0 \in \mathbb{N}$ la dimensión del espacio (cantidad de variables de

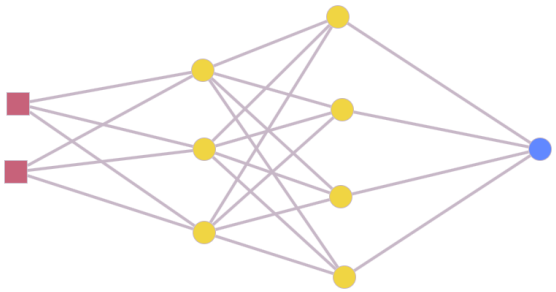


Figura 1.1: Red neuronal (estándar) con tres capas (dos capas ocultas) y 8 neuronas, dimensión de entrada 2, dimensión de salida 1

la función a implementar) y por N_l la dimensión del *output* de la l-esima capa, $l = 1, \ldots, L$. Así, si $\sigma : \mathbb{R} \to \mathbb{R}$ es la función de activación, $T_l : \mathbb{R}^{N_{l-1}} \to \mathbb{R}^{N_l}$ es la transformación afín de la capa l, el cálculo en esa capa puede definirse por la función

$$f_l : R^{N_{l-1}} \to \mathbb{R}^{N_l} \qquad f_l(x) = \sigma(T_l(x))$$

para $l = 1 \ldots, L-1$ y

$$f_L : R^{N_{L-1}} \to \mathbb{R}^{N_l} \qquad f_L(x) = T_L(x).$$

Aclaramos que la función ρ se aplica componente a componente. Finalmente, la red neuronal profunda implementa entonces la función

$$f_L \circ f_{L-1} \circ \cdots \circ f_1 : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}.$$

Notemos que la función de activación actúa en todas las capas salvo la última. Los parámetros que definen las transformaciones afines se llaman *pesos*. La suma $\sum_{k=0}^{L} N_k$ es la cantidad de neuronas dado que cada coordenada de la salida de f_l puede pensarse como una pequeña unidad de cálculo, similar a una neurona en el cerebro. La *complejidad* de una red neuronal se mide por la cantidad de capas, pesos y neuronas.

En la figura 1.1 puede verse una representación gráfica de una red neuronal de $L = 3$ capas (dos capas ocultas). En este caso la dimensión de entrada es $N_0 = 2$, la dimensión de salida $N_3 = 1$ y tiene $N_1 = 3$ neuronas en la capa 1 y $N_2 = 4$ neuronas en la capa 2.

Consideraremos también arquitecturas más generales que permiten conexiones entre capas no vecinas, llamadas, *skip connections*. En estas, la entrada de una capa consiste de la salida de todas las capas precedentes. Presentamos un ejemplo en Figura 1.2.

Una de las funciones de activación más usadas es la *Rectified Linear Unit*, ReLU, definida por

$$\rho : \mathbb{R} \to \mathbb{R}, \qquad \rho(x) = \max\{0, x\}.$$

Esta Tesis, siguiendo el trabajo [6], se enfoca en redes ReLu con arquitectura *feed-forward* y se permiten *skip connections*. Si bien la definición de redes neuronales es independiente de la función de activación, el uso de ReLU es fundamental para la obtención de los resultados de esta Tesis. Por ejemplo, la función de activación ReLu permite realizar la función Identidad (ver Lema 2.2.2) y con esto definir la

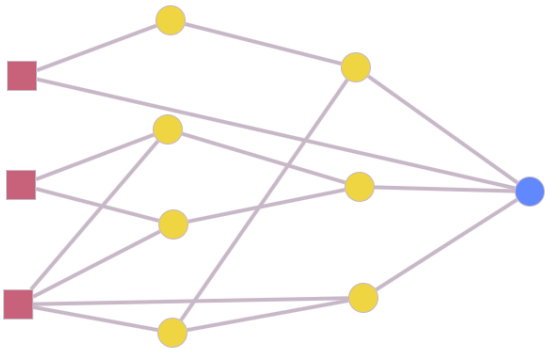


Figura 1.2: Una red neuronal con *skip connections*. En azul, conexiones entre capas no vecinas

concatenación dispersa (ver Definición 2.2.2) y demostrar que toda función realizada por una red neuronal puede realizarse por una red neuronal estándar de complejidad similar (ver Lemma 2.3.1). Para la validez de la Proposición 3.1.2 también resulta fundamental que $\sigma = \rho$. Esta Proposición afirma que la función x^2 puede ser aproximada tanto como se quiera en norma $W^{1,\infty}(0,1)$ por redes neuronales, y es el punto de partida para la obtención de los resultados de esta Tesis.

1.2. Expresividad de funciones en normas clásicas

Uno de los primeros resultados que aseguran que un espacio de funciones puede aproximarse por redes neuronales es el Teorema de aproximación universal que puede encontrarse por ejemplo en [3]. Sea $\sigma : \mathbb{R} \to \mathbb{R}$ una función continua. Decimos que σ es *sigmoide* si

$$\sigma(t) \to \begin{cases} 1 & \text{si } t \to +\infty \\ 0 & \text{si } t \to -\infty \end{cases}$$

. Entonces podemos enunciar un teorema de aproximación universal (ver [3, Teorema 2]).

Teorema 1.2.1. *Sea σ cualquier función sigmoide continua. Dada una función $f \in \mathcal{C}([0,1]^d)$ y $\varepsilon > 0$ existe una red neuronal Φ con función de activación σ, con dos capas, tal que*

$$|f(x) - R_\sigma(\Phi)(x)| < \varepsilon, \qquad \forall x \in [0,1]^d.$$

En el mismo paper se demuestra un resultado similar para la aproximación de funciones en $L^1([0,1]^d)$, con redes neuronales de dos capas y con función de activación medible, acotada y sigmoide, extendiendo la definición de sigmoide a funciones medibles (ver [3, Teorema 4]).

Estos resultados se generalizan en el paper [8]. Por un lado se consideran funciones de activación no necesariamente de tipo sigmoide. Y por otro lado se considera la aproximación de funciones en espacios L^p, $1 \leq p < \infty$ Más precisamente se sigue de allí el siguiente resultado.

Teorema 1.2.2. *Sea σ una función de acotada y no constante. Dada $f \in L^p([0,1]^d)$ y $\varepsilon > 0$ existe una red neuronal Φ con función de activación σ y dos capas tal que*

$$\|f - R_\sigma(\Phi)\|_{L^p([0,1]^d)} < \varepsilon.$$

En el mismo paper se extiende el resultado anterior a la aproximación de funciones continuas en norma $\|\cdot\|_{L^\infty([0,1]^d)}$, agregando la hipótesis de continuidad sobre la función de activación (ver [8, Teorema 2]). También se prueba un teorema similar de aproximación de funciones $\mathcal{C}^m([0,1]^d)$ en normas de Sobolev $W^{m,p}([0,1]^d)$ con funciones de activación $\mathcal{C}^m(\mathbb{R}^d)$ acotadas y no constantes (ver [8, Teorema 3]).

En [9] se presenta una generalización notable de Teoremas 1.2.1 y 1.2.2, que impone condiciones mínimas sobre la función de activación y contiene, como casos especiales, prácticamente todos los casos de funciones de activación reportados en la literatura.

Teorema 1.2.3. *[9, Teorema 1] Para todo compacto $K \subset \mathbb{R}^d$, $f \in \mathcal{C}(\mathbb{R}^d)$ y $\epsilon > 0$, existe una red neuronal Φ con dos capas tal que*

$$\|f - R_\sigma(\Phi)\|_{L^\infty(K)} < \varepsilon$$

si y solo si la función de activación σ no es un polinomio.

Notar que en los resultados mencionados, las redes neuronales tienen dos capas, pero no hay una estimación, o no se tiene control, de la cantidad de neuronas y pesos de las redes. Uno de los primeros análisis de la complejidad necesaria para obtener un grado de aproximación se presenta en [13]. Se $\mathcal{W}^{n,\infty}([0,1]^d)$ el espacio de funciones que junto con sus derivadas hasta orden n están en $L^\infty([0,1]^d)$, con la norma

$$\|f\|_{\mathcal{W}^{n,\infty}([0,1]^d)} = \max_{\alpha:|\alpha|\leq n} \operatorname*{ess\,sup}_{x\in[0,1]^d} |D^\alpha f(x)|.$$

Allí se analiza la capacidad de redes neuronales para aproximar en norma $\|\cdot\|_{L^\infty([0,1]^d)}$ el conjunto de funciones

$$F_{n,d} = \{f \in \mathcal{W}^{n,\infty}([0,1]^d) : \|f\|_{\mathcal{W}^{n,\infty}([0,1]^d)} < 1\}$$

Teorema 1.2.4. *ver [13, Teorema 1] Sean $d, n \in \mathbb{N}$ y $\varepsilon > 0$. Sea $f \in F_{d,n}$. Existe una red neuronal Φ con función de activación ReLU tal que*

$$\|f - R_\sigma(\Phi)\|_{L^\infty([0,1]^d)} < \varepsilon$$

con cantidad de capas de a lo sumo $c(\log(1/\varepsilon)+1)$ y a lo sumo $c\varepsilon^{-d/n}(\log(1/\varepsilon)+1)$ neuronas y pesos, siendo $c = c(n,d)$.

En realidad [13, Teorema 1] contiene un resultado más fuerte, afirmando que dada la tolerancia ε existe una arquitectura de red neuronal con la complejidad enunciada tal que toda función en $F_{d,n}$ puede aproximarse por una red neuronal con dicha arquitectura con error menor que ε (ver Definición 2.1.3 de Arquitectura).

En esta Tesis, siguiendo el trabajo [6], nos interesa la aproximación simultánea de la función y sus derivadas de orden ≤ 1 (que pueden ser de orden fraccionario). Así se extiende el resultado del Teorema 1.2.4 al caso en que la aproximación se mide en la norma $W^{s,p}((0,1)^d)$ con $s \in (0,1)$ y $1 \leq p \leq \infty$. Especíificamente:

- *Cotas superiores de complejidad.* Para funciones $W^{n,p}((0,1)^d)$ con norma ≤ 1, en Teorema 3.1.1 se demuestra que se puede obtener una aproximación con tolerancia ε en norma $W^{s,p}((0,1)^d)$ por redes neuronales ReLU con a lo sumo $c\log(\varepsilon^{-n/(n-s)})$ capas y a lo sumo $c\varepsilon^{-d/(n-s)}\log(\varepsilon^{-n/(n-s)})$ neuronas y pesos.
- *Trade-off.* Como en aproximaciones con otros métodos clásicos, vemos que existe un compromiso entre la regularidad n de las funciones a aproximar y la regularidad s de la norma de aproximación. Para asegurar una tolerancia de ε en norma s para una función con regularidad n, la complejidad de la red involucra potencias de $1/\varepsilon$ con un exponente $1/(n-s)$. Así, cuanto más fina es la norma de aproximación, es mayor la complejidad requerida de la red, mientras que si aumenta regularidad de la función a aproximar disminuye dicho costo.

Enunciamos además un resultado (ver Teorema 3.2.1) que da cotas inferiores de complejidad. Establece que si una arquitectura de red neuronal es capaz de aproximar en la norma $W^{1,\infty}$ con tolerancia ε cualquier función de $W^{n,\infty}((0,1)^n)$ y norma ≤ 1, entonces la cantidad de pesos de la arquitectura debe ser de al menos $c\varepsilon^{-d/(2(n-1))}$.

1.3. El curso (o *la maldición*) de la dimensión

Como vemos, las cotas superiores e inferiores de complejidad para asegurar una tolerancia ε en la aproximación en norma $W^{1,\infty}$ de funciones $W^{n,\infty}((0,1)^d)$ dependen exponencialmente de la dimensión. Esta característica suele expresarse como *curso de la dimensión*. La cota inferior, en particular, demuestra que para aproximaciones uniformes en norma $W^{1,\infty}$ no puede evitarse el curso de la dimensión.

Algo similar ocurre con otros métodos clásicos. Por ejemplo, consideremos la aproximación de un problema elíptico en $(0,1)^d$ por el método de elementos finitos. Supongamos que elementos d-lineales en una malla uniforme de d-rectángulos de lado h. En este caso el método de elementos finitos genera una solución numérica u_h que es *casi* la mejor aproximación a la solución exacta en el espacio discreto V_h, que consiste de todas las funciones continuas y d-lineales en cada rectángulo de la malla. Si la solución exacta está en el espacio $W^{n,p}((0,1)^d)$, resulta que el error de aproximación puede estimarse por

$$\|u - u_h\|_{W^{1,p}((0,1)^d)} \leq Ch^{n-1}\|u\|_{W^{n,p}(0,1)^d}.$$

con una constante C independiente de la complejidad. La complejidad del método se mide por la cantidad N de grados de libertad del espacio discreto, esto es, la dimensión de dicho espacio, que en este caso simple puede verse que es

$$N = \left(\frac{1}{h}\right)^d.$$

Llamando $\varepsilon = Ch^{n-1}$, si lo escribimos en términos de N tenemos $\varepsilon = CN^{-(n-1)/d}$ y por lo tanto $N = C\varepsilon^{-d/(n-1)}$ (posiblemente con una diferente constante C). Resumiendo, para asegurar una cota de error

$$\|u - u_h\|_{W^{1,p}((0,1)^d)} \leq C\varepsilon,$$

(suponiendo $\|u\|_{W^{n,p}(0,1)^d} \leq 1$), requerimos el uso de $N = C\varepsilon^{-d/(n-1)}$ grados de libertad, o complejidad. Así observamos que el método de elementos finitos también sufre el curso de la dimensión.

Cuando la función de activación es la ReLU, puede verse que, en general, las funciones del espacio V_h son funciones generadas por las redes neuronales (ver [7, Teorema 3.1] para un resultado relacionado), sin embargo, el espacio de funciones generadas por estas es infinitamente más grande. Es por eso que se esperaría que, en teoría, la aproximación por redes neuronales sea superior. Sin embargo ambos métodos sufren el curso de la dimensión de manera análoga.

Existen estrategias para evitar en teoría, o disminuir el impacto, del curso de la dimensión en el caso de aproximaciones por redes neuronales. Un inconvenientes con los resultados mencionados anteriormente es que buscan aproximar uniformemente todas las funciones en una bola de $W^{n.p}((0,1)^n)$. Pero muchas veces la función de interés está en un espacio más reducido. Una de las estrategias consiste describir las funciones a aproximar mediante *modelos de esparcidad*, que de manera muy somera, dado un espacio vectorial con una *base* infinita, la clase de funciones a aproximar está dada por las funciones que pueden expresarse como combinación lineal de a lo sumo $n < \infty$ elementos de la base. Se trata de un espacio no lineal. La aproximación por redes neuronales de funciones de esta clase puede lograrse con complejidad independiente de la dimensión. De cualquier manera, los modelos de esparcidad no son inmunes al curso de la dimensión, puesto que la base del espacio vectorial subyacente es infinita. Referimos a [4, Sección 8] para una explicación mucho más completa de esta estrategia, sus ventajas y desventajas.

1.4. Bosquejo de la técnica utilizada

En el Capítulo 3 se presenta el principal resultado de la Tesis, Teorema 3.1.1, que asegura que, dado $\varepsilon > 0$ existe una arquitectura de red neuronal $\mathcal{A}_\varepsilon$ con una complejidad acotada en términos de ε tal que toda función $W^{n,p}((0,1)^d)$ (con norma ≤ 1) puede aproximarse por una red neuronal Φ de arquitectura $\mathcal{A}_\varepsilon$ con error $W^{s,p}((0,1)^d)$ acotado por ε.

La idea de la demostración consiste en introducir una partición de la unidad ϕ_m, $m \in 0, 1, \ldots, M^d$ de $(0,1)^d$ de funciones con soportes de medida acotada por M^{-d}, de tal manera que cada ϕ_m es realizada como el producto de la salida de una red neuronal Φ_m:

$$\phi_m(x) = \prod_{l=1}^{d} \left[R_\rho(\Phi_m)\right]_l$$

(ver Lema 3.1.1). Dada una $f \in W^{n,p}((0,1)^d)$ se construyen, utilizando polinomios de Taylor promediados, polinomios p_m de grado $< n$ de manera que

$$\left\| f - \sum_m \phi_n p_m \right\|_{W^{s,p}(0,1)^d} = \left\| \sum_m (f - p_m)\, \phi_m \right\|_{W^{s,p}(0,1)^d} < \frac{\varepsilon}{2}.$$

El paso importante es ahora encontrar una red neuronal $\Phi_{P,\varepsilon}$, con arquitectura prescrita, tal que

$$\left\| \sum_m \phi_n p_m - R_\rho(\Phi_{P,\varepsilon}) \right\|_{W^{s,p}(0,1)^d} < \frac{\varepsilon}{2}.$$

Esto se logra teniendo en cuenta que $\phi_m p_m$ es una suma de *monomios localizados* $\phi_m x^\alpha$ con $|\alpha| < n$. Cada monomio x^α puede realizarse como

$$x^\alpha = \prod_{l=1}^{|\alpha|} \left[R_\rho(\Phi_\alpha)\right]_l (x)$$

Luego aplicando la operación de paralelización $P(\Phi_m, \Phi_\alpha)$ a las redes Φ_m y Φ_α se obtiene una red $\Phi_{m,\alpha}$ tal que

$$\phi_m x^\alpha = \prod_{l=1}^{|\alpha|+d} \left[R_\rho(\Phi_{m,\alpha})\right]_l (x)$$

Si bien no podemos encontrar una red que realice este producto, sí se puede encontrar una que esté cerca, es decir, existe una red, denotada por $\Phi_{\varepsilon,(m,\alpha)}$ tal que

$$\left\|\phi_m x^\alpha - R_\rho(\Phi_{\varepsilon,(m,\alpha)})\right\|_{W^{k,p}((0,1)^d)} \leq c(N,n,d)\varepsilon.$$

Teniendo en cuenta que

$$\sum \phi_m p_m = \sum_{\alpha,m} c_{m,\alpha}\phi_m x^\alpha$$

y que cada término $\phi_m x^\alpha$ es *casi* realizado por $\Phi_{\varepsilon,(m,\alpha)}$, el problema se reduce a hallar una red neuronal $\Phi_{P,\varepsilon}$ que realice la combinación lineal

$$R_\rho(\Phi_{P,\varepsilon}) = \sum_{m,\alpha} c_{m,\alpha} R_\rho\left(\Phi_{\varepsilon,(m,\alpha)}\right).$$

Esto último se logra componiendo dos redes neuronales, una que realiza una combinación lineal de las componentes de la entrada, y otra que es la operación de paralelización P aplicada a las redes $\Phi_{\varepsilon,(m,\alpha)}$ con $m \in \{1,\ldots,N\}^d$ y $|\alpha| < n$.

Finalmente basta ver que la red $\Phi_{P,\varepsilon}$ es la buscada, y para esto tenemos que obtener la acotación

$$\left\|\sum \phi_m p_m - R_\rho(\Phi_{P,\varepsilon})\right\|_{W^{s,p}((0,1)^d)} < \frac{\varepsilon}{2},$$

que se logra eligiendo adecuadamente el parámetro N.

Vemos en esta construcción que dada una red neuronal Φ con salida m dimensional, resulta fundamental hallar otra red $\Psi_{\varepsilon,\Phi}$ que *casi realice* el producto de las componentes de la salida de Φ, esto es

$$\left\|R_\rho\left(\Psi_{\varepsilon,\Phi}\right) - \prod_{l=1}^{m}\left[R_\rho\left(\Phi\right)\right]_l\right\|_{W^{k,p}((0,1)^d)} < \varepsilon$$

Esto se logra usando un argumento inductivo esencialmente en la dimensión m (ver Lema 3.1.3). En el caso particular $m = 2$, la red $\Psi_{\varepsilon,\Phi}$ resulta de la composición de Φ con $\widetilde{\times}$, siendo esta última una red neuronal, con entrada bidimensional y salida unidimensional, que aproxima a la función xy en un cuadrado $[-M,M]^2$ con complejidad dependiendo logarítmicamente del grado ε de aproximación (ver Proposición 3.1.4).

Notablemente, este resultado de aproximación, es consecuencia de

- Que la función x^2 puede aproximarse con error menor que ε en norma $W^{1,\infty}(0,1)$ por la salida de una red Φ_ε^{sq} con complejidad dependiendo logarítmicamente de ε (Proposición 3.1.2)

- La *identidad de polarización* que permite escribir a xy como combinación lineal de cuadrados
$$xy = \frac{1}{2}\left((x+y)^2 - x^2 - y^2\right).$$

Después de esta Introducción, en el Capítulo 2 introducimos la definición formal de red neuronal y las operaciones fundamentales de paralelización y concatenación. En el Capítulo 3 se presentan los resultados principales sobre las cotas superiores e inferiores de complejidad para la aproximación por redes neuronales de funciones en espacios de Sobolev. Se presenta una demostración constructiva, completa, para la existencia de una arquitectura de red neuronal que aproxima una función arbitraria en la bola unitaria de un espacio de Sobolev con tolerancia prescrita, dando una cota superior de la complejidad de tal arquitectura. Y solo se enuncia un resultado sobre una cota inferior de la complejidad de una arquitectura con las carácterísticas mencionadas. En el Apéndice A presentamos resultados útiles sobre Interpolación de Operadores en espacios de Banach, Espacios de Sobolev y Aproximación Polinomial.

Capítulo 2

Redes Neuronales

2.1. Definiciones básicas

En esta sección hacemos referencia a las definiciones y los resultados sobre redes neuronales del [6, Capitulo 2].

Definición 2.1.1. Sean $d, L \in \mathbb{N}$

1. Una red neuronal Φ con dimensión de entrada d y L capas es una secuencia de t-uplas matriz-vector:

$$\Phi = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$$

donde $N_0 = d$ y $N_1, \ldots, N_L \in \mathbb{N}$, y donde cada A_l es una matriz de dimensión $N_l \times \sum_{k=0}^{l-1} N_k$ y $b_l \in \mathbb{R}^{N_l}$.

2. Sea Φ red neuronal y $\sigma : \mathbb{R} \to \mathbb{R}$. Definimos entonces la realización asociada a Φ con función de activación σ como la aplicación $R_\sigma(\Phi) : \mathbb{R}^d \to \mathbb{R}^{N_L}$ tal que

$$R_\sigma(\Phi)(x) = x_L$$

donde x_L resulta del siguiente esquema:

$$\begin{aligned} x_0 &:= x, \\ x_l &:= \sigma\left(A_l[x_0^T|\ldots|x_{l-1}^T]^T + b_l\right) \quad \text{para } l = 1, \ldots, L-1, \\ x_L &:= A_L[x_0^T|\ldots|x_{L-1}^T]^T + b_L. \end{aligned}$$

Aquí, σ actúa componente a componente, es decir, $\sigma(y) = [\sigma(y^1), \ldots, \rho(y^m)]$ para $y = [y^1, \ldots, y^m] \in \mathbb{R}^m$. A veces escribimos A_l en forma de matriz por bloques como

$$A_l = [A_{l,x_0}|\ldots|A_{l,x_{l-1}}],$$

donde A_{l,x_k} es una matriz de dimensión $N_l \times N_k$ para $k = 0, \ldots, l-1$ y $l = 1, \ldots, L$. Entonces,

$$\begin{aligned} x_l &= \sigma\left(A_{l,x_0}x_0 + \cdots + A_{l,x_{l-1}}x_{l-1} + b_l\right) \quad \text{para } l = 1, \ldots, L-1, \\ x_L &= A_{L,x_0}x_0 + \cdots + A_{L,x_{L-1}}x_{L-1} + b_L. \end{aligned}$$

3. Llamamos
$$N(\Phi) := d + \sum_{j=1}^{L} N_j$$
el número de neuronas de la red Φ,
$$L = L(\Phi)$$
el número de capas, y
$$M(\Phi) := \sum_{j=1}^{L} (\|A_j\|_{\ell^0} + \|b_j\|_{\ell^0})$$
denota el número de entradas no nulas de todas las matrices A_l y vectores b_l, el cual llamamos el número de pesos de Φ. Además, nos referimos a N_L como la dimensión de la capa de salida de Φ.

Ahora vamos a definir redes neuronales donde solo son permitidas conexiones entre capas vecinas. Estas redes son un caso particular de la definición 2.1.1.

Definición 2.1.2. Si $\Phi = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$ es una red neuronal como la definida en Definición 2.1.1 y
$$A_l = [A_{l,x_0} | \ldots | A_{l,x_{l-1}}],$$
satisface $A_{l,x_i} = 0$ para $i = 1, \ldots, l-2$ y $l = 0, \ldots, L$, entonces la llamamos red neuronal estándar. En este caso, se reduce a resolver el esquema
$$\begin{aligned} x_0 &:= x, \\ x_l &:= \sigma\left(A_{l,x_{l-1}} x_{l-1} + b_l\right) \quad \text{para } l = 1, \ldots, L-1, \\ x_L &:= A_{L,x_{L-1}} x_{L-1} + b_L. \end{aligned}$$

En práctica, antes de entrenar una red neuronal se debe elegir una arquitectura de red neuronal para usar. Esto quedará claro, con la siguiente definición que da la noción de arquitectura de red neuronal.

Definición 2.1.3. Sean $d, L \in \mathbb{N}$

1. Una arquitectura de red neuronal $\mathcal{A}$ con dimensión de entrada d y L capas es una secuencia de tuplas matriz-vector:
$$\mathcal{A} = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$$
donde $N_0 = d$ y $N_1, \ldots, N_L \in \mathbb{N}$, y donde cada A_l es una matriz de dimensión $N_l \times \prod_{k=0}^{l-1} N_k$ con elementos en $\{0, 1\}$ y $b_l \in \mathbb{R}^{N_l}$ con elementos en $\{0, 1\}$. Así, una arquitectura de red neuronal tiene pesos binarios.

2. Llamamos $N(\mathcal{A}) := d + \sum_{j=1}^{L} N_j$ el número de neuronas de la arquitectura $\mathcal{A}$, $L = L(\mathcal{A})$ el número de capas, y

$$M(\mathcal{A}) := \sum_{j=1}^{L} (\|A_j\|_{\ell^0} + \|b_j\|_{\ell^0})$$

denota el número de entradas no nulas de todas las matrices A_l y vectores b_l, lo cual llamamos el número de pesos de $\mathcal{A}$. Más aún, nos referimos a N_L como la dimensión de la capa de salida de $\mathcal{A}$.

3. Decimos que una red neuronal $\Phi = ((A'_1, b'_1), (A'_2, b'_2), \ldots, (A'_L, b'_L))$ tiene arquitectura $\mathcal{A}$ si tiene dimensión de entrada d, L capas y

   a) $N_l(\Phi) = N_l$ para todo $l = 1, \ldots, L$ y

   b) $[A_l]_{i,j} = 0$ implica que $[A'_l]_{i,j} = 0$ y $[b_l]_i = 0$ implica $[b'_l]_i = 0$ para todo $l = 1, \ldots, N$, con $i = 1, \ldots, N_l$ y $j = 1, \ldots, \sum_{k=1}^{l-1} N_k$.

4. Decimos que $\mathcal{A}$ es una arquitectura de red neuronal estándar si todos los pesos conectando neuronas de capas no adyacentes son cero.

5. Una vez fijado un orden de las entradas no nulas en A_l para $l = 1, \ldots, L$, podemos identificar de manera unívoca cada vector w de $\mathbb{R}^{M(\mathcal{A})}$ con una red neuronal (estándar) que tiene arquitectura (estándar) $\mathcal{A}$. Denotamos esta red neuronal (estándar) por $\mathcal{A}(w)$.

Observamos que en general, una arquitectura de una red neuronal no necesita ser única, es decir, una red neuronal Φ puede tener arquitecturas $\mathcal{A}_1$ y $\mathcal{A}_2$ con $\mathcal{A}_1 \neq \mathcal{A}_2$.

2.2. Operaciones con redes neuronales

En esta sección nos centraremos en ver las operaciones Concatenación y Paralelización entre redes neuronales. Recordemos que ρ es la función ReLU definida en la Introducción.

2.2.1. Concatenación de redes

La concatenación de dos redes nos permite componer las realizaciones de las mismas. En detalle, si Φ^1 y Φ^2 son dos redes, estamos interesados en la red $\Phi^1 \bullet \Phi^2$ tal que $R_\sigma(\Phi^1 \bullet \Phi^2)(x) = R_\sigma(\Phi^1) \circ R_\sigma(\Phi^2)(x)\ \forall x \in \mathbb{R}^d$. Construimos una red $\Phi^1 \bullet \Phi^2$ para una función de activación general σ. En el caso particular donde la función de activación es ρ que es la función ReLU, usamos esto para encontrar una red que realice la función identidad y con esta construir una llamada Concatenación dispersa que nos da control sobre la complejidad de la red resultante.

Definición 2.2.1. Sean $L_1, L_2 \in \mathbb{N}$ y sea

$$\Phi^1 = ((A^1_1, b^1_1), (A^1_2, b^1_2), \ldots, (A^1_{L_1}, b^1_{L_1})), \Phi^2 = ((A^2_1, b^2_1), (A^2_2, b^2_2), \ldots, (A^2_{L_2}, b^2_{L_2}))$$

dos redes, tal que la dimensión de entrada de la red Φ^1 coincide con la dimensión de salida de Φ^2. Entonces, $\Phi^1 \bullet \Phi^2$ denota la siguiente red neuronal de $L_1 + L_2 - 1$ capas:

$$\Phi^1 \bullet \Phi^2 := ((A_1^2, b_1^2), \ldots, (A_{L_2-1}^2, b_{L_2-1}^2), (\tilde{A}_1^1, \tilde{b}_1^1), \ldots, (\tilde{A}_{L_1}^1, \tilde{b}_{L_1}^1))$$

donde

$$\tilde{A}_l^1 := [A_{l,x_0}^1 A_{L_2}^2 | \cdots | A_{l,x_{l-1}}^1], \quad \text{y} \quad \tilde{b}_l^1 := A_{l,x_0}^1 b_{L_2}^2 + b_l^1$$

para $l = 1, \ldots, L_1$. Llamamos $\Phi^1 \bullet \Phi^2$ la concatenación de Φ^1 y Φ^2.

Lema 2.2.1. *Sean Φ^1 y Φ^2 dos redes neuronales tales que la dimensión de entrada de la red Φ^1 coincide con la dimensión de salida de Φ^2. Entonces*

$$R_\sigma(\Phi^1 \bullet \Phi^2)(x) = R_\sigma(\Phi^1) \circ R_\sigma(\Phi^2)(x), \forall x \in \mathbb{R}^d.$$

La demostración del Lema anterior resulta de un simple cálculo. A modo de ejemplo lo verificamos en el caso en que Φ_1 y Φ_2 tienen tres capas:

$$\Phi^1 = ((A_1^1, b_1^1), (A_2^1, b_2^1), (A_3^1, b_3^1)), \Phi^2 = ((A_1^2, b_1^2), (A_2^2, b_2^2), (A_3^2, b_3^2)).$$

Entonces, $\Phi^1 \bullet \Phi^2$ denota la siguiente red neuronal de $3 + 3 - 1 = 5$ capas:

$$\Phi^1 \bullet \Phi^2 := ((A_1^2, b_1^2), (A_2^2, b_2^2), (\tilde{A}_1^1, \tilde{b}_1^1), (\tilde{A}_2^1, \tilde{b}_2^1), (\tilde{A}_3^1, \tilde{b}_3^1)),$$

donde

$$\tilde{A}_1^1 = A_{1,x_0}^1 A_3^2, \qquad \tilde{A}_2^1 = [A_{2,x_0}^1 A_3^2 | A_{2,x_1}^1], \qquad \tilde{A}_3^1 = [A_{3,x_0}^1 A_3^2 | A_{3,x_1}^1 | A_{3,x_2}^1],$$

y

$$\tilde{b}_1^1 = A_{1,x_0}^1 b_3^2 + b_1^1, \qquad \tilde{b}_2^1 = A_{2,x_0}^1 b_3^2 + b_2^1, \qquad \tilde{b}_3^1 = A_{3,x_0}^1 b_3^2 + b_3^1.$$

Luego, sea $x \in \mathbb{R}^d$. Entonces

$$\begin{aligned}
x_0 &= x \\
x_1 &= \sigma\left(A_1^2 x_0 + b_1^2\right), \\
x_2 &= \sigma\left(A_{2,x_0}^2 x_0 + A_{2,x_1}^2 x_1 + b_2^2\right), \\
x_3 &= A_{3,x_0}^2 x_0 + A_{3,x_1}^2 x_1 + A_{3,x_2}^2 x_2 + b_3^2.
\end{aligned}$$

y

$$\begin{aligned}
y_0 &= x_3 \\
y_1 &= \sigma\left(A_{1,x_0}^1 y_0 + b_1^1\right), \\
y_2 &= \sigma\left(A_{2,x_0}^1 y_0 + A_{2,x_1}^1 y_1 + b_2^1\right), \\
y_3 &= A_{3,x_0}^1 y_0 + A_{3,x_1}^1 y_1 + A_{3,x_2}^1 y_2 + b_3^1.
\end{aligned}$$

Así resulta, $y_3 = R_\sigma(\Phi^1)(x_3) = R_\sigma(\Phi^1)(R_\sigma(\Phi^2))(x)$ usando que $x_3 = R_\sigma(\Phi^2)(x)$ Además,

$$\begin{aligned}
z_0 &= x \\
z_1 &= \sigma\left(A_{1,x_0}^2 z_0 + b_1^2\right) = x_1, \\
z_2 &= \sigma\left(A_{2,x_0}^2 z_0 + A_{2,x_1}^2 z_1 + b_2^2\right) = x_2, \\
z_3 &= \sigma\left(\tilde{A}_1^1 \begin{bmatrix} z_0 \\ z_1 \\ z_2 \end{bmatrix} + \tilde{b}_1^1\right) = \sigma\left(A_{1,x_0}^1 (A_{3,x_0}^2 x_0 + A_{3,x_1}^2 x_1 + A_{3,x_2}^2 x_2) + A_{1,x_0}^1 b_3^2 + b_1^1\right) \\
&= \sigma\left(A_{1,x_0}^1 \left(A_{3,x_0}^2 x_0 + A_{3,x_1}^2 x_1 + A_{3,x_2}^2 x_2 + b_3^2\right) + b_1^1\right) = \sigma\left(A_{1,x_0}^1 y_0 + b_1^1\right) = y_1.
\end{aligned}$$

Similarmente obtenemos

$$\begin{aligned}
z_4 &= \sigma\left(A^1_{2,x_0}(A^2_{3,x_0}z_0 + A^2_{3,x_1}z_1 + A^2_{3,x_2}z_2) + A^1_{2,x_1}z_3 + A^1_{2,x_0}b^2_3 + b^1_2\right)\\
&= \sigma\left(A^1_{2,x_0}\left(A^2_{3,x_0}x_0 + A^2_{3,x_1}x_1 + A^2_{3,x_2}x_2 + b^2_3\right) + A^1_{2,x_1}y_1 + b^1_2\right)\\
&= \sigma\left(A^1_{2,x_0}y_0 + A^1_{2,x_1}y_1 + b^1_2\right) = y_2,\\
z_5 &= A^1_{3,x_0}(A^2_{3,x_0}z_0 + A^2_{3,x_1}z_1 + A^2_{3,x_2}z_2) + A^1_{3,x_1}z_3 + A^1_{3,x_2}z_4 + A^1_{3,x_0}b^2_3 + b^1_3\\
&= A^1_{3,x_0}(A^2_{3,x_0}x_0 + A^2_{3,x_1}x_1 + A^2_{3,x_2}x_2 + b^2_3) + A^1_{3,x_1}y_1 + A^1_{3,x_2}y_2 + b^1_3\\
&= A^1_{3,x_0}y_0 + A^1_{3,x_1}y_1 + A^1_{3,x_2}y_2 + b^1_3 = y_3.
\end{aligned}$$

Por lo tanto, $z_5 = y_3 = R_\sigma(\Phi^1)(R_\sigma(\Phi^2))(x)$.

Lema 2.2.2. *Sea $d \in \mathbb{N}$. Definimos*

$$\Phi^{Id} := ((A_1, b_1), (A_2, b_2))$$

con función de activación ρ con

$$A_1 := \begin{bmatrix} Id_{\mathbb{R}^d} \\ -Id_{\mathbb{R}^d} \end{bmatrix}, b_1 := 0, A_2 := [0_{\mathbb{R}^{d,d}} | Id_{R^d}| - Id_{R^d}], b_2 := 0.$$

Entonces $R_\rho(\Phi^{Id}) = Id_{\mathbb{R}^d}$.

Demostración. Sea $x \in \mathbb{R}^d$. Luego,

$$\begin{aligned}
x_1 &= \rho(A_1[x^T]^T) + b_1) = \rho\left(\begin{bmatrix} Id_{\mathbb{R}^d} \\ -Id_{\mathbb{R}^d} \end{bmatrix} x\right) + 0 = \begin{bmatrix} \rho(x) \\ \rho(-x) \end{bmatrix}\\
x_2 &= A_2[x^T, x_1^T]^T + b_2 = [0_{\mathbb{R}^{d,d}} | Id_{\mathbb{R}^d}| - Id_{\mathbb{R}^d}] \begin{bmatrix} x^T \\ \rho(x)^T \\ \rho(-x)^T \end{bmatrix} = \rho(x) - \rho(-x)\\
&= x.
\end{aligned}$$

.

Pues, $\text{máx}(x^i, 0) - \text{máx}(-x^i, 0) = \text{máx}(x^i, 0) + \text{mín}(x^i, 0) = x^i + 0 = x^i$, $\forall i = 1, \ldots, d$.

Por lo tanto, $R_\rho(\Phi^{Id}) = Id_{\mathbb{R}^d}$. □

Usando la Φ_{Id} del lema anterior, podemos definir una concatenación alternativa que nos permite controlar la complejidad de la red resultante para el caso de la función ReLU.

Definición 2.2.2. Sean $L_1, L_2 \in \mathbb{N}$ y

$$\Phi^1 = ((A^1_1, b^1_1), (A^1_2, b^1_2), \ldots, (A^1_{L_1}, b^1_{L_1})), \Phi^2 = ((A^2_1, b^2_1), (A^2_2, b^2_2), \ldots, (A^2_{L_2}, b^2_{L_2}))$$

dos redes con función de activación ρ tal que la dimensión de entrada de la red Φ^1 coincide con la dimensión de salida de Φ^2. Sean Φ^{Id} la del lema 2.2.2. Entonces, definimos la concatenación dispersa de Φ^1 y Φ^2 como

$$\Phi^1 \odot \Phi^2 = \Phi^1 \bullet \Phi^{Id} \bullet \Phi^2.$$

Observación 2.2.1. Tenemos que

$$\Phi^1 \odot \Phi^2 = \left((A_1^2, b_1^2), \dots, (A_{L_2-1}^2, b_{L_2-1}^2), \left(\begin{bmatrix} A_{L_2}^2 \\ -A_{L_2}^2 \end{bmatrix}, \begin{bmatrix} b_{L_2}^2 \\ -b_{L_2}^2 \end{bmatrix} \right), (\tilde{A}_1^1, b_1^1), \dots (\tilde{A}_{L_1}^1, b_{L_1}^1) \right)$$

donde

$$\tilde{A}_l^1 := [0_{\mathbb{R}^{N_l, \sum_{k=1}^{L_2-1} N_k^2}} | A_{l,x_0}^1 | - A_{l,x_0}^1 | A_{l,x_1}^1 | \dots | A_{l,x_{l-1}}^1]$$

para $l = 1, \dots L_1$. Más aún, se tiene que $L(\Phi^1 \odot \Phi^2) = L(\Phi^1) + L(\Phi^2)$ y $M(\Phi^1 \odot \Phi^2) \leq 2M(\Phi^1) + 2M(\Phi^2)$ y $N(\Phi^1 \odot \Phi^2) \leq 2N(\Phi^1) + 2N(\Phi^2)$.

2.2.2. Paralelización de redes

En algunos casos es útil combinar redes en una red más larga. En particular, si Φ^1 y Φ^2 son dos redes, entonces buscamos una red neuronal $P(\Phi^1, \Phi^2)$ tal que $R_\sigma(P(\Phi^1, \Phi^2))(x) = (R_\sigma(\Phi^1)(x), R_\sigma(\Phi^2)(x))$ para todo $x \in \mathbb{R}^d$.

Definición 2.2.3. Sean $L_1, L_2 \in \mathbb{N}$ y sea

$$\Phi^1 = ((A_1^1, b_1^1), (A_2^1, b_2^1), \dots, (A_{L_1}^1, b_{L_1}^1)), \Phi^2 = ((A_1^2, b_1^2), (A_2^2, b_2^2), \dots, (A_{L_2}^2, b_{L_2}^2))$$

dos redes con dimensión de entrada d. Si $L_1 \leq L_2$, entonces definimos

$$P(\Phi^1, \Phi^2) := ((\tilde{A}_1, \tilde{b}_1), \dots, (\tilde{A}_{L_2}, \tilde{b}_{L_2}))$$

donde

$$\tilde{A}_l := \left[\begin{array}{c|cc|c|cc} A_{l,x_0}^1 & A_{l,x_1}^1 & 0 & \cdots & A_{l,x_{l-1}}^1 & 0 \\ A_{l,x_0}^2 & 0 & A_{l,x_1}^2 & & 0 & A_{l,x_{l-1}}^2 \end{array} \right], \tilde{b}_l := \begin{bmatrix} b_l^1 \\ b_l^2 \end{bmatrix}$$

para $1 \leq l < L_1$ y

$$\tilde{A}_l := \left[A_{l,x_0}^2 | 0 \quad A_{l,x_1}^2 | \cdots | 0 \quad A_{l,x_{L_1-1}}^2 | \cdots | A_{l,x_{L_1}}^2 | A_{l,x_{l-1}}^2 \right], \tilde{b}_l := b_l^2$$

para $L_1 \leq l < L_2$, y

$$\tilde{A}_l := \left[\begin{array}{c|cc|c|cc|c|c} A_{L_1,x_0}^1 & A_{L_1,x_1}^1 & 0 & \cdots & A_{L_1,x_{L_1-1}}^1 & 0 & 0 & \cdots & 0 \\ A_{L_2,x_0}^2 & 0 & A_{L_2,x_1}^2 & & 0 & A_{L_2,x_{L_1-1}}^2 & A_{L_2,x_{L_1}}^2 & & A_{L_2,x_{L_2-1}}^2 \end{array} \right]$$

y

$$\tilde{b}_{L_2} := \begin{bmatrix} b_{L_1}^1 \\ b_{L_2}^2 \end{bmatrix}.$$

De manera similar se puede definir para el caso $L_1 > L_2$. Entonces $P(\Phi^1, \Phi^2)$ es una red neuronal con dimensión de entrada d y máx $\{L_1, L_2\}$ capas, que llamaremos paralelización de Φ^1 y Φ^2.

Lema 2.2.3. *Sean Φ^1 y Φ^2 dos redes con dimensión de entrada d. Entonces $M(P(\Phi^1, \Phi^2)) = M(\Phi^1) + M(\Phi^2)$, $N(P(\Phi^1, \Phi^2)) = N(\Phi^1) + N(\Phi^2) - d$ y*

$$R_\sigma(P(\Phi^1, \Phi^2))(x) = (R_\sigma(\Phi^1)(x), R_\sigma(\Phi^2)(x)), \text{ para todo } x \in \mathbb{R}^d.$$

Observación 2.2.2. La Definición 2.2.3 puede garantizar la paralización para un número arbitrario de redes neuronales. Sea $n \in \mathbb{N}$ y para $i = 1, \ldots, n$, sea Φ^i redes neuronales de dimensión de entrada d y L_i capas. La red

$$P(\Phi^i : i = 1, \ldots, n) := P(\Phi^1, \ldots, \Phi^n) := P(\Phi^1, P(\Phi^2, P(\ldots, P(\Phi^{n-1}\Phi^n)\ldots)))$$

con dimensión d de entrada y $L = \max\{L_1, \ldots, L_n\}$ capas es llamada paralelización de $\Phi^1, \ldots, \Phi^n$. Tenemos que

$$M(P(\Phi^1, \ldots, \Phi^n)) = \sum_{i=1}^{n} M(\Phi^i), \quad \text{y} \quad N((P(\Phi^1, \ldots, \Phi^n)) = \sum_{i=1}^{n} N(\Phi^i) - (n-1)d,$$

y $R_\sigma(P(\Phi^1, \ldots, \Phi^n))(x) = (R_\sigma(\Phi^1)(x), \ldots, R_\sigma(\Phi^n)(x))$ para todo $x \in \mathbb{R}^d$. Esto se puede probar usando el principio de inducción sobre n, usando el Lema 2.2.3 para el caso base. Luego, suponer que vale para $n-1$ y luego usar el Lema 2.2.3 de nuevo pero ahora con $P(\Phi^1, \ldots, \Phi^{n-1})$ y Φ^n en lugar de Φ^1 y Φ^2, para después usar la hipótesis de inducción para obtener el resultado para n.

2.3. Redes neuronales estándar

Lema 2.3.1. *Existen constantes absolutas $C_1, C_2 > 0$ tales que para cualquier red neuronal Φ con función de activación ρ, dimensión de entrada d, L capas, $N = N(\Phi)$ neuronas y $M = M(\Phi)$ pesos no nulos, existe una red neuronal estándar Φ_{st} con dimensión de entrada d y L capas, a lo sumo C_1LN neuronas y a lo sumo $C_2(LN + M)$ pesos no nulos tal que*

$$R_\rho(\Phi)(x) = R_\rho(\Phi_{st})(x)$$

para todo $x \in \mathbb{R}^d$

Demostración.

Si $L = 1$, $\Phi_{st} = \Phi$ ya está.

Para $L > 1$, podemos definir las matrices,

$$A^{st}_{1,x_0} := \begin{bmatrix} Id_{\mathbb{R}^{N_0}} \\ -Id_{\mathbb{R}^{N_0}} \\ A_{1,x_0} \end{bmatrix}, b^{st}_1 := \begin{bmatrix} 0 \\ 0 \\ b_1 \end{bmatrix}.$$

y

$$A^{st}_{l,x_{l-1}} := \begin{bmatrix} Id_{\mathbb{R}^{2N_0}} & 0 \\ 0 & Id_{\mathbb{R}^{\Sigma_{k=1}^{l-1} N_k}} \\ A_{l,x_0} - A_{l,x_0} & A_{l,x_1} \ldots A_{l,x_{l-1}} \end{bmatrix}, b^{st}_l := \begin{bmatrix} 0 \\ 0 \\ b_l \end{bmatrix}$$

para $l = 2, \ldots, L-1$ y

$$A^{st}_{L,x_{L-1}} := \left[A_{L,x_0}| - A_{L,x_0}|A_{L,x_1}| \ldots |A_{L,x_{L-1}}\right], b^{st}_L := b_L.$$

Ahora pongamos

$$A^{st}_l := \left[0| \ldots |0|A^{st}_{l,x_{l-1}}\right], \quad \text{para} \quad l = 1, \ldots, L$$

y

$$\Phi_{st} \coloneqq ((A_1^{st}, b_1^{st}), (A_2^{st}, b_2^{st}), \ldots, (A_L^{st}, b_L^{st})).$$

Entonces Φ_{st} es una red neuronal estándar con dimensión de entrada d y L capas, y se puede ver que

$$R_\rho(\Phi)(x) = R_\rho(\Phi_{st})(x), \forall x \in \mathbb{R}^d.$$

Para el número de neuronas en la capa l $N_l^{st} = 2N_0 + \sum_{k=1}^{l} N_k \leq 2N(\Phi) = 2N$. Por lo que, $N(\Phi_{st}) = \sum_{k=1}^{l} N_k^{st} + N_0^{st} \leq 2NL$.

Para estimar el número de pesos, vemos que

$$\left\|A_l^{st}\right\|_{\ell^0} \leq 2N + 2\left\|A_l\right\|_{\ell^0}, \left\|b_l^{st}\right\|_{\ell^0} = \left\|b_l\right\|_{\ell^0}$$

para $l = 1, \ldots, L$. De donde concluimos que $M(\Phi_{st}) \leq \sum_{l=1}^{L} 2N + 2\left\|A_l\right\|_{\ell^0} + \left\|b_l\right\|_{\ell^0} \leq 2NL + 2M = 2(NL + M)$. Por lo que obtenemos el resultado considerando $C_1 = C_2 = 2$. □

Ilustramos el Lema anterior con el siguiente ejemplo, donde Φ es una red neuronal con dimensión de entrada $d = 1$ y dimensión de salida 1 de 2 capas:

$$\Phi = ((A_1, b_1), (A_2, b_2)),$$

donde $A_1 = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}, b_1 = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}, A_2 = \begin{bmatrix} c_4 & c_5 & c_6 & c_7 \end{bmatrix} c, b_2 = d_4$, donde $c_1, c_2, c_3, c_4, c_5, c_6, c_7, d_1, d_2, d_3, d_4 \in \mathbb{R}$.

Entonces por el Lema 2.3.1, $\Phi^{st} = ((A_1^{st}, b_1^{st}), (A_2^{st}, b_2^{st}))$ es una red neuronal estándar donde

$$A_1^{st} = \begin{bmatrix} 1 \\ -1 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix}, b_1^{st} = \begin{bmatrix} 0 \\ 0 \\ d_1 \\ d_2 \\ d_3 \end{bmatrix}, A_2^{st} = \begin{bmatrix} 0 & 0 & 0 & c_4 & -c_4 & c_5 & c_6 & c_7 \end{bmatrix}, b_2^{st} = d_4$$

Verifiquemos que $R(\Phi_\rho^{st})(x) = R_\rho(\Phi)(x), \quad \forall x \in \mathbb{R}$.

Sea $x \in \mathbb{R}$. Para Φ se tiene que

$$\begin{aligned}
&x_0 = x \\
&x_1 = \rho\left(A_1 x_0 + b_1\right) = \rho\left(\begin{bmatrix} c_1 x + d_1 \\ c_2 x + d_2 \\ c_3 x + d_3 \end{bmatrix}\right) = \begin{bmatrix} \rho(c_1 x + d_1) \\ \rho(c_2 x + d_2) \\ \rho(c_3 x + d_3) \end{bmatrix}, \\
&x_2 = A_{2,x_0} x + A_{2,x_1} x_1 + b_2 = \\
&\quad = c_4 x + c_5 \rho(c_1 x + d_1) + c_6 \rho(c_2 x + d_2) + c_7 \rho(c_3 x + d_3) + d_4,
\end{aligned}$$

Por lo que resulta, $R_\rho(\Phi)(x) = x_3 = c_4 x + c_5 \rho(c_1 x + d_1) + c_6 \rho(c_2 x + d_2) + c_7 \rho(c_3 x + d_3) + d_4$.

Para Φ^{st} se tiene que

$$x_0^{st} = x$$

$$x_1^{st} = \rho\left(A_1^{st}x_0 + b_1^{st}\right) = \rho\left(\begin{bmatrix} x \\ -x \\ c_1x + d_1 \\ c_2x + d_2 \\ c_3x + d_3 \end{bmatrix}\right) = \begin{bmatrix} \rho(x) \\ \rho(-x) \\ \rho(c_1x + d_1) \\ \rho(c_2x + d_2) \\ \rho(c_3x + d_3) \end{bmatrix},$$

$$\begin{aligned} x_2^{st} &= A_{2,x_1}^{st}x_1 + b_2^{st} = \\ &= c_4\rho(x) - = c_4\rho(-x) + c_5\rho(c_1x + d_1) + c_6\rho(c_2x + d_2) + c_7\rho(c_3x + d_3) + d_4 = \\ &= c_4(\rho(x) - \rho(-x)) + c_5\rho(c_1x + d_1) + c_6\rho(c_2x + d_2) + c_7\rho(c_3x + d_3) + d_4 = \\ &= c_4x + c_5\rho(c_1x + d_1) + c_6\rho(c_2x + d_2) + c_7\rho(c_3x + d_3) + d_4, \end{aligned}$$

usando el hecho de que $\rho(x) - \rho(-x) = x$, que ya lo hemos en la demostración del 2.2.2 para d variables. Y así, resulta $R_\rho(\Phi)(x) = x_2^{st} = c_4x + c_5\rho(c_1x + d_1) + c_6\rho(c_2x + d_2) + c_7\rho(c_3x + d_3) + d_4$.

Entonces, de lo anterior se tiene

$$R_\rho(\Phi)(x) = x_3 = c_4x + c_5\rho(c_1x + d_1) + c_6\rho(c_2x + d_2) + c_7\rho(c_3x + d_3) + d_4 = R_\rho(\Phi)(x).$$

Por lo tanto, como $x \in \mathbb{R}$ es arbitrario, $R(\Phi_\rho^{st})(x) = R_\rho(\Phi)(x), \quad \forall x \in \mathbb{R}$.

Observación 2.3.1. Observemos que en el ejemplo anterior se usan propiedades particulares de la función ReLU, es por eso que el Lema 2.3.1 se restringe a esta función de activación.

Corolario 2.3.1. *Sean Φ^1 y Φ^2 redes neuronales con función de activación ρ, con la misma arquitectura $\mathcal{A} = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$ con L capas, N neuronas, y M pesos no nulos. Entonces las correspondientes redes neuronales estándar del Lema 2.3.1 Φ_{st}^1 y Φ_{st}^2 pueden tomarse con una misma arquitectura estándar $\mathcal{A}_{st}$ con a lo sumo C_1LN neuronas y a lo sumo $C_2(LN + M)$ pesos no nulos, donde C_1 y C_2 coinciden con las del Lema mencionado.*

Demostración. Sean $\Phi^1 = ((\tilde{A}_1^1, \tilde{b}_1^1), \ldots, (\tilde{A}_L^1, \tilde{b}_L^1))$ y $\Phi^2 = ((\tilde{A}_1^2, \tilde{b}_1^2), \ldots, (\tilde{A}_L^2, \tilde{b}_L^2))$ tienen la misma arquitectura $\mathcal{A} = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$.

Sean $\Phi_{st}^1 = ((\tilde{A}_1^{1,st}, \tilde{b}_1^{1,st}), (\tilde{A}_2^{1,st}, \tilde{b}_2^{1,st}), \ldots, (\tilde{A}_L^{1,st}, \tilde{b}_L^{1,st}))$ y $\Phi_{st}^2 = ((\tilde{A}_1^{2,st}, \tilde{b}_1^{2,st}), (\tilde{A}_2^{2,st}, \tilde{b}_2^{2,st}), \ldots, (\tilde{A}_L^{2,st}, \tilde{b}_L^{2,st}))$ las redes estándar construidas en el Lema 2.3.1 para Φ^1 y Φ^2 respectivamente.

Veamos que existe una arquitectura estándar $\mathcal{A}_{st}$ tal que las redes neuronales estándar Φ_{st}^1 y Φ_{st}^2 tienen dicha arquitectura.

Para esto, primero construyamos la arquitectura estándar $\mathcal{A}_{st}$.

Si $L = 1$, $\Phi_{st}^1 = \Phi^1$, $\Phi_{st}^2 = \Phi^2$, $\mathcal{A}_{st} := \mathcal{A}$ y ya está.

Para $L > 1$.

Usaremos una construcción parecida a la demostración del Lema 2.3.1. Definimos

$$A_{1,x_0}^{st} := \begin{bmatrix} Id_{\mathbb{R}^{N_0}} \\ Id_{\mathbb{R}^{N_0}} \\ A_{1,x_0} \end{bmatrix}, b_1^{st} := \begin{bmatrix} 0 \\ 0 \\ b_1 \end{bmatrix}.$$

y

$$A^{st}_{l,x_{l-1}} := \begin{bmatrix} Id_{\mathbb{R}^{2N_0}} & 0 \\ 0 & Id_{\mathbb{R}^{\sum_{k=1}^{l-1} N_k}} \\ A_{l,x_0} \quad A_{l,x_0} & A_{l,x_1} \dots A_{l,x_{l-1}} \end{bmatrix}, b^{st}_l := \begin{bmatrix} 0 \\ 0 \\ b_l \end{bmatrix}$$

para $l = 2, \dots, L-1$ y

$$A^{st}_{L,x_{L-1}} := [A_{L,x_0}|A_{L,x_0}|A_{L,x_1}| \dots |A_{L,x_{L-1}}], b^{st}_L := b_L.$$

Ahora pongamos

$$A^{st}_l := [0| \dots |0|A^{st}_{l,x_{l-1}}], \quad \text{para} \quad l = 1, \dots, L$$

y

$$\mathcal{A}_{st} \coloneqq ((A^{st}_1, b^{st}_1), (A^{st}_2, b^{st}_2), \dots, (A^{st}_L, b^{st}_L).)$$

Primero vemos que por cómo la definimos, $\mathcal{A}_{st}$ es una red neuronal estándar con entrada de dimensión d y L capas. Además, las componentes de las matrices $A^{st}_1, \dots, A^{st}_L$ son las componentes de las matrices $Id_{\mathbb{R}^{N_0}}, Id_{\mathbb{R}^{2N_0}}, 0, A_1, \dots, A_L$ que son 1 o 0 pues $\mathcal{A}$ es arquitectura, y las otras matrices identidades o nulas, y las componentes de los vectores $b^{st}_1, \dots, b^{st}_L$ son las componentes de los vectores $b_1, \dots, b_L$ que son 1 o 0 pues $\mathcal{A}$ es arquitectura y el vector 0 que son 0. Por lo que, las componentes de las matrices $A^{st}_1, \dots, A^{st}_L$ y las componentes de los vectores $b^{st}_1, \dots, b^{st}_L$ son 1 o 0. Luego, $\mathcal{A}_{st}$ es una arquitectura estándar.

Falta ver Φ^1_{st} y Φ^2_{st} tengan la misma arquitectura $\mathcal{A}_{st}$.

Para esto verificamos para Φ^1_{st} y Φ^2_{st} las propiedades $a), b)$ de la Definición 2.1.3.

Primero por ser que Φ^1 y Φ^2 tienen la misma arquitectura $\mathcal{A}$, $N_l(\Phi^k) = N_l(\mathcal{A})$ con $k = 1, 2$ $\forall l = 1, \dots, L$. Luego, usando lo anterior y por cómo construimos $\mathcal{A}_{st}$ resulta $N_l(\Phi^k_{st}) = N_l(\mathcal{A}_{st})$ con $k = 1, 2$, $\forall l = 1, \dots, L$. Por lo que se verifica $a)$ para Φ^1_{st} y Φ^2_{st}.

Para ver $b)$, para $l = 1$, si alguna de las componentes de la matriz A^{st}_1 es nula, entonces es componente nula de la matriz identidad, o es una componente nula de la matriz A_1. Si la componente es de la matriz identidad, por construcción, esa componente va a ser nula para las matrices $\tilde{A}^{1,st}_1$, $\tilde{A}^{2,st}_1$. Si la componente es de la matriz A_1, como Φ^1 y Φ^2 tienen la misma arquitectura $\mathcal{A}$, esa componente va a ser nula para las matrices $\tilde{A}^1_1$, $\tilde{A}^2_1$ y por lo cual, esa componente va a ser nula para las matrices $\tilde{A}^{1,st}_1$, $\tilde{A}^{2,st}_1$, por construcción.

Para $l > 1$, si alguna de las componentes de la matriz A^{st}_l es nula entonces es la componente nula de la matriz nula o de la matriz identidad, o es la componente nula de la matriz A_{l,x_i} para algún $i = 1, \dots, l-1$, es decir, de la matriz A_l. Si es componente de la matriz nula o la matriz identidad, por construcción esa componente va a ser nula para las matrices $\tilde{A}^{1,st}_l$, $\tilde{A}^{2,st}_l$. Si la componente es de la matriz A_l, como Φ_1 y Φ_2 tienen la misma arquitectura $\mathcal{A}$, esa componente va a ser nula para las matrices $\tilde{A}^1_l$, $\tilde{A}^2_l$ y por lo cual, esa componente va a ser nula para las matrices $\tilde{A}^{1,st}_l$, $\tilde{A}^{2,st}_l$, por construcción.

Por lo que, de lo anterior podemos concluir que $[A^{st}_l]_{i,j} = 0$ implica que $\left[\tilde{A}^{k,st}_l\right]_{i,j} = 0$ para todo $l = 1, \dots, N$, con $i = 1, \dots, N^{st}_l$ y $j = 1, \dots, \sum_{r=1}^{l-1} N^{st}_r$, y $k = 1, 2$.

De manera parecida a lo anterior podemos ver que y $[b^{st}_l]_i = 0$ implica $\left[\tilde{b}^{k,st}_l\right]_i = 0$ para todo $l = 1, \dots, L$, con $i = 1, \dots, N_l$ y $k = 1, 2$. Con lo cual se verifica $b)$ para Φ^1_{st} y Φ^2_{st}.

En conclusión, obtenemos que Φ^1_{st} y Φ^2_{st} tienen la misma arquitectura $\mathcal{A}_{st}$.

De manera análoga a la demostración del Lema 2.3.1 se puede probar que $\mathcal{A}^{st}$ tiene a lo sumo $C_1 LN$ neuronas y a lo sumo $C_2(LN + M)$ pesos no nulos. □

Capítulo 3

Aproximación de funciones en normas de Sobolev

Este Capítulo contiene los resultados principales de la Tesis, sobre cotas de complejidad superiores e inferiores para la aproximación por redes neuronales de funciones en espacios de Sobolev. Referimos al Apéndice A.2 para resultados preliminares sobre espacios de Sobolev que se usan aquí.

3.1. Cotas superiores

Presentamos el siguiente resultado, que son cotas superiores de complejidad para aproximaciones de funciones de ciertos espacios de Sobolev mediante redes neuronales profundas con función de activación ReLU.

De ahora en adelante usaremos la función de activación ρ que es la ReLU. Nos interesa la aproximación de funciones definidas en subconjuntos de Espacios de Sobolev $W^{n,p}((0,1)^d)$ con realización de redes neuronales. Para esto definimos el conjunto:

$$\mathcal{F}_{n,d,p,B} \coloneqq \{f \in W^{n,p}((0,1)^d) : ||f||_{W^{n,p}((0,1)^d)} \leq B\}.$$

Teorema 3.1.1. *Sea $d \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty, B > 0, 0 \leq s \leq 1$. Entonces, existe una constante $c = c(d,n,p,B,s) > 0$ con las siguientes propiedades:*

Para todo $\varepsilon \in (0, \frac{1}{2})$ existe una arquitectura de red neuronal $\mathcal{A}_\varepsilon = \mathcal{A}_\varepsilon(d,n,p,B,s,\varepsilon)$ con dimensión de entrada d y dimensión de salida 1 tal que $\forall f \in \mathcal{F}_{n,d,p,B}$ hay una red neuronal Φ_ε^f con arquitectura $\mathcal{A}_\varepsilon$ tal que

$$\left\| R_\rho(\Phi_\varepsilon^f) - f \right\|_{W^{s,p}((0,1)^d)} \leq \varepsilon$$

y

(I) $L(\mathcal{A}_\varepsilon) \leq c \log_2(\varepsilon^{-n/(n-s)})$;

(II) $M(\mathcal{A}_\varepsilon) \leq c\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)})$;

(III) $N(\mathcal{A}_\varepsilon) \leq c\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)})$.

La demostración de este Teorema se hará al final de este capitulo en 3.1.3, pues para esto usaremos los resultados que demostraremos en las subsecciones 3.1.1 y 3.1.2.

Corolario 3.1.1. *Sea $d \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty, B > 0, 0 \leq s \leq 1$. Entonces, existe una constante $c = c(d, n, p, B, s) > 0$ con las siguientes propiedades:*

Para todo $\varepsilon \in (0, \frac{1}{2})$ existe una arquitectura de red neuronal estándar $\mathcal{A}_{st,\varepsilon} = \mathcal{A}_{st,\varepsilon}(d, n, p, B, s, \varepsilon)$ con dimensión de entrada d y dimensión de salida 1 tal que $\forall f \in \mathcal{F}_{n,d,p,B}$ hay una red neuronal estándar $\Phi^f_{st,\varepsilon}$ con arquitectura $\mathcal{A}_{st,\varepsilon}$ tal que

$$\left\| R_\rho(\Phi^f_{st,\varepsilon}) - f \right\|_{W^{s,p}((0,1)^d)} \leq \varepsilon$$

y

(I) $L(\mathcal{A}_{st,\varepsilon}) \leq c \log_2(\varepsilon^{-n/(n-s)})$;

(II) $M(\mathcal{A}_{st,\varepsilon}) \leq c\varepsilon^{-d/(n-s)} \log_2^2(\varepsilon^{-n/(n-s)})$;

(III) $N(\mathcal{A}_{st,\varepsilon}) \leq c\varepsilon^{-d/(n-s)} \log_2^2(\varepsilon^{-n/(n-s)})$.

Demostración. Sea $\varepsilon \in (0, \frac{1}{2})$, por Teorema 3.1.1, existe una arquitectura de red neuronal $\mathcal{A}_\varepsilon = \mathcal{A}_\varepsilon(d, n, p, B, s, \varepsilon)$ con dimensión de entrada d y dimensión de salida 1 tal que $\forall f \in \mathcal{F}_{n,d,p,B}$ hay una red neuronal Φ^f_ε con arquitectura $\mathcal{A}_\varepsilon$ tal que

$$\left\| R_\rho(\Phi^f_\varepsilon) - f \right\|_{W^{s,p}((0,1)^d)} \leq \varepsilon$$

y para $c' = c'(d, n, p, B, s) > 0$ constante,

(I) $L(\mathcal{A}_\varepsilon) \leq c' \log_2(\varepsilon^{-n/(n-s)})$;

(II) $M(\mathcal{A}_\varepsilon) \leq c'\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)})$;

(III) $N(\mathcal{A}_\varepsilon) \leq c'\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)})$.

Por el Corolario 2.3.1, existe una red neuronal estándar $\Phi^f_{st,\varepsilon}$ con una arquitectura estándar $\mathcal{A}_{st,\varepsilon}$ que depende de $\mathcal{A}$ y así $d, n, p, B, s, \varepsilon$ con a lo sumo $C_1 L(\mathcal{A}_\varepsilon) N(\mathcal{A}_\varepsilon)$ neuronas y a lo sumo $C_2(L(\mathcal{A}_\varepsilon)N(\mathcal{A}_\varepsilon) + M(\mathcal{A}_\varepsilon))$ pesos no nulos, $C_1, C_2 > 0$ constantes absolutas. Y además,

$$R_\rho(\Phi^f)(x) = R_\rho(\Phi^f_{st,\varepsilon})(x), \forall x \in \mathbb{R}^d.$$

Por lo que, existe una arquitectura de red neuronal estándar $\mathcal{A}_{st,\varepsilon} = \mathcal{A}_{st,\varepsilon}(d, n, p, B, s, \varepsilon)$ con dimensión de entrada d y dimensión de salida 1 tal que $\forall f \in \mathcal{F}_{n,d,p,B}$ hay una red neuronal estándar $\Phi^f_{st,\varepsilon}$ con arquitectura $\mathcal{A}_{st,\varepsilon}$ tal que

$$\|R_\rho(\Phi^f_{st,\varepsilon}) - f\|_{W^{s,p}((0,1)^d)} = \|R_\rho(\Phi^f_\varepsilon) - f\|_{W^{s,p}((0,1)^d)} \leq \varepsilon$$

Y además, tiene a

(I) $L(\mathcal{A}_{st,\varepsilon}) = L(\mathcal{A}_\varepsilon) \leq c' \log_2(\varepsilon^{-n/(n-s)})$;

(II) $M(\mathcal{A}_{st,\varepsilon}) \leq C_2(L(\mathcal{A}_\varepsilon)N(\mathcal{A}_\varepsilon)+M(\mathcal{A}_\varepsilon)) \leq C_2(c'^2 \log_2(\varepsilon^{-n/(n-s)})\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)})$ $+ c'\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)})) \leq C_2 \max\{c', c'^2\} \cdot \varepsilon^{-d/(n-s)} \cdot \log_2^2(\varepsilon^{-n/(n-s)})$;

(III) $N(\mathcal{A}_{st,\varepsilon}) \leq C_1 L(\mathcal{A}_\varepsilon) N(\mathcal{A}_\varepsilon) \leq C_1 c' \log_2(\varepsilon^{-n/(n-s)}) c'\varepsilon^{-d/(n-s)} \log_2(\varepsilon^{-n/(n-s)}) =$ $C_1 c'^2 \varepsilon^{-d/(n-s)} \log_2^2(\varepsilon^{-n/(n-s)})$

usando el hecho de $\log_2(\varepsilon^{-n/(n-s)}) \geq 1$, que es fácil de ver.

Tomando $c = \max\{c', C_1 c'^2, C_2 \max\{c', c'^2\}\}$ obtenemos el resultado. □

3.1.1. Algunos casos particulares: aproximación de x^2 y xy

Para realizar la demostración del Teorema 3.1.1 se busca lo siguiente : La aproximación polinomial P_f de una función $f \in \mathcal{F}_{n,d,p,B}$ por una red neuronal Φ_{P_f}. Empezamos con la construcción de una red neuronal que aproxime la función cuadrática en el $(0,1)$ con un error de aproximación que de a lo sumo ε . El error es medido en la norma $W^{1,\infty}$ (Proposición 3.1.2). Este resultado se utiliza luego para obtener una aproximación de un operador de multiplicación (Proposición 3.1.4). Luego, derivamos una partición de la unidad (Lema 3.1.1) como el producto de funciones invariantes que pueden ser realizadas por redes neuronales. Usando esta construcción, entonces creamos una aproximación de Taylor localizada para una función $f \in \mathcal{F}_{n,d,p,B}$ en las normas L^p y $W^{1,\infty}$. Un argumento de interpolación muestra la misma construcción para la norma $W^{s,p}$ donde $0 \leq s \leq 1$ (Lema 3.1.2). En el siguiente paso, establecemos la base para la aproximación de productos de la partición de la unidad y monomios, es decir, monomios localizados, con redes ReLU (Lema 3.1.3) en la norma $W^{k,\infty}$ para $k = 0, 1$. Por un embebimiento, este resultado es entonces usado para la aproximación de polinomios localizados en $W^{k,p}$ con redes ReLU (Lema 3.1.4).

Todos los espacios de Sobolev usados en estas pruebas son dominios abiertos, acotados y convexos.

Para probar la Proposición 3.1.2, vamos a usar la siguiente Proposición que se puede encontrar en [13, Proposición 2].

Proposición 3.1.1. *La función $f(x) = x^2$ en el segmento $[0,1]$ puede ser aproximada con cualquier error $\varepsilon > 0$ por una red neuronal que tiene capas, número de pesos y neuronas del orden $\mathcal{O}\left(\log_2\left(\frac{1}{\varepsilon}\right)\right)$.*

Demostración. Consideremos la función $g : [0,1] \to [0,1]$,

$$g(x) = \begin{cases} 2x & \text{si} \quad 0 \leq x < \frac{1}{2}, \\ 2(1-x) & \text{si} \quad \frac{1}{2} \leq x \leq 1, \end{cases}$$

y las funciones

$$g_s(x) = \underbrace{g \circ g \circ \cdots \circ g}_{s}.$$

Por ejemplo,

$$g_1(x) = g(x),$$

y

$$g_2(x) = \begin{cases} 2x & \text{si} \quad 0 \leq x < \frac{1}{2}, \\ 2(1-x) & \text{si} \quad \frac{1}{2} \leq x \leq 1, \end{cases}$$

Luego, se puede ver que g_s que

$$\begin{cases} 2^s\left(x - \frac{2k}{2^s}\right) & \text{si} \quad x \in [\frac{2k}{2^s}, \frac{2k+1}{2^s}], k = 0, 1, \ldots, 2^{s-1} - 1, \\ 2^s\left(\frac{2k}{2^s} - x\right) & \text{si} \quad \frac{1}{2}x \in [\frac{2k-1}{2^s}, \frac{2k}{2^s}], k = 0, 1, \ldots, 2^{s-1}, \end{cases}$$

lo que se puede encontrar en [12, Lema 2.4].

Lo que observamos ahora es que la función $f(x) = x^2$ puede ser aproximada por la combinación lineal de funciones g_s. Sea f_m la interpolación lineal de f con $2^m + 1$

puntos uniformemente distribuidos $\frac{k}{2^m}, k = 0, \dots, 2^m$ en el intervalo $[0, 1]$: esto es, f_m verifica

$$f_m\left(\frac{k}{2^m}\right) = \left(\frac{k}{2^m}\right)^2, \qquad k = 0, \dots, 2^m,$$

y es lineal en cada intervalo $[\frac{k}{2^m}, \frac{k+1}{2^m}]$. Es fácil ver que f_m aproxima a f con un error $\varepsilon_m = 2^{-2m-2}$ (en norma $\|\cdot\|_\infty$). Además, f_{m-1} y f_m se diferencian en una función proporcional a g_m:

$$f_{m-1}(x) - f_m(x) = \frac{g_m(x)}{2^{2m}}.$$

Entonces, teniendo en cuenta que $f_0(x) = x$ obtenemos

$$f_m(x) = x - \sum_{s=1}^{m} \frac{g_s(x)}{2^{2s}}.$$

Construiremos redes neuronales Φ_m tal que $R_\rho(\Phi_m)(x) = f_m(x)$.

Para $m = 1$.

Primero $f_1(x) = x - \frac{g_1(x)}{2^2} = x - \frac{g(x)}{4}$. Para f_1, podemos armar, usando que $g(x) = 2\rho(x) - 4\rho(x - \frac{1}{2}) + 2\rho(x - 1)$ en $[0, 1]$ que se comprueba fácilmente. Notar que esta igualdad vale en $\mathbb{R}$ si extendemos a g por 0 fuera de $[0, 1]$. Definimos $\Phi_1 \coloneqq ((A_1^1, b_1^1), (A_2^1, b_2^1))$ donde

$$A_1^1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, b_1^1 = \begin{bmatrix} 0 \\ -\frac{1}{2} \\ -1 \end{bmatrix},$$

y

$$A_2^1 = \begin{bmatrix} 1 & -\frac{1}{2} & 1 & -\frac{1}{2} \end{bmatrix}, b_2^1 = 0.$$

Luego, sea $x \in \mathbb{R}$

$$x_0 = x,$$

$$x_1 = \rho\left(A_1^1 x + b_1^1\right) = \rho\left(\begin{bmatrix} x \\ x \\ x \end{bmatrix} - \begin{bmatrix} 0 \\ -\frac{1}{2} \\ -1 \end{bmatrix}\right) = \begin{bmatrix} \rho(x) \\ \rho\left(x - \frac{1}{2}\right) \\ \rho(x - 1) \end{bmatrix},$$

$$x_2 = A_2^1 \begin{bmatrix} x \\ x_1 \end{bmatrix} + b_2^1 = \begin{bmatrix} 1 & -\frac{1}{2} & 1 & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} x \\ \rho(x) \\ \rho(x - \frac{1}{2}) \\ \rho(x - 1) \end{bmatrix} + 0 =$$

.

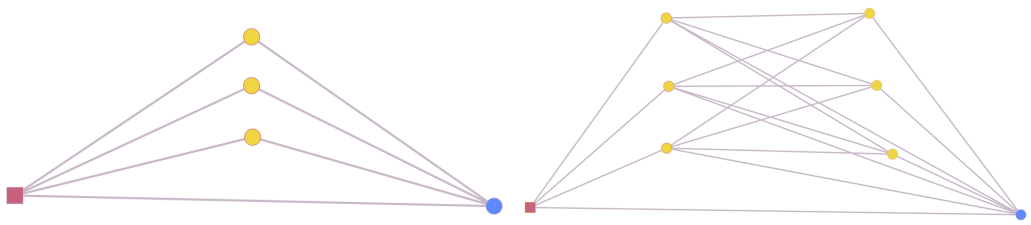
Así resulta,

$$R_\rho(\Phi_1)(x) = x_2 = x - \frac{\rho(x)}{2} + \rho\left(x - \frac{1}{2}\right) + \frac{\rho(x - 1)}{2}.$$

Entonces Φ_1 es una red neuronal de dimensión de entrada 1 y dimensión de salida 1 con $2 = 1 + 1$ capas, $5 = 3 \cdot 1 + 2$ neuronas, y $9 \leq 11 = 15 \cdot 1 - 4$ pesos no nulos tal que $R_\rho(\Phi_1)(x) = x - g(x) = f_1(x), \quad \forall x \in [0, 1]$.

Ahora, para f_2. Podemos ver que $f_2(x) = x - \frac{g_1(x)}{2^2} - \frac{g_2(x)}{2^4} = x - \frac{g(x)}{4} - \frac{g_2(x)}{16}$, que se ve fácilmente, definimos $\Phi_2 \coloneqq ((A_1^2, b_1^2), (A_2^2, b_2^2), (A_3^2, b_3^2))$ donde

$$A_1^2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, b_2^1 = \begin{bmatrix} 0 \\ -\frac{1}{2} \\ -1 \end{bmatrix},$$



(a) Red neuronal Φ_1 (b) Red neuronal Φ_2

Figura 3.1: Redes neuronales Φ_1,Φ_2

$$A_2^2 = \begin{bmatrix} 0 & 2 & -4 & 2 \\ 0 & 2 & -4 & 2 \\ 0 & 2 & -4 & 2 \end{bmatrix}, b_2^2 = \begin{bmatrix} 0 \\ -\frac{1}{2} \\ -1 \end{bmatrix},$$

y

$$A_3^2 = \begin{bmatrix} 1 & -\frac{2}{4} & \frac{4}{4} & -\frac{2}{4} & -\frac{1}{8} & \frac{1}{4} & -\frac{1}{8} \end{bmatrix}, b_3^2 = 0.$$

Luego, sea $x \in \mathbb{R}$

$$x_0 = x,$$

$$x_1 = \rho\left(A_1^2 x + b_1^2\right) = \rho\left(\begin{bmatrix} x \\ x \\ x \end{bmatrix} - \begin{bmatrix} 0 \\ -\frac{1}{2} \\ -1 \end{bmatrix}\right) = \begin{bmatrix} \rho(x) \\ \rho\left(x - \frac{1}{2}\right) \\ \rho(x-1) \end{bmatrix},$$

$$x_2 = \rho\left(A_2^2 \begin{bmatrix} x \\ x_1 \end{bmatrix} + b_2^2\right) = \rho\left(\begin{bmatrix} 0 & 2 & -4 & 2 \\ 0 & 2 & -4 & 2 \\ 0 & 2 & -4 & 2 \end{bmatrix} \begin{bmatrix} x \\ \rho(x) \\ \rho(x-\frac{1}{2}) \\ \rho(x-1) \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{1}{2} \\ -1 \end{bmatrix}\right) =$$

$$= \begin{bmatrix} \rho(2\rho(x) - 4\rho(x-\frac{1}{2}) + 2\rho(x-1) - 0) \\ \rho(2\rho(x) - 4\rho(x-\frac{1}{2}) + 2\rho(x-1) - \frac{1}{2}) \\ \rho(2\rho(x) - 4\rho(x-\frac{1}{2}) + 2\rho(x-1) - 1) \end{bmatrix} = \begin{bmatrix} \rho(g(x)) \\ \rho\left(g(x) - \frac{1}{2}\right) \\ \rho(g(x)-1) \end{bmatrix},$$

$$x_3 = A_2^3 \begin{bmatrix} x \\ x_1 \\ x_2 \end{bmatrix} + b_2^3 = \begin{bmatrix} 1 & -\frac{2}{4} & \frac{4}{4} & -\frac{2}{4} & -\frac{1}{8} & \frac{1}{4} & -\frac{1}{8} \end{bmatrix} \begin{bmatrix} x \\ \rho(x) \\ \rho(x-\frac{1}{2}) \\ \rho(x-1) \\ \rho(g(x)) \\ \rho\left(g(x)-\frac{1}{2}\right) \\ \rho(g(x)-1) \end{bmatrix} + 0 =$$

$$= x - \frac{\rho(x)}{2} + \rho\left(x - \frac{1}{2}\right) - \frac{\rho(x-1)}{2} - \frac{\rho(g(x))}{8} + \frac{\rho(g(x)-\frac{1}{2})}{4} - \frac{\rho(g(x)-1)}{8}.$$

Así resulta,

$$R_\rho(\Phi_2)(x) = x_3 = x - \frac{\rho(x)}{2} + \rho\left(x - \frac{1}{2}\right) - \frac{\rho(x-1)}{2} - \frac{\rho(g(x))}{8} + \frac{\rho(g(x)-\frac{1}{2})}{4} - \frac{\rho(g(x)-1)}{8}.$$

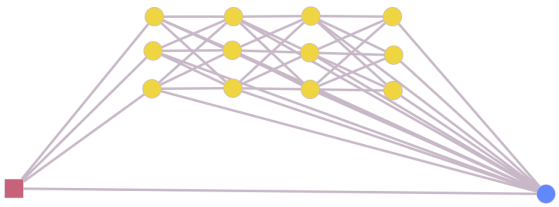


Figura 3.2: Red neuronal Φ_4

Entonces Φ_2 es una red neuronal de dimensión de entrada 1 y dimensión de salida 1 con $3 = 2 + 1$ capas, $8 = 3 \cdot 2 + 2$ neuronas y $23 \leq 26 = 15 \cdot 2 - 4$ pesos no nulos, tal que $R_\rho(\Phi_2)(x) = x - \frac{g_1(x)}{2^2} - \frac{g_2(x)}{2^4} = f_2(x), \forall x \in [0, 1]$.

El análisis anterior para f_1 y f_2 puede generalizarse a f_m, ver por ejemplo el caso $m = 4$ en Figura 3.2, concluyendo que existe una red neuronal Φ_m con función realización $R_\rho(\Phi_m) = f_m$ y con $m + 1$ capas, $3m + 2 = \mathcal{O}(m)$ neuronas, y a lo sumo $3 + 9(m - 1) + 3m + 1 + 3m + 1 = 15m - 4 = \mathcal{O}(m)$ pesos no nulos. □

La próxima proposición es fundamental para demostrar el resultado principal de este Capítulo. Se prueba que la función x^2 puede aproximarse en norma $W^{1,\infty}(0, 1)$ por redes neuronales con complejidad dependiendo logarítmicamente del grado de aproximación. El resultado similar para norma $L^\infty(0, 1)$ está demostrado en [13, Proposoción 2].

Proposición 3.1.2. *Por la Proposición 3.1.1, existen constantes $c_1, c_2, c_3, c_4 > 0$ tal que $\forall \varepsilon \in (0, \frac{1}{2})$ existe una red neuronal Φ_ϵ^{sq} con a lo sumo $c_1 \log_2(1/\epsilon)$ pesos no nulos, a lo sumo $c_2 \log_2(1/\epsilon)$ capas, a lo sumo $c_3 \log_2(1/\epsilon)$ neuronas con 1 dimensión de entrada y de salida tal que*

$$\left\| R_\rho(\Phi_\varepsilon^{sq})(x) - x^2 \right\|_{W^{1,\infty}((0,1),dx)} \leq \varepsilon \tag{3.1}$$

y $R_\rho(\Phi_\varepsilon^{sq})(0) = 0$. Más aún, vale que

$$|R_\rho(\Phi_\varepsilon^{sq})(x)|_{W^{1,\infty}((0,1),dx)} \leq c_4 \tag{3.2}$$

Demostración. En , se muestra que existen constantes $c_1, c_2, c_3 > 0$ tal que para todo $m \in \mathbb{N}$ existe una red neuronal Φ_m con a lo sumo $c_1 m$ pesos no nulos, a lo sumo $c_2 m$ capas, a lo sumo $c_3 m$ neuronas cuya función realización es una interpolación lineal por partes de la función $f : x \mapsto x^2$ sobre el intervalo $(0, 1)$. En detalle, se muestra que existe una red Φ_m satisfaciendo para todo $k \in \{0, \dots 2^m - 1\}$ y todo $x \in \left[\frac{k}{2^m}, \frac{k+1}{2^m}\right]$

$$R_\rho(\Phi_m)(x) = \left(\frac{(k+1)^2}{2^m} - \frac{k^2}{2^m}\right)\left(x - \frac{k}{2^m}\right) + \left(\frac{k}{2^m}\right)^2. \tag{3.3}$$

Así, $R_\rho(\Phi_m)$ es una interpolante lineal por partes de f con $2^m + 1$ puntos uniformemente distribuidos $\frac{k}{2^m}$, $k = 0, \dots, 2^m$. En particular, como $0 \in [0, \frac{1}{2^m}]$ entonces $k = 0$ y $R_\rho(\Phi_m)(0) = \left(\frac{(1)^2}{2^m} - \frac{0^2}{2^m}\right)\left(0 - \frac{0}{2^m}\right) + \left(\frac{0}{2^m}\right)^2 = \frac{1}{2^m}0 = 0$. Más aún, en 3.1.1 mostramos que

$$\left\| R_\rho(\Phi_m)(x) - x^2 \right\|_{L^\infty((0,1))} \leq 2^{-2-2m}. \tag{3.4}$$

Mostraremos que

$$|R_\rho(\Phi_m)(x) - x^2|_{W^{1,\infty}((0,1),dx)} \leq 2^{-m}. \tag{3.5}$$

Para esto, sea $k \in \{0, \ldots, 2^m - 1\}$ usando la Ecuación (3.3) se tiene que

$$|R_\rho(\Phi_m)(x) - x^2|_{W^{1,\infty}((k/2^m,(k+1)/2^m),dx)} = \left|\left|\frac{d}{dx}(R_\rho(\Phi_m)(x) - x^2)\right|\right|_{L^\infty((k/2^m,(k+1)/2^m),dx)} =$$
$$= \left|\left|\frac{(k+1)^2}{2^m} - \frac{k^2}{2^m} - 2x\right|\right|_{L^\infty((k/2^m,(k+1)/2^m),dx)} = \left|\left|\frac{k^2+2k+1-k^2}{2^m} - 2x\right|\right|_{L^\infty((k/2^m,(k+1)/2^m),dx)} =$$
$$= \left|\left|\frac{2k+1}{2^m} - 2x\right|\right|_{L^\infty((k/2^m,(k+1)/2^m),dx)} = \text{máx}\left\{\left|\frac{2k+1}{2^m} - 2\frac{k+1}{2^m}\right|, \left|\frac{2k+1}{2^m} - 2\frac{k}{2^m}\right|\right\} =$$
$$= \text{máx}\left\{\left|\frac{-1}{2^m}\right|, \left|\frac{1}{2^m}\right|\right\} = 2^{-m}$$

Luego, de lo anterior, como el $k \in \{0, \ldots 2^m - 1\}$ es arbitrario,

$$|R_\rho(\Phi_m)(x) - x^2|_{W^{1,\infty}((0,1),dx)} \leq 2^{-m}$$

De las ecuaciones (3.4) y (3.5) obtenemos que

$$\|R_\rho(\Phi_m)(x) - x^2\|_{W^{1,\infty}((0,1),dx)} = \text{máx}\left\{\|R_\rho(\Phi_m)(x) - x^2\|_{L^\infty((0,1),dx)}, |R_\rho(\Phi_m)(x) - x^2|_{W^{1,\infty}((0,1),dx)}\right\}$$
$$\leq \text{máx}\left\{2^{-m}, 2^{-2-2m}\right\} \leq 2^{-m}$$

Claramente la derivada débil de $R_\rho(\Phi_m)$ es una función constante a trozos que alcanza su máximo en el intervalo $\left[\frac{2^m-1}{2^m}, 1\right]$.Por lo que,

$$|R_\rho(\Phi_m)(x)|_{W^{1,\infty}((0,1),dx)} \leq \frac{((2^m-1)+1)^2}{2^m} - \frac{(2^m-1)^2}{2^m} =$$
$$= \frac{(2^{2m}) - 2^{2m} + 2 \cdot 2^m - 1}{2^m} = \frac{2^{m+1}}{2^m} - \frac{1}{2^m} = 2 - \frac{1}{2^m} \leq 2. \tag{3.6}$$

Sea $\varepsilon \in (0, \frac{1}{2})$, y sea $m = \lceil \log_2(1/\varepsilon) \rceil$.
Ahora, $\Phi_\varepsilon^{sq} \coloneqq \Phi_m$ satisface

$$\|R_\rho(\Phi_\epsilon^{sq})(x) - x^2\|_{W^{1,\infty}((0,1)^d,dx)} \leq 2^{-\lceil \log_2(1/\varepsilon) \rceil} \leq \varepsilon$$

pues

$$\left\lceil \log_2\left(\frac{1}{\varepsilon}\right) \right\rceil \geq \log_2\left(\frac{1}{\varepsilon}\right) \Rightarrow -\left\lceil \log_2\left(\frac{1}{\varepsilon}\right) \right\rceil \leq -\log_2\left(\frac{1}{\varepsilon}\right) \Rightarrow$$
$$\Rightarrow -\log(2)\left\lceil \log_2\left(\frac{1}{\varepsilon}\right) \right\rceil \leq -\log(2)\log_2\left(\frac{1}{\varepsilon}\right) = -\log\left(\frac{1}{\varepsilon}\right) = \log(\varepsilon) \Rightarrow$$
$$\Rightarrow 2^{-m} = \exp\left(-\log(2)\left\lceil \log_2\left(\frac{1}{\varepsilon}\right) \right\rceil\right) \leq \exp(\log(\varepsilon)) = \varepsilon$$

usando que la función exponencial es creciente. Por lo que, Φ_ε^{sq} satisface la ecuación (3.1). Y además, $R_\rho(\Phi_\varepsilon^{sq})(0) = R_\rho(\Phi_m)(0) = 0$.

Consideramos $c_4 = 2$, de la ecuación (3.6) se tiene que

$$|R_\rho(\Phi_\epsilon^{sq})(x)|_{W^{1,\infty}((0,1),dx)} = |R_\rho(\Phi_m)(x)|_{W^{1,\infty}((0,1),dx)} \leq 2 = c_4.$$

Por lo que, Φ_ε^{sq} verifica la estimación (3.2).

Como

$$0 < \varepsilon < \frac{1}{2} \Rightarrow \frac{1}{\varepsilon} > 2 \Rightarrow \log_2\left(\frac{1}{\varepsilon}\right) > \log_2(2) = 1,$$

pues la función logaritmo es creciente, el número de pesos puede ser acotado por

$$M(\Phi_\varepsilon^{sq}) = M(\Phi_m) \leq c_1 m \leq c_1\left(\log_2\left(\frac{1}{\varepsilon}\right) + 1\right) \leq 2c_1 \log_2\left(\frac{1}{\varepsilon}\right),$$

el número de capas puede ser acotado por

$$L(\Phi_\varepsilon^{sq}) = L(\Phi_m) \leq c_2 m \leq c_2\left(\log_2\left(\frac{1}{\varepsilon}\right) + 1\right) \leq 2c_2 \log_2\left(\frac{1}{\varepsilon}\right),$$

y el número de neuronas puede ser acotado por

$$N(\Phi_\varepsilon^{sq}) = N(\Phi_m) \leq c_3 m \leq c_3\left(\log_2\left(\frac{1}{\varepsilon}\right) + 1\right) \leq 2c_3 \log_2\left(\frac{1}{\varepsilon}\right).$$

Renombrando respectivamente $2c_1$, $2c_2$, $2c_3$ por c_1, c_2, c_3 se obtiene el resultado. □

En la siguiente Proposición demostramos que la suma de las funciones realizadas por dos redes neuronales también puede realizarse por una red neuronal.

Proposición 3.1.3. *Sean Φ_1 y Φ_2 dos redes neuronales con función de activación σ con las mismas dimensiones de entrada y de salida. Entonces existe una red neuronal Φ con función de activación σ tal que $R_\sigma(\Phi)(x) = R_\sigma(\Phi_1)(x) + R_\sigma(\Phi_2)(x)$. La cantidad de capas, neuronas y pesos de Φ es menor o igual a la suma de las cantidades de capas, neuronas y pesos de Φ_1 y Φ_2.*

Demostración. Para simplificar, suponemos que Φ_1 y Φ_2 tienen ambas 4 capas,

$$\Phi_1 = ((A_1, a_1), (A_2, a_2), (A_3, a_3), (A_4, a_4))$$
$$\Phi_2 = ((B_1, b_1), (B_2, b_2), (B_3, b_3), (B_4, b_4)).$$

Entonces basta tomar

$$\Phi = ((C_1, c_1), (C_2, c_2), (C_3, c_3), (C_4, c_4))$$

con

$$C_1 = \begin{bmatrix} A_1 \\ B_1 \end{bmatrix} \qquad c_1 = \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}$$
$$C_2 = \begin{bmatrix} A_{2,x_0} & A_{2,x_1} & 0 \\ B_{2,x_0} & 0 & B_{2,x_1} \end{bmatrix} \qquad c_2 = \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}$$
$$C_3 = \begin{bmatrix} A_{3,x_0} & A_{3,x_1} & 0 & A_{3,x_2} & 0 \\ B_{3,x_0} & 0 & B_{3,x_1} & 0 & B_{3,x_2} \end{bmatrix} \qquad c_3 = \begin{bmatrix} a_3 \\ b_3 \end{bmatrix}$$
$$C_4 = \begin{bmatrix} A_{4,x_0} + B_{4,x_0} | A_{4,x_1} | B_{4,x_1} | A_{4,x_2} | B_{4,x_2} | A_{4,x_3} | B_{4,x_3} \end{bmatrix} \qquad c_4 = a_4 + b_4.$$

Sea $x \in \mathbb{R}^d$.
Para Φ_1 se tiene,

$$\begin{aligned}
x_0 &= x \\
x_1 &= \sigma\left(A_1 x_0 + a_1\right) = \sigma\left(A_1 x + a_1\right), \\
x_2 &= \sigma\left(A_{2,x_0} x_0 + A_{2,x_1} x_1 + a_2\right), \\
x_3 &= \sigma\left(A_{3,x_0} x_0 + A_{3,x_1} x_1 + A_{3,x_2} x_2 + a_3\right), \\
x_4 &= A_{4,x_0} x_0 + A_{4,x_1} x_1 + A_{4,x_2} x_2 + A_{4,x_3} x_3 + a_4.
\end{aligned}$$

Para Φ_2 se tiene,

$$\begin{aligned}
y_0 &= x \\
y_1 &= \sigma\left(B_1 y_0 + b_1\right) = \sigma\left(B_1 x + b_1\right), \\
y_2 &= \sigma\left(B_{2,x_0} y_0 + B_{2,x_1} y_1 + b_2\right), \\
y_3 &= \sigma\left(B_{3,x_0} y_0 + B_{3,x_1} y_1 + B_{3,x_2} y_2 + b_3\right), \\
y_4 &= B_{4,x_0} y_0 + B_{4,x_1} y_1 + B_{4,x_2} y_2 + B_{4,x_3} y_3 + b_4.
\end{aligned}$$

Para Φ se tiene,

$$\begin{aligned}
z_0 &= x \\
z_1 &= \sigma\left(C_1 z_0 + c_1\right) = \begin{bmatrix} \sigma\left(A_1 x + a_1\right) \\ \rho\left(B_1 x + b_1\right) \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}, \\
z_2 &= \left(C_{2,x_0} z_0 + C_{2,x_1} z_1 + c_2\right) \\
&= \sigma\left(\begin{bmatrix} A_{2,x_0} \\ B_{2,x_0} \end{bmatrix} x + \begin{bmatrix} A_{2,x_1} & 0 \\ 0 & B_{2,x_1} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} a_2 \\ b_2 \end{bmatrix}\right) \\
&= \begin{bmatrix} \sigma\left(A_{2,x_0} x + A_{2,x_1} x_1 + a_2\right) \\ \sigma\left(B_{2,x_0} x + B_{2,x_1} y_1 + b_2\right) \end{bmatrix} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix}, \\
z_3 &= \sigma\left(C_{3,x_0} z_0 + C_{3,x_1} z_1 + C_{3,x_2} z_2 + c_3\right) \\
&= \sigma\left(\begin{bmatrix} A_{3,x_0} \\ B_{3,x_0} \end{bmatrix} x + \begin{bmatrix} A_{3,x_1} & 0 \\ 0 & B_{3,x_1} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} A_{3,x_2} & 0 \\ 0 & B_{3,x_2} \end{bmatrix} \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} + \begin{bmatrix} a_3 \\ b_3 \end{bmatrix}\right) \\
&= \begin{bmatrix} \sigma\left(A_{3,x_0} x + A_{3,x_1} x_1 + A_{3,x_2} x_2 + a_3\right) \\ \sigma\left(B_{3,x_0} x + B_{3,x_1} y_1 + B_{3,x_2} y_2 + b_3\right) \end{bmatrix}, \\
z_4 &= C_{4,x_0} z_0 + C_{4,x_1} z_1 + C_{4,x_2} z_2 + C_{4,x_3} z_3 + c_4 \\
&= \left(A_{4,x_0} + B_{4,x_0}\right) x + A_{4,x_1} x_1 + B_{4,x_1} y_1 + A_{4,x_2} x_2 + B_{4,x_2} y_2 + \\
&\qquad A_{4,x_3} x_3 + B_{4,x_3} y_3 + a_4 + b_4 \\
&= A_{4,x_0} x + A_{4,x_1} x_1 + A_{4,x_2} x_2 + A_{4,x_3} x_3 + a_4 + B_{4,x_0} x + \\
&\qquad B_{4,x_1} y_1 + B_{4,x_2} y_2 + B_{4,x_3} y_3 + b_4 \\
&= x_4 + y_4.
\end{aligned}$$

Así resulta,

$$\begin{aligned}
R_\sigma(\Phi_1)(x) &= x_4 = A_{4,x_0} x_0 + A_{4,x_1} x_1 + A_{4,x_2} x_2 + A_{4,x_3} x_3 + a_4, \\
R_\sigma(\Phi_2)(x) &= y_4 = B_{4,x_0} y_0 + B_{4,x_1} y_1 + B_{4,x_2} y_2 + B_{4,x_3} y_3 + b_4, \\
R_\sigma(\Phi)(x) &= z_4 = x_4 + y_4
\end{aligned}$$

Por lo que, de lo anterior obtenemos

$$R_\sigma(\Phi)(x) = x_4 + y_4 = R_\sigma(\Phi_1)(x) + R_\sigma(\Phi_2)(x).$$

Por lo tanto, como $x \in \mathbb{R}^d$ es arbitrario, resulta

$$R_\rho(\sigma)(x) = R_\sigma(\Phi_1)(x) + R_\sigma(\Phi_2)(x), \quad \forall x \in \mathbb{R}^d.$$

Además, Φ tiene

$$L(\Phi) = 4 = \max(L(\Phi_1), L(\Phi_2))$$

capas,

$$\begin{aligned} M(\Phi) &= \sum_{j=1}^{4} (\|C_j\|_{\ell^0} + \|c_j\|_{\ell^0}) \\ &= \sum_{j=1}^{4} (\|A_j\|_{\ell^0} + \|B_j\|_{\ell^0} + \|a_j\|_{\ell^0} + \|b_j\|_{\ell^0}) = M(\Phi_1) + M(\Phi_2) \end{aligned}$$

pesos no nulos, y a lo sumo

$$\begin{aligned} N(\Phi) &= d + N_1 + N_2 + N_3 + N_4 = \\ &= d + N_1^1 + N_2^1 + N_3^1 + N_4^1 + N_1^2 + N_2^2 + N_3^2 + N_4^2 \leq N(\Phi_1) + N(\Phi_2). \end{aligned}$$

neuronas. □

En la siguiente proposición vamos a definir una red tal que la función realización aproxima la función multiplicación donde el error es medido en la norma $W^{1,\infty}$. Para esto, usamos la proposición anterior y la identidad de polarización.

$$xy = \frac{1}{2}((x+y)^2 - x^2 - y^2), \quad \text{para} \quad x, y \in \mathbb{R}. \tag{3.7}$$

Proposición 3.1.4. *Para cualquier $M \geq 1$ existen constantes $c, c_1 > 0$ y $c_2 = c_2(M) > 0$ tal que $\forall \varepsilon \in (0, \frac{1}{2})$ existe una red neuronal, $\tilde{\times}$ con dimensión 2 de entrada y una dimensión de salida que satisface las siguientes propiedades:*

(I) $\left\|R_\rho(\tilde{\times})(x,y) - xy\right\|_{W^{1,\infty}((-M,M)^2, dxdy)} \leq \varepsilon$;

(II) si $x = 0$ o $y = 0$, entonces $R_\rho(\tilde{\times})(x,y) = 0$;

(III) $\left|R_\rho(\tilde{\times})(x,y)\right|_{W^{1,\infty}((-M,M)^2, dxdy)} \leq cM$;

(IV) el número de capas, de pesos y neuronas en $\tilde{\times}$ es a lo sumo $c_1 \log_2\left(\frac{1}{\varepsilon}\right) + c_2$.

Demostración. Sea $\varepsilon \in (0, \frac{1}{2})$, y $\delta := \varepsilon/(6M^2C)$ donde C es la constante del Corolario A.2.1 para $d = 2$ y $m = 1$ y sea Φ_δ^{sq} la red neuronal de la Proposición 3.1.2 tal que

$$\left\|R_\rho(\Phi_\delta^{sq})(x) - x^2\right\|_{W^{1,\infty}((0,1),dx)} \leq \delta. \tag{3.8}$$

Definimos las siguientes funciones

$$u_{x,y} : (-M, M)^2 \to (0,1), (x,y) \mapsto \frac{|x+y|}{2M},$$
$$u_x : (-M, M)^2 \to (0,1), (x,y) \mapsto \frac{|x|}{2M},$$
$$u_y : (-M, M)^2 \to (0,1), (x,y) \mapsto \frac{|y|}{2M}.$$

Usando el hecho que $|x| = \rho(x) + \rho(-x)$ podemos construir tres redes neuronales $\Phi_{u_x}, \Phi_{u_y}, \Phi_{u_{x,y}}$ de dimensión de entrada 2 y dimensión de salida 1 tal que la aplicación realización sea $R_\rho(\Phi_{u_x})(x,y) = \frac{|x|}{2M}, R_\rho(\Phi_{u_y})(x,y) = \frac{|y|}{2M}, \text{y} R_\rho(\Phi_{u_{x,y}})(x,y) = \frac{|x+y|}{2M}$ para todo $(x,y) \in (-M, M)^2$.

En efecto definimos $\Phi_{u_x} \coloneqq ((A_1^{u_x}, b_1^{u_x}), (A_2^{u_x}, b_2^{u_x}))$ tal que

$$A_1^{u_x} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, b_1^{u_x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$A_2^{u_x} = \begin{bmatrix} 0 & 0 & \frac{1}{2M} & \frac{1}{2M} \end{bmatrix}, b_2^{u_x} = 0,$$

y obtenemos

$$x_0 = \begin{bmatrix} x \\ y \end{bmatrix},$$
$$x_1 = \rho(A_1^{u_x} x_0 + b_1^{u_x}) = \rho\left(\begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} \rho(x) \\ \rho(-x) \end{bmatrix},$$
$$x_2 = A_2^{u_x} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} + b_2^{u_x} = \begin{bmatrix} 0 & 0 & \frac{1}{2M} & \frac{1}{2M} \end{bmatrix} \begin{bmatrix} x \\ y \\ \rho(x) \\ \rho(-x) \end{bmatrix} + 0 = \frac{1}{2M}\rho(x) + \frac{1}{2M}\rho(-x) =$$
$$= \frac{1}{2M}(\rho(x) + \rho(-x)) = \frac{|x|}{2M}.$$

Por lo que, $R_\rho(\Phi_{u_x}))(x,y) = x_2 = \frac{|x|}{2M}$ y así obtenemos lo que queríamos.

Definiendo $\Phi_{u_y} \coloneqq ((A_1^{u_y}, b_1^{u_y}), (A_2^{u_y}, b_2^{u_y}))$ tal que

$$A_1^{u_y} = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix}, b_1^{u_y} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$A_2^{u_y} = \begin{bmatrix} 0 & 0 & \frac{1}{2M} & \frac{1}{2M} \end{bmatrix}, b_2^{u_y} = 0,$$

y obtenemos

$$x_0 = \begin{bmatrix} x \\ y \end{bmatrix},$$
$$x_1 = \rho(A_1^{u_y} x_0 + b_1^{u_y}) = \rho\left(\begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} \rho(y) \\ \rho(-y) \end{bmatrix},$$
$$x_2 = A_2^{u_y} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} + b_2^{u_y} = \begin{bmatrix} 0 & 0 & \frac{1}{2M} & \frac{1}{2M} \end{bmatrix} \begin{bmatrix} x \\ y \\ \rho(y) \\ \rho(-y) \end{bmatrix} + 0 = \frac{1}{2M}\rho(y) + \frac{1}{2M}\rho(-y) = \frac{|y|}{2M}.$$

Por lo que, $R_\rho(\Phi_{u_y})(x,y) = x_2 = \frac{|y|}{2M}$.

Definiendo $\Phi_{u_{xy}} = ((A_1^{u_y}, b_1^{u_{sy}}), (A_2^{u_y}, b_2^{u_{sy}}))$ tal que

$$A_1^{u_{xy}} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, b_1^{u_{xy}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

$$A_2^{u_{xy}} = \begin{bmatrix} 0 & 0 & \frac{1}{2M} & \frac{1}{2M} \end{bmatrix}, b_2^{u_{xy}} = 0,$$

y obtenemos

$$\begin{aligned}
x_0 &= \begin{bmatrix} x \\ y \end{bmatrix}, \\
x_1 &= \rho\left(A_1^{u_{xy}} x_0 + b_1^{u_{xy}}\right) = \rho\left(\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix}\right) = \begin{bmatrix} \rho(x+y) \\ \rho(-(x+y)) \end{bmatrix}, \\
x_2 &= A_2^{u_{xy}} \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} + b_2^{u_{xy}} = \rho(\begin{bmatrix} 0 & 0 & \frac{1}{2M} & \frac{1}{2M} \end{bmatrix} \begin{bmatrix} x \\ y \\ \rho(x+y) \\ \rho(-(x+y)) \end{bmatrix} + 0 = \\
&= \frac{1}{2M}\rho(x+y) + \frac{1}{2M}\rho(-(x+y)) = \frac{|x+y|}{2M}.
\end{aligned}$$

Por lo que, $R_\rho(\Phi_{u_{xy}})(x,y) = x_2 = \frac{|x+y|}{2M}$.

Primero formamos las redes $\Phi_\delta^{sq} \odot \Phi_{u_x}$, $\Phi_\delta^{sq} \odot \Phi_{u_y}$, $\Phi_\delta^{sq} \odot \Phi_{u_{xy}}$, que verifican $R_\rho(\Phi_\delta^{sq} \odot \Phi_{u_x})(x,y) = R_\rho(\Phi_\delta^{sq})(\frac{|x|}{2M})$, $R_\rho(\Phi_\delta^{sq} \odot \Phi_{u_y})(x,y) = R_\rho(\Phi_\delta^{sq})(\frac{|y|}{2M})$, y $R_\rho(\Phi_\delta^{sq} \odot \Phi_{u_{xy}})(x,y) = R_\rho(\Phi_\delta^{sq})(\frac{|x+y|}{2M})$. Y con esto, usando la Proposición 3.1.3 dos veces, se puede construir la red neuronal $\tilde{\times}$ con dimensión de entrada 2 y dimensión de salida 1, de manera que

$$R_\rho(\tilde{\times})(x,y) = 2M^2\left(R_\rho(\Phi_\delta^{sq})\left(\frac{|x+y|}{2M}\right) - R_\rho(\Phi_\delta^{sq})\left(\frac{|x|}{2M}\right) - R_\rho(\Phi_\delta^{sq})\left(\frac{|y|}{2M}\right)\right).$$

Por la Proposición 3.1.2, existe una constante $c_0 > 0$ (tomando c_0 el máximo de las constantes c_1, c_2, c_3 de la Proposición 3.1.2) tal que $\tilde{\times}$ tiene a lo sumo $c_0 \log_2(1/\delta) + 3 = c_0 \log_2(1/\varepsilon) + c_0 \log_2(6CM^2) + 3 \leq c_0 \log_2(1/\varepsilon) + c_1$ capas, $3c_0 \log_2(1/\delta) + 9 = 3c_0 \log_2(1/\varepsilon) + 3c_0 \log_2(6CM^2) + 9 \leq c' \log_2(1/\varepsilon) + c_2$ neuronas y $3c_0 \log_2(1/\delta) + 17 = 3c_0 \log_2(1/\varepsilon) + 3c_0 \log_2(6CM^2) + 17 \leq c'' \log_2(1/\varepsilon) + c_3$ pesos no nulos. Aquí $c', c'' > 0$ $c_1 = c_1(M), c_2 = c_2(M), c_3 = c_3(M) > 0$ son constantes. Luego si renombramos a c_1, $c_2 = c_2(M)$ como máx(c, c') , máx(c_1, c_2, c_3), respectivamente se tiene (IV).

Como $R_\rho(\Phi_\delta^{sq})(0) = 0$,

$$R_\rho(\tilde{\times})(0,y) = 2M^2\left(R_\rho(\Phi_\delta^{sq})\left(\frac{|y|}{2M}\right) - R_\rho(\Phi_\delta^{sq})\left(\frac{|0|}{2M}\right) - R_\rho(\Phi_\delta^{sq})\left(\frac{|y|}{2M}\right)\right) = 2M^2 0 = 0,$$

$$R_\rho(\tilde{\times})(x,0) = 2M^2\left(R_\rho(\Phi_\delta^{sq})\left(\frac{|x|}{2M}\right) - R_\rho(\Phi_\delta^{sq})\left(\frac{|x|}{2M}\right) - R_\rho(\Phi_\delta^{sq})\left(\frac{|0|}{2M}\right)\right) = 2M^2 0 = 0.$$

Así se verifica (II). Usando la identidad de polarización, de la ecuación (3.7)

$$\begin{aligned} xy &= 4M^2 \frac{x}{2M}\frac{y}{2M} \\ &= 4M^2 \frac{1}{2}\left(\left(\frac{x}{2M}+\frac{y}{2M}\right)^2 - \left(\frac{x}{2M}\right)^2 - \left(\frac{y}{2M}\right)^2\right) \\ &= 2M^2\left(\left(\frac{|x+y|}{2M}\right)^2 - \left(\frac{|x|}{2M}\right)^2 - \left(\frac{|y|}{2M}\right)^2\right). \end{aligned}$$

Sea $f:(0,1)\to\mathbb{R}, x\mapsto x^2$ y usando lo anterior y las definiciones de las funciones u_x, u_y, u_{xy} obtenemos que

$$xy = 2M^2\left(f\circ u_{xy}(x,y) - f\circ u_x(x,y) - f\circ u_y(x,y)\right).$$

Ahora podemos estimar

$$\begin{aligned} &||R_\rho(\tilde{\times})(x,y) - xy||_{W^{1,\infty}((-M,M)^2,dxdy)} = \\ &\quad = ||2M^2(R_\rho(\Phi_\delta^{sq})\circ u_{xy} - R_\rho(\Phi_\delta^{sq})\circ u_x - R_\rho(\Phi_\delta^{sq})\circ u_y) \\ &\quad\quad - 2M^2(f\circ u_{xy} - f\circ u_x - f\circ u_y)||_{W^{1,\infty}((-M,M)^2,dxdy)} \\ &\quad\quad \le 2M^2(||(R_\rho(\Phi_\delta^{sq}) - f)\circ u_{xy}||_{W^{1,\infty}((-M,M)^2,dxdy)} \\ &\quad\quad\quad + ||(R_\rho(\Phi_\delta^{sq}) - f)\circ u_x||_{W^{1,\infty}((-M,M)^2,dxdy)} \\ &\quad\quad\quad\quad + ||(R_\rho(\Phi_\delta^{sq}) - f)\circ u_y)||_{W^{1,\infty}((-M,M)^2,dxdy)}) \end{aligned} \tag{3.9}$$

Podemos ver que

$$|u_{xy}|_{W^{1,\infty}((-M,M)^2)} = |u_x|_{W^{1,\infty}((-M,M)^2)} = |u_y|_{W^{1,\infty}((-M,M)^2)} = \frac{1}{2M}, \tag{3.10}$$

pues

$$\frac{\partial}{\partial x}u_{xy}(x,y) = \frac{\partial}{\partial x}\left(\frac{|x+y|}{2M}\right) = \begin{cases} \frac{1}{2M} & \text{si} \quad x \ge -y \\ -\frac{1}{2M} & \text{si} \quad x < -y \end{cases}$$

Luego, $\left|\frac{\partial u_{xy}}{\partial x}\right| = \frac{1}{2M}$. De manera análoga, se puede ver que $\left|\frac{\partial u_{xy}}{\partial y}\right| = \frac{1}{2M}$. Y así obtenemos que,

$$|u_{xy}|_{W^{1,\infty}((-M,M)^2)} = \text{máx}\left\{\left|\frac{\partial u_{xy}}{\partial x}\right|, \left|\frac{\partial u_{xy}}{\partial y}\right|\right\} = \frac{1}{2M}.$$

De la misma forma podemos ver que $|u_x|_{W^{1,\infty}((-M,M)^2)} = |u_y|_{W^{1,\infty}((-M,M)^2)} = \frac{1}{2M}$.

Por lo que, usando el Corolario A.2.1 con $\Omega_1 = (-M,M)^2 \subset \mathbb{R}^2, \Omega_2 = (0,1)$ son abiertos acotados convexos, las funciones $u_x, u_y, u_{xy}, R_\rho(\Phi_\delta^{sq}) - f$ son Lipschitz continuas, y obtenemos

$$\begin{aligned} &||R_\rho(\tilde{\times})(x,y) - xy||_{W^{1,\infty}((-M,M)^2,dxdy)} \\ &\quad \le 2M^2 3C\,\text{máx}\left\{||R_\rho(\Phi_\delta^{sq}) - f||_{L^\infty((0,1))}, |R_\rho(\Phi_\delta^{sq}) - f|_{W^{1,\infty}((0,1))}\cdot 1/2M\right\} \\ &\quad \le 6M^2C||R_\rho(\Phi_\delta^{sq}) - f||_{W^{1,\infty}((0,1))} \le 6M^2C\frac{\varepsilon}{6M^2C} = \varepsilon, \end{aligned}$$

usando la ecuación (3.8) y la definición de δ, para la tercer desigualdad.

Por lo tanto, vale (I).

Para ver (III), usamos la otra estimación dada en el Corolario A.2.1 y obtenemos

$$\begin{aligned}|R_\rho(\tilde{\times})|_{W^{1,\infty}((-M,M)^2)} &\\ &\leq 2M^2C \sum_{z\in\{x,y,xy\}} |R_\rho(\Phi_\delta^{sq})|_{W^{1,\infty}((0,1))}|u_z|_{W^{1,\infty}((-M,M)^2)}\\ &\leq 2M^2C3C'\frac{1}{2M} = 2CC'M,\end{aligned}$$

usando la ecuación (3.9), y C' la constante de la ecuación (3.2) de la Proposición 3.1.2. □

3.1.2. Algunos lemas técnicos

Ahora construiremos una partición de la unidad que puede definirse como producto de funciones lineales por partes, tal que cada factor del producto es realizado por una red neuronal. Usaremos esta construcción en el lema como una generalización de la Proposición 3.1.2.

Lema 3.1.1. *Para cualquier $d, N \in \mathbb{N}$, existe una colección de funciones*

$$\Psi = \Psi(d,N) = \{\phi_m : m \in \{0,\dots,N\}^d\}$$

con $\phi_m : \mathbb{R}^d \to \mathbb{R}$ para todo $m \in \{0,\dots,N\}^d$ y con las siguientes propiedades:

(I) $0 \leq \phi_m(x) \leq 1$ *para todo* $\phi_m \in \Psi$ *y todo* $x \in \mathbb{R}^d$*;*

(II) $\sum_{\phi_m\in\Psi} \phi_m(x) = 1$ *para todo* $x \in [0,1]^d$*;*

(III) $\operatorname{supp}\phi_m \subset B_{\frac{1}{N},\|\cdot\|_{\ell^\infty}}(\frac{m}{N})$ *para todo* $\phi_m \in \Psi$*;*

(IV) Existe una constante $c \geq 1$ tal que $\|\phi_m\|_{W^{k,\infty}(\mathbb{R}^d)} \leq (cN)^k$ para $k \in \{0,1\}$;

(V) Existen constantes absolutas $C, c \geq 1$ tales que para cada $\phi_m \in \Psi$, existe una red neuronal Φ_m con dimensión de entrada d y dimensión de salida d, con a lo sumo tres capas, Cd pesos no nulos y Cd^2 neuronas, que satisface

$$\prod_{l=1}^{d}[R_\rho(\Phi_m)]_l = \phi_m$$

y $\|[R_\rho(\Phi_m)]_l\|_{W^{k,\infty}(\mathbb{R}^d)} \leq (cN)^k$ para todo $l = 1,\dots,d$ y $k \in \{0,1\}$.

Demostración. Definimos la función

$$\psi : \mathbb{R} \to \mathbb{R}, \psi(x) = \begin{cases} 1 & \text{si} \quad |x| < 1 \\ 0 & \text{si} \quad 2 < |x| \\ 2 - |x| & \text{si} \quad 1 \leq |x| \leq 2 \end{cases}$$

y $\phi_m : \mathbb{R}^d \to \mathbb{R}$ como

$$\phi_m(x) := \prod_{l=1}^{d} \psi\left(3N\left(x_l - \frac{m_l}{N}\right)\right)$$

para $m = (m_1, \dots, m_d) \in \{0, \dots, N\}^d$. Entonces, como $0 \leq \psi(x) \leq 1 \ \forall x \in \mathbb{R}$ se satisface (I) .

Probemos (III).

Sea $m \in \{0, \dots, N\}^d$.

$$x \in \operatorname{supp} \phi_m \Rightarrow \phi_m(x) \neq 0 \Rightarrow \psi\left(3N\left(x_l - \frac{m_l}{N}\right)\right) \neq 0, \forall l = 1, \dots, d \Rightarrow$$
$$\Rightarrow \left|3N\left(x_l - \frac{m_l}{N}\right)\right| < 2, \forall l = 1, \dots, d \Rightarrow \left|x_l - \frac{m_l}{N}\right| < \frac{2}{3N}, \forall l = 1, \dots, d \Rightarrow$$
$$\Rightarrow \left\|x - \frac{m}{N}\right\|_{\ell^\infty} < \frac{2}{3N} < \frac{1}{N} \Rightarrow x \in B_{\frac{1}{N}, \|.\|_{\ell^\infty}}\left(\frac{m}{N}\right).$$

Por lo que vale (III).

Veamos que vale (IV).

Si $k = 0$ tenemos que ver que $\|\phi_m\|_{L^\infty(\mathbb{R}^d)} = \|\phi_m\|_{W^{0,\infty}(\mathbb{R}^d)} \leq (cN)^0 = 1$.

Como por (I), $0 \leq \phi_m(x) \leq 1$ para todo $\phi_m \in \Psi$ y todo $x \in \mathbb{R}^d$, se tiene que $0 \leq \sup_{x \in \mathbb{R}^d} \phi_m(x) = \sup_{x \in \mathbb{R}^d} |\phi_m(x)| \leq 1$ para todo $\phi_m \in \Psi$, por lo que $\|\phi_m\|_{L^\infty(\mathbb{R}^d)} \leq 1 = (cN)^0$ para todo $\phi_m \in \Psi$ y esto vale para cualquier constante c.

Si $k = 1$ queremos ver que existe una constante $c \geq 1$ tal que $\|\phi_m\|_{W^{1,\infty}(\mathbb{R}^d)} \leq (cN)^1$.

Veamos primero que $|\phi_m|_{W^{1,\infty}(\mathbb{R}^d)} \leq 3N$.

Sea $l \in \{1, \dots, d\}$ y $x \in \mathbb{R}^d$. Entonces

$$\left|\frac{\partial}{\partial x_l}\phi_m(x)\right| = \left|\prod_{i=1, i \neq l}^{d} \psi\left(3N\left(x_i - \frac{m_i}{N}\right)\right)\right| \left|\psi'\left(3N\left(x_l - \frac{m_l}{N}\right)\right)\right| 3N$$

Como $|\psi'\left(3N\left(x_l - \frac{m_l}{N}\right)\right)| \leq 1$, para casi todo $x \in \mathbb{R}^d$. Además, se tiene que $\left|\prod_{i=1, i \neq l}^{d} \psi\left(3N\left(x_i - \frac{m_i}{N}\right)\right)\right| \leq 1$, $\forall x \in \mathbb{R}^d$, pues $0 \leq \psi(x) \leq 1 \ \forall x \in \mathbb{R}$. Así se tiene que,

$$\left|\frac{\partial}{\partial x_l}\phi_m\right| \leq 3N, \forall x \in \mathbb{R}^d \Rightarrow \left\|\frac{\partial}{\partial x_l}\phi_m\right\|_{L^\infty(\mathbb{R}^d)} \leq 3N.$$

Luego,

$$|\phi_m|_{W^{1,\infty}(\mathbb{R}^d)} = \max_{l=1,\dots,d}\left\{\left\|\frac{\partial}{\partial x_l}\phi_m(x)\right\|\right\} \leq 3N.$$

Con esto, si consideramos $c = 4 \geq 1$ y usando que $N \geq 1$ tenemos que,

$$\|\phi_m\|_{W^{1,\infty}(\mathbb{R}^d)} = \|\phi_m\|_{L^\infty(\mathbb{R}^d)} + |\phi_m|_{W^{1,\infty}(\mathbb{R}^d)} \leq 1 + 3N \leq 4N = cN = (cN)^1$$

Por lo que obtenemos (IV).

Falta ver (V).

Para esto construimos una red Φ_ψ que realice la función ψ. Sea $\Phi_\psi \coloneqq ((A_1, b_1), (A_2, b_2))$ donde

$$A_1 \coloneqq \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, b_1 \coloneqq \begin{bmatrix} 2 \\ 1 \\ -1 \\ -2 \end{bmatrix}, A_2 \coloneqq \begin{bmatrix} 0 | & 1 & -1 & -1 & 1 \end{bmatrix}, b_2 \coloneqq 0.$$

Entonces Φ_ψ es una red neuronal de dos capas con dimensión de entrada 1 y dimensión de salida 1 con $4 + 4 + 4 = 12$ pesos no nulos $N_0 + N_1 + N_2 = 1 + 4 + 1 = 6$ neuronas tal que

$$R_\rho(\Phi_\psi)(x) = x_2 = A_2 \begin{bmatrix} x \\ \rho(A_1 x + b_1) \end{bmatrix} + b_2 = \begin{bmatrix} 0 | & 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ \rho(x+2) \\ \rho(x+1) \\ \rho(x-1) \\ \rho(x-2) \end{bmatrix} + 0 =$$

$$= \rho(x+2) - \rho(x+1) - \rho(x-1) + \rho(x-2).$$

Se puede ver que $\rho(x+2) - \rho(x+1) - \rho(x-1) + \rho(x-2) = \psi(x)$. Y así resulta $R_\rho(\Phi_\psi)(x) = \psi(x)$.

Para $l = 1, \ldots, d$ queremos construir una red que denotaremos $\Phi_{m,l}$ tal que $R_\rho(\Phi_{m,l})(x) = 3N\left(x_l - \frac{m_l}{N}\right)$.

Para esto ponemos

$$A_1^l \coloneqq \begin{bmatrix} 0 & \ldots & 0 & 3N & 0 & \ldots & 0 \end{bmatrix}, b_1^l \coloneqq -3m_l,$$

A_1 es una matriz $1 \times d$ no nula y con $3N$ en la posición l y 0 en las restantes. Definimos $\Phi_{m,l} \coloneqq ((A_1^l, b_1^l))$. Entonces $\Phi_{m,l}$ es una red neuronal de una capa con dimensión de entrada d y dimensión de salida 1 con $1 + 1 = 2$ pesos no nulos $N_0 + N_1 = d + 1$ neuronas tal que

$$R_\rho(\Phi_{m,l})(x) = x_1 = A_1^l x + b_1^l = \begin{bmatrix} 0 & \ldots & 0 & 3N & 0 & \ldots & 0 \end{bmatrix} x - 3m_l =$$
$$= 3Nx_l - 3m_l = 3N\left(x_l - \frac{m_l}{N}\right)$$

para todo $x \in \mathbb{R}^d$ y $m \in \{0, \ldots, N\}^d$.

Ahora definimos

$$\Phi_m \coloneqq P(\Phi_\psi \odot \Phi_{m,l} : l = 1, \ldots, d)$$

que tiene $L(\Phi_m) = \max_{l=1,\ldots,d}\{L(\Phi_\psi \odot \Phi_{m,l})\} = \max_{l=1,\ldots,d}\{2+1\} = 3$ capas, con dimensión de entrada d y dimensión de salida d, con a lo sumo

$$M(\Phi_m) = \sum_{l=1}^{d} M(\Phi_\psi \odot \Phi_{m,l}) \leq \sum_{l=1}^{d} 2M(\Phi_\psi) + 2M(\Phi_{m,l}) = \sum_{l=1}^{d} 2 \cdot 12 + 2 \cdot 2 = 28d$$

pesos no nulos y a lo sumo

$$N(\Phi_m) = \sum_{l=1}^{d} N(\Phi_\psi \odot \Phi_{m,l}) - (d-1)d \leq \sum_{l=1}^{d} (2N(\Phi_\psi) + 2N(\Phi_{m,l})) - (d-1)d =$$
$$= (\sum_{l=1}^{d} 2 \cdot 6 + 2 \cdot (d+1)) - d^2 + d = 14d + 2d^2 - d^2 + d = d^2 + 14d \leq 15d^2$$

neuronas. Por propiedad de Paralización y Concatenación Dispersa,

$$[R_\rho(\Phi_m)(x)]_l = R_\rho(\Phi_\psi \odot \Phi_{m,l})(x) = R_\rho(\Phi_\psi)(R_\rho(\Phi_{m,l})(x)) = \psi\left(3N\left(x_l - \frac{m_l}{N}\right)\right)$$

para todo $x \in \mathbb{R}^d$ y $l = 1, \ldots, d$. Con esto se puede ver por la definición de ϕ_m que

$$\prod_{l=1}^{d} [R_\rho(\Phi_m)]_l = \prod_{l=1}^{d} \psi\left(3N\left(x_l - \frac{m_l}{N}\right)\right) = \phi_m$$

Usando un razonamiento similar a como demostramos (IV) se puede probar que $\|[R_\rho(\Phi_m)]_l\|_{W^{k,\infty}} \leq (cN)^k$ para todo $l = 1, \ldots, d$ y $k \in \{0, 1\}$. Por lo tanto, vale (V).

Para probar (II), dado que $\sum_{\phi_m \in \Psi} \phi_m$ es una función continua (por ser suma de las ϕ_m que son continuas, por ser producto de funciones continuas) y d-lineal en cada d-rectángulo con vértices $\frac{m}{N}$ vecinos, con $m \in \{0, \ldots, N\}^d$, basta probar que $\sum_{\phi_m \in \Psi} \phi_m(\frac{m'}{N}) = 1$ para cada $m' \in \{0, \ldots, N\}^d$.

Para esto, veamos que $\phi_m(\frac{m'}{N}) = 1$ si $m' = m$ y $\phi_m(\frac{m'}{N}) = 0$ si $m' \neq m$ donde $m' \in \{0, \ldots, N\}^d$. Sea $m' = (m'_1, m'_2, \ldots, m'_d) \in \{0, \ldots, N\}^d$

Luego,

$$\phi_m\left(\frac{m'}{N}\right) = \prod_{l=1}^{d} \psi\left(3N\left(\frac{m'_l}{N} - \frac{m_l}{N}\right)\right) = \prod_{l=1}^{d} \psi\left(3\left(m'_l - m_l\right)\right)$$

.

Si $m' = m$, entonces $m'_l = m_l, \forall l = 1, \ldots, d$, y obtenemos $\phi_m(\frac{m'}{N}) = \prod_{l=1}^{d} \psi(0) = \prod_{l=1}^{d} 1 = 1$

Y si $m' \neq m$, entonces $m'_l \neq m_l$ para algún l, entonces $|m'_l - m_l| \geq 1$, así $3|m'_l - m_l| \geq 3 > 2$ y por definición de ψ, $\psi(3(m'_l - m_l)) = 0$. Por lo que resulta $\phi_m\left(\frac{m'}{N}\right) = 0$.

Por lo tanto,

$$\sum_{\phi_m \in \Psi} \phi_m\left(\frac{m'}{N}\right) = \phi_{m'}\left(\frac{m'}{N}\right) + \sum_{\phi_m \in \Psi, m \neq m'} \phi_m\left(\frac{m'}{N}\right) = 1 + 0 = 1$$

como se quería demostrar. □

El siguiente lema usa la partición de la unidad del lema 3.1.1 y el lema de Bram-Hilbert (A.2.4) en la forma clásica, para derivar una aproximación con polinomios localizados en la norma L^p y $W^{1,\infty}$. Usando un argumento de interpolación, podemos generalizar este resultado para el caso donde la aproximación es respecto de la norma $W^{s,p}$ donde $0 \leq s \leq 1$.Por lo que, en este resultado no estamos usando redes neuronales.

Lema 3.1.2. *Sea $d, N \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty$, y $\Psi = \Psi(d, N) = \{\phi_m : m \in \{0, \ldots, N\}^d\}$ la partición de la unidad del Lema 18. Entonces existe una constante $C = C(d, n, p) > 0$ tal que $f \in W^{n,p}((0,1)^d)$, existe polinomios $p_{f,m}(x) = \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} x^\alpha$ para $m \in \{0, \ldots, N\}^d$ con las siguientes propiedades:*

Sea $0 \leq s \leq 1$ y sea $f_N \coloneqq \sum_{m \in \{0,\ldots,N\}^d} \phi_m p_{f,m}$ entonces el operador $T_s : W^{n,p}((0,1)^d) \to W^{s,p}((0,1)^d)$ con $T_s f = f - f_N$ es lineal y acotada con

$$\|T_s f\|_{W^{s,p}((0,1)^d)} \leq C\left(\frac{1}{N}\right)^{n-s} \|f\|_{W^{n,p}((0,1)^d)}.$$

Más aún, existe una constante $c = c(d, n) > 0$ tal que $\forall f \in W^{n,p}((0,1)^d)$ los coeficientes de los polinomios $p_{f,m}$ satisfacen

$$|c_{f,m,\alpha}| \leq c\|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})} N^{d/p}$$

$\forall \alpha \in \mathbb{N}_0^d$ con $|\alpha| \leq n - 1$ y $m \in \{0, \ldots, N\}^d$, donde $\Omega_{m,N} \coloneqq B_{\frac{1}{N}, \|.\|_{\ell^\infty}}(\frac{m}{N})$ y $\tilde{f} \in W^{n,p}(\mathbb{R}^d)$ es una extensión de f.

Demostración. Usaremos propiedades de aproximación de polinomios promediados de Taylor para obtener una estimación de derivada local y entonces, combinando esta con el uso de la partición de la unidad, para obtener una estimación global. Como estrategia, definiremos un operador extensión. Para esto, sea $E : W^{n,p}((0,1)^d) \to W^{n,p}(\mathbb{R}^d)$ el operador extensión enunciado en el Teorema A.2.1 y sea $\tilde{f} \coloneqq E(f)$. Notamos que para $\Omega \subset \mathbb{R}^d$ arbitrario y $1 \leq k \leq n$ se tiene

$$|\tilde{f}|_{W^{k,p}(\Omega)} \leq \|\tilde{f}\|_{W^{n,p}(\mathbb{R}^d)} \leq C_E ||f||_{W^{n,p}((0,1)^d)} \tag{3.11}$$

donde $C_E = C_E(n,p,d)$ es la norma del operador extensión.

Paso 1:Polinomio de Taylor promediado Para cada $m \in \{0,\dots,N\}^d$, definimos

$$\Omega_{m,N} \coloneqq B_{\frac{1}{N},\|\cdot\|_\ell^\infty}\left(\frac{m}{N}\right), B_{m,N} \coloneqq B_{\frac{3}{4N},|\cdot|}\left(\frac{m}{N}\right)$$

y denotemos $p_m = p_{m,f}$ el polinomio de Taylor de orden n de $\tilde{f}$ promediado sobre $B_{m,N}$ (Definición A.2.6) pues $\tilde{f} \in W^{n-1,p}(\Omega)$, $\overline{B_{m,N}} \subset \Omega_{m,N}$, y $x_0 = \frac{m}{N} \in \Omega_{m,N}$. Del Lema A.2.3 con $\Omega = \Omega_{m,N}, B = B_{m,N}$, $x_0 = \frac{m}{N} \in \Omega_{m,N}$, $R = 2 \geq 1$ y $r = \frac{3}{4N} > 0$.

Veamos que se verifican las hipótesis de la Definición y del Lema.

Veamos que $\overline{B_{m,N}} \subset \Omega_{m,N}$, y que $B_{m,N} \subset B_{2,\|\cdot\|_\ell^\infty}(0)$.

$$y \in \overline{B_{m,N}} \Rightarrow \left|y - \frac{m}{N}\right| \leq \frac{3}{4N} < \frac{1}{N} \Rightarrow \left|y_l - \frac{m_l}{N}\right| \leq \left|y - \frac{m}{N}\right| < \frac{1}{N}, \forall l = 1,\dots,d \Rightarrow$$
$$\Rightarrow \left\|y - \frac{m}{N}\right\|_{\ell^\infty} = \max_{l=1,\dots,d}\left\{\left|y_l - \frac{m_l}{N}\right|\right\} < \frac{1}{N} \Rightarrow y \in \Omega_{m,N}.$$

Por lo que, $\overline{B_{m,N}} \subset \Omega_{m,N}$.

$$y \in B_{m,N} \Rightarrow \left|y - \frac{m}{N}\right| < \frac{3}{4N} \Rightarrow \left|y_l - \frac{m_l}{N}\right| \leq \left|y - \frac{m}{N}\right| < \frac{3}{4N}, \forall l = 1,\dots,d \Rightarrow$$
$$\Rightarrow |y_l| - \left|\frac{m_l}{N}\right| \leq \left|y_l - \frac{m_l}{N}\right| < \frac{3}{4N}, \forall l = 1,\dots,d \Rightarrow |y_l| < \frac{3}{4N} + \left|\frac{m_l}{N}\right| < 2, \forall l = 1,\dots,d$$
$$\Rightarrow \|y\|_{\ell^\infty} = \max_{l=1,\dots,d}\{|y_l|\} < 2 \Rightarrow y \in B_{2,\|\cdot\|_\ell^\infty}(0).$$

Por lo que, $B_{m,N} \subset B_{2,\|\cdot\|_\ell^\infty}(0)$. Luego, del lema se tiene que $p_m(x) = \sum_{|\alpha|\leq n-1} c_{f,m,\alpha} x^\alpha$ para $m \in \{0,\dots,N\}^d$ y existe una constante $c' = c'(n,d) > 0$ tal que

$$|c_{f,m,\alpha}| \leq c'\|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})}\left(\frac{3}{4N}\right)^{-d/p} \leq c'\|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})}\left(\frac{4N}{3}\right)^{d/p} =$$
$$= c'\left(\frac{4}{3}\right)^{d/p}\|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})}N^{d/p} = c''\|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})}N^{d/p}$$

donde $c'' = c''(n,d,p) = c'(n,d)\left(\frac{4}{3}\right)^{d/p}$ $\forall \alpha \in \mathbb{N}_0^d$ con $|\alpha| \leq n-1$ y $m \in \{0,\dots,N\}^d$. Con esto probamos el paso 1.

Paso 2 (Estimación local en $\|\cdot\|_{W^{k,p}}$ para $k \in \{0,1\}$): Para chequear las condiciones del Lema de Bram-Hilbert A.2.4 se satisfacen ya vimos que $\overline{B_{m,N}} \subset \Omega_{m,N}$. Más aún, $B_{m,N}$ es una bola en $\Omega_{m,N}$ tal que $B_{m,N}$ es estrellada con respecto a $\Omega_{m,N}$, además tenemos $\operatorname{diam}_{|\cdot|}(\Omega_{m,N}) = \frac{2\sqrt{d}}{N}$, y $r^*_{\text{máx}}(\Omega_{m,N}) = \frac{1}{N}$, pues $B_{m,N}$ es una bola de radio $\frac{3}{4N}$ centrado en $\frac{m}{N}$, y $\Omega_{m,N}$ es un cubo d dimensional donde la longitud de cada lado es $\frac{2}{N}$ centrado en $\frac{m}{N}$. Por lo que,

$$r_{|\cdot|}(B_{m,N}) = \frac{3}{4N} > \frac{1}{2}\frac{1}{N} = \frac{1}{2}r^*_{\text{máx}}(\Omega_{m,N}).$$

Finalmente, tenemos que el parámetro de Chunkiness es

$$\gamma(\Omega_{m,N}) = \operatorname{diam}(\Omega_{m,N}) \cdot \frac{1}{r^*_{\text{máx}}(\Omega_{m,N})} = \frac{2\sqrt{d}}{N} \cdot N = 2\sqrt{d}.$$

Además, $\tilde{f} \in W^{n,p}(\Omega_{m,N})$ y $1 \leq p \leq \infty$.

Aplicando el Lema de Bram-Hilbert A.2.4, con $h = \operatorname{diam}(\Omega_{m,N})$ para cada $m \in \{0, \ldots, N\}^d$ obtenemos la estimación local con $k = 0$

$$\begin{aligned} ||\tilde{f} - p_m||_{L^p(\Omega_{m,N})} = ||\tilde{f} - p_m||_{W^{0,p}(\Omega_{m,N})} \leq C_1 h^{n-0} |\tilde{f}|_{W^{n,p}(\Omega_{m,N})} = \\ = C_1 \left(\frac{2\sqrt{d}}{N}\right)^n |\tilde{f}|_{W^{n,p}(\Omega_{m,N})} \leq C_1 (2\sqrt{d})^n \left(\frac{1}{N}\right)^n \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} \\ = C_2 \left(\frac{1}{N}\right)^n \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}. \quad (3.12) \end{aligned}$$

donde $C_1 = C_1(n, d, \gamma) = C_1(n, d, 2\sqrt{d}) = C_1(n, d) > 0$ es la constante del Lema donde solo depende de n, d, pues γ depende solo de d, y $C_2 = C_2(n, d) = C_1 \cdot (2\sqrt{d})^n$, además usamos que $|\tilde{f}|_{W^{n,p}(\Omega_{m,N})} \leq \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}$.

Con $k = 1$, usando nuevamente $|\tilde{f}|_{W^{n,p}(\Omega_{m,N})} \leq \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}$ que obtenemos la estimación local

$$\begin{aligned} ||\tilde{f} - p_m||_{W^{1,p}(\Omega_{m,N})} \leq \tilde{C}_3 h^{n-1} |\tilde{f}|_{W^{n,p}(\Omega_{m,N})} = \tilde{C}_3 \left(\frac{2\sqrt{d}}{N}\right)^{n-1} |\tilde{f}|_{W^{n,p}(\Omega_{m,N})} \\ \leq \tilde{C}_3 (2\sqrt{d})^{n-1} \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} = C_3 \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}, \quad (3.13) \end{aligned}$$

donde $\tilde{C}_3 = \tilde{C}_3(n, d) > 0$ es la constante del Lema donde solo depende de n, d, pues γ depende solo de d, y $C_3 = C_3(n, d) = \tilde{C}_3 \cdot (2\sqrt{d})^{n-1}$.

El primer paso hacia una estimación global es combinar los pasos anteriores con las funciones cut-off de la partición de la unidad del Lema 3.1.1. Usando el Lema 3.1.1 ítem (IV), para $k = 0$ y ecuación (3.12)

$$\begin{aligned} \|\phi_m(\tilde{f} - p_m)\|_{L^p(\Omega_{m,N})} \leq \|\phi_m\|_{L^\infty(\Omega_{m,N})} \cdot \|\tilde{f} - p_m\|_{L^p(\Omega_{m,N})} \leq \\ \leq 1 \cdot \|\tilde{f} - p_m\|_{L^p(\Omega_{m,N})} \leq C_2 \left(\frac{1}{N}\right)^n \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}. \end{aligned} \quad (3.14)$$

Más aún, usando Lema A.2.2 para $\phi_m \in W^{1,\infty}(\Omega_{m,N})$, $\tilde{f} - p_m \in W^{1,p}(\Omega_{m,N})$, entonces $\phi_m(\tilde{f} - p_m) \in W^{1,\infty}(\Omega_{m,N})$ y existe $C' = C'(n, d) > 0$ tal que

$$\begin{aligned} |\phi_m(\tilde{f} - p_m)|_{W^{1,p}(\Omega_{m,N})} \leq C' |\phi_m|_{W^{1,\infty}(\Omega_{m,N})} \cdot \|\tilde{f} - p_m\|_{L^p(\Omega_{m,N})} \\ + C' \|\phi_m\|_{L^\infty(\Omega_{m,N})} \cdot |\tilde{f} - p_m|_{W^{1,p}(\Omega_{m,N})}. \quad (3.15) \end{aligned}$$

En la ecuación (3.15) usando el Lema 3.1.1 ítem (IV) con constante $\tilde{c} \geq 1$, para $k = 0$, $k = 1$ y las ecuaciones (3.12),(3.13) obtenemos

$$\begin{aligned} C' |\phi_m|_{W^{1,\infty}(\Omega_{m,N})} \cdot \|\tilde{f} - p_m\|_{L^p(\Omega_{m,N})} + C' \|\phi_m\|_{L^\infty(\Omega_{m,N})} \cdot |\tilde{f} - p_m|_{W^{1,p}(\Omega_{m,N})} \leq \\ \leq C' \cdot \tilde{c} N C_2 \left(\frac{1}{N}\right)^n \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} + C' \cdot C_3 \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} = \\ = (C' \tilde{c} C_2 + C' C_3) \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} = C_4 \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}. \end{aligned}$$

donde $C_4 = C_4(n,d,p) = C'\tilde{c}C_2 + C'C_3 > 0$. Y así resulta

$$|\phi_m(\tilde{f} - p_m)|_{W^{1,p}(\Omega_{m,N})} \leq C_4 \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}. \tag{3.16}$$

Ahora usando las ecuaciones (3.14) y (3.16), y que $\frac{1}{N} \leq 1$ se tiene que

$$\begin{aligned}\|\phi_m(\tilde{f} - p_m)\|_{W^{1,p}(\Omega_{m,N})} &= \|\phi_m(\tilde{f} - p_m)\|_{L^p(\Omega_{m,N})} + |\phi_m(\tilde{f} - p_m)|_{W^{1,p}(\Omega_{m,N})} \leq \\ &\leq C_2 \left(\frac{1}{N}\right)^{n} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} + C_4 \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} \leq \\ &\leq (C_2 + C_4) \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} = C_5 \left(\frac{1}{N}\right)^{n-1} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})}\end{aligned} \tag{3.17}$$

donde $C_5 = C_5(n,d,p) = C_2 + C_4 > 0$.

Sea $C_6 = C_6(n,d,p) = \max\{C_4, C_5\} > 0$ de los cotas para las normas $\|\cdot\|_{L^p}$ y $\|\cdot\|_{W^{1,p}}$ se tiene que

$$\|\phi_m(\tilde{f} - p_m)\|_{W^{k,p}(\Omega_{m,N})} \leq C_6 \left(\frac{1}{N}\right)^{n-k} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} \tag{3.18}$$

con $k \in \{0,1\}$.

Paso 3 (Estimación Global en $\|\cdot\|_{W^{k,p}}, k \in \{0,1\}$): Para derivar una estimación global, empezamos notando que por la propiedad (II) del Lema 3.1.1, tenemos

$$\tilde{f}(x) = 1 \cdot \tilde{f}(x) = \sum_{\phi_m \in \Psi} \phi_m(x)\tilde{f}(x), \text{para casi todo} \quad x \in (0,1)^d.$$

Usando que $\tilde{f}$ es la extensión de f

$$\begin{aligned}\left\| f - \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m \right\|_{W^{k,p}((0,1)^d)} &= \\ \left\| \tilde{f} - \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m \right\|_{W^{k,p}((0,1)^d)} &= \left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(\tilde{f} - p_m) \right\|_{W^{k,p}((0,1)^d)} \leq \\ &\leq \sum_{\tilde{m}\in\{0,\dots,N\}^d} \left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})}\end{aligned} \tag{3.19}$$

Donde el último paso se sigue de $(0,1)^d \subset \bigcup_{\tilde{m}\in\{0,\dots,N\}^d} \Omega_{\tilde{m},N}$. Ahora obtenemos para cada $\tilde{m} \in \{0,\dots,N\}^d$:

$$\begin{aligned}\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(\tilde{f} - p_m) \right\|_{W^{k,p}(\Omega_{\tilde{m},N})} &\leq \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{ell^\infty}\leq 1}} \|\phi_m(f - p_m)\|_{W^{k,p}(\Omega_{\tilde{m},N})} \\ &\leq \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{ell^\infty}\leq 1}} \|\phi_m(f - p_m)\|_{W^{k,p}(\Omega_{m,N})},\end{aligned} \tag{3.20}$$

por desigualdad triangular y pues $\phi_m(x) \neq 0$ si $x \in \operatorname{supp}(\phi_m)$. Luego, para que no se anule, $x \in \operatorname{supp}(\phi_m) \cap \Omega_{\tilde{m},N} \subset \Omega_{m,N} \cap \Omega_{\tilde{m},N} \subset \Omega_{m,N}$. En particular, tomamos $x = \frac{\tilde{m}}{N} \in \Omega_{m,N} \cap \Omega_{\tilde{m},N} \Rightarrow \|m - \tilde{m}\|_{\ell^\infty} \leq 1$.

Luego, de las acotaciones (3.18) y (3.20) , se tiene que

$$\sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{\ell^\infty}\leq 1}} \|\phi_m(f-p_m)\|_{W^{k,p}(\Omega_{m,N})} \leq C_6 \left(\frac{1}{N}\right)^{n-k} \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{\ell^\infty}\leq 1}} \|\tilde{f}\|_{W^{n,p}(\Omega_{m,N})} \quad (3.21)$$

Finalmente, la afirmación de acotación se sigue del uso de la definición de f_N y combinando con las ecuaciones (3.19) y (3.21)

$$\|f - f_N\|^p_{W^{k,p}((0,1)^d)} \leq \sum_{\tilde{m}\in\{0,\dots,N\}^d} C_6^p \left(\frac{1}{N}\right)^{(n-k)p} \times \left(\sum_{\substack{m\in\{0,\dots,N\}^d \\ |m-\tilde{m}|_\infty\leq 1}} \|f\|_{W^{n,p}(\Omega_{m,N})} \right)^p$$

(Desigualdad de Hölder; $q = 1 - 1/p$)

$$\leq \sum_{\tilde{m}\in\{0,\dots,N\}^d} C_6^p \left(\frac{1}{N}\right)^{(n-k)p} \times \left(\left(\sum_{\substack{m\in\{0,\dots,N\}^d \\ |m-\tilde{m}|_\infty\leq 1}} \|f\|^p_{W^{n,p}(\Omega_{m,N})} \right)^{\frac{1}{p}} \left(\sum_{\substack{m\in\{0,\dots,N\}^d \\ |m-\tilde{m}|_\infty\leq 1}} 1^q \right)^{\frac{1}{q}} \right)^p$$

$$= C_6^p \left(\frac{1}{N}\right)^{(n-k)p} \times \sum_{\tilde{m}\in\{0,\dots,N\}^d} \sum_{\substack{m\in\{0,\dots,N\}^d \\ |m-\tilde{m}|_\infty\leq 1}} \|f\|^p_{W^{n,p}(\Omega_{m,N})} (3^d)^{p/q}$$

$$\leq C_6^p 3^{dp/q} \left(\frac{1}{N}\right)^{(n-k)p} 3^d \sum_{m\in\{0,\dots,N\}^d} \|f\|^p_{W^{n,p}(\Omega_{m,N})}$$

$$\leq C_6^p 3^{dp/q} \left(\frac{1}{N}\right)^{(n-k)p} 3^d 2^d \|f\|^p_{W^{n,p}(\bigcup_{\tilde{m}\in\{0,\dots,N\}^d} \Omega_{m,N})}.$$

$$= C_6^p 3^{(dp+1)/q} 2^d \left(\frac{1}{N}\right)^{(n-k)p} \|f\|^p_{W^{n,p}(\bigcup_{\tilde{m}\in\{0,\dots,N\}^d} \Omega_{m,N})}.$$

Donde los últimos dos pasos se siguen de la definición de $\Omega_{\tilde{m},N}$. Por lo tanto, tenemos:

$$\|f - f_N\|_{W^{k,p}((0,1)^d)} \leq C_7 \left(\frac{1}{N}\right)^{n-k} \|f\|_{W^{n,p}(\mathbb{R}^d)}$$

$$\leq C_7 C_E \left(\frac{1}{N}\right)^{n-k} \|f\|_{W^{n,p}((0,1)^d)} = C_8 \left(\frac{1}{N}\right)^{n-k} \|f\|_{W^{n,p}((0,1)^d)}$$

para $k \in \{0, 1\}$, donde la Ecuación (3.11) se usó en el primer y segundo paso. Aquí $C_7 = C_7(n, d, p) = C_6^p 3^{(dp+1)/q} 2^d > 0$ y $C_8 = C_8(n, d, p) = C_7 C_E$ son constantes. La

linealidad de T_k, $k \in \{0,1\}$ es una consecuencia de la linealidad del polinomio de Taylor promediado.

Paso 4: Interpolación:

Para $0 < s < 1$ usamos un argumento de interpolación de espacios de Banach. Sea

$$A_0 = A_1 = W^{n,p}((0,1)^d)$$

con

$$B_0 = L^p((0,1)^d), B_2 = W^{n,p}((0,1)^d).$$

$$T : A_0 \to B_0, \|T\|_{B_0} \leq C\left(\frac{1}{N}\right)^n$$

,

$$T : A_1 \to B_1, \|T\|_{B_1} \leq C\left(\frac{1}{N}\right)^{n-1},$$

entonces, tomando el operador T con $Tf \coloneqq f - f_N$, como $T \in \mathcal{L}(A_0, B_0) \cap \mathcal{L}(A_1, B_1)$ por el Teorema A.1.1 obtenemos una constante $c = c(n,d,p) > 0$ tal que

$$\|T\|_{\mathcal{L}(W^{n,p}((0,1)^n),W^{s,p}((0,1)^d))} \leq \|T\|^{1-s}_{\mathcal{L}(W^{n,p}((0,1)^d),L^p((0,1)^n))} \cdot \|T\|^{s}_{\mathcal{L}(W^{n,p}((0,1)^d),W^{1,p}((0,1)^d))}$$
$$\leq \left(C_8\left(\frac{1}{N}\right)^n\right)^{1-s}\left(C_8\left(\frac{1}{N}\right)^{n-1}\right)^s = C_8\left(\frac{1}{N}\right)^{n-ns+ns-s} = C_8\left(\frac{1}{N}\right)^{n-s}$$

□

El siguiente lema técnico cae en la formulación para la aproximación de monomios localizados con una red neuronal . Usando la notación y estamento (IV) del Lema 3.1.1, un monomio $\phi_m(x)x^\alpha$ puede expresarse como el producto de componentes de salida de una red $\Phi_{(m,\alpha)}$ con dimensión de salida n, i.e.

$$\phi_m(x)x^\alpha = \prod_{i=1}^{n}[R(\Phi_{m,\alpha})]_l(x)$$

Dada una red Φ con dimensión de salida n construimos una red Ψ_Φ que aproxime el producto de las componentes de la salida de Φ.

Lema 3.1.3. *Sean $d, m, k \in \mathbb{N}$ y $N \geq 1$ arbitrarios. Entonces existe una constante $c = c(m) > 0$ tal que lo siguiente se mantiene:*

Para cada $\varepsilon \in \left(0, \frac{1}{2}\right)$ y cualquier red neuronal Φ con dimensión de entrada d y dimensión de salida n donde $n \leq m$, y con número de capas, neuronas y pesos acotados por K, tal que

$$\|[R_\rho(\Phi)]_l\|_{W^{k,\infty}((0,1)^d)} \leq N^k \quad para \quad k \in \{0,1\}, y \quad l = 1, \ldots, n. \tag{3.22}$$

existe una red neuronal $\Psi_{\varepsilon,\Phi}$ con dimensión de entrada d y dimensión de salida 1, con número de capas, neuronas, pesos acotados por $KC\log_2(\frac{1}{\varepsilon})$ talque

$$\left\| R_\rho(\Psi_{\varepsilon,\Phi}) - \prod_{i=1}^{n}[R_\rho(\Phi)]_l \right\|_{W^{k,\infty}((0,1)^d)} \leq cN^k\varepsilon \tag{3.23}$$

para $k \in \{0,1\}$ y alguna constante $c = c(d,m,k) > 0$. Más aún,

$$R_\rho(\Psi_{\varepsilon,\Phi})(x) = 0 \quad si \quad \prod_{i=1}^{n}[R_\rho(\Phi)(x)]_l = 0, \forall x \in (0,1)^d. \tag{3.24}$$

Demostración. Sea $d, m, k \in \mathbb{N}$ y $N \geq 1$. Haremos la prueba por inducción sobre $m \in \mathbb{N}$.

Adicionalmente mostraremos que la constante $c = c(d, m, k) = m^{1-k}c_1^k$ donde $c_1 = c_1(d, m) > 0$ y que la red $\Psi_{\varepsilon,\Phi}$ puede ser elegida tal que las primeras $L(\Phi) - 1$ capas de $\Psi_{\varepsilon,\Phi}$ y Φ coincidan y $|R_\rho(\Psi_{\varepsilon,\Phi})| \leq C_1 N$ para alguna constante $C_1 = C_1(d.m) > 0$.

Si $m = 1$, elegimos $\Psi_{\varepsilon,\Phi} = \Phi \ \forall \varepsilon \in (0, \frac{1}{2})$. Por lo que, vale para el caso base.

Ahora supongamos que vale para m y veamos que vale para $m + 1$.

Para esto, sea $\epsilon \in (0, \frac{1}{2})$ y sea $\Phi = ((A_1, b_1), (A_2, b_2), \ldots, (A_L, b_L))$ es una red neuronal con dimensión de entrada d y dimensión de salida n con $n \leq m + 1$ con número de capas, neuronas y pesos acotados por K, donde A_l es una matriz de dimensiones $N_l \times \sum_{k=0}^{l-1} N_k$, y $b_l \in \mathbb{R}^{N_l}$ para $l = 1, \ldots, L$.

Caso 1: Si $n \leq m$ entonces por hipótesis de inducción resulta que existe una constante $C_0 = C_0(m) > 0$ y una red neuronal $\Psi_{\varepsilon,\Phi}$ con dimensión de entrada d y dimensión de salida 1, con número de capas, neuronas, pesos acotados por $KC_0 \log_2(\frac{1}{\varepsilon})$ tal que

$$\left\| R_\rho(\Psi_{\varepsilon,\Phi}) - \prod_{i=1}^{n}[R_\rho(\Phi)]_l \right\|_{W^{k,\infty}((0,1)^d)} \leq m^{1-k}c_1^k N^k \varepsilon \leq (m+1)^{1-k}c_1^k N^k \varepsilon$$

para $k \in \{0, 1\}$ y alguna constante $c_1 = c_1(d, m) > 0$, usando que $m^0 = 1 = (m+1)^0$ y que $m^1 \leq (m + 1)^1$. Más aún,

$$R_\rho(\Psi_{\varepsilon,\Phi})(x) = 0 \quad \text{si} \quad \prod_{i=1}^{n}[R_\rho(\Phi)(x)]_l = 0, \forall x \in (0,1)^d.$$

Y además, $|R_\rho(\Psi_{\varepsilon,\Phi})|_{W^{k,\infty}((0,1)^d)} \leq C_1 N$ para alguna constante $C_1 = C_1(d.m) > 0$.

Caso 2: Ahora, supongamos que $n = m+1$ y mostramos que existen constantes $\tilde{C}_0, \tilde{c}_1$ y $\tilde{C}_1$ constantes dependiendo de $m + 1$, posiblemente diferentes de las constantes C_0, c_1, y C_1 del caso 1, respectivamente. El máximo de estas constantes dan el resultado para redes con dimensión de salida $n \leq m + 1$.

Denotamos Φ_m red neuronal con dimensión de entrada d y dimensión de salida m que resulta de Φ removiendo la última neurona de salida y los pesos correspondientes. En detalle, escribimos

$$A_L = \begin{bmatrix} A_L^{(1,m)} \\ a_L^{(m+1)} \end{bmatrix}, \text{y } b_L = \begin{bmatrix} b_L^{(1,m)} \\ b_L^{(m+1)} \end{bmatrix}$$

donde $A_L^{(1,m)}$ es una matriz de dimensiones $m \times \sum_{k=0}^{l-1} N_k$ y $a_L^{(m+1)}$ es un vector $1 \times \sum_{k=0}^{l-1} N_k$, y $b_L^{(1,m)} \in \mathbb{R}^n$, $b_L^{(m+1)} \in \mathbb{R}$.

Ahora definimos

$$\Phi_m = ((A_1, b_1), (A_2, b_2), \ldots, (A_L^{(1,m)}, b_L^{(1,m)}))$$

. Usando la hipótesis de inducción y las constantes C_0, c_1, y C_1 del caso 1, obtenemos que existe una red neuronal $\Psi_{\varepsilon,\Phi_m} = ((A'_1, b'_1), (A'_2, b'_2), \ldots, (A'_{L'}, b'_{L'}))$ con dimensión

de entrada d y dimensión de salida 1, con número de capas, neuronas, pesos acotados por $KC_0 \log_2(\frac{1}{\varepsilon})$ tal que

$$\left\| R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^{m} [R_\rho(\Phi_m)]_l \right\|_{W^{k,\infty}((0,1)^d)} \leq m^{1-k} c_1^k N^k \varepsilon \tag{3.25}$$

para $k \in \{0, 1\}$ y alguna constante $c_1 = c_1(d, m) > 0$. Más aún,

$$R_\rho(\Psi_{\varepsilon,\Phi})(x) = 0 \quad \text{si} \quad \prod_{i=1}^{n} [R_\rho(\Phi)(x)]_l = 0, \forall x \in (0,1)^d. \tag{3.26}$$

Y además, podemos asumir por hipótesis inductiva que $|R_\rho(\Psi_{\varepsilon,\Phi})| \leq C_1 N$ para alguna constante $C_1 = C_1(d.m) > 0$ y que las primeras $L(\Phi_m) - 1 = L(\Phi) - 1$ capas de Φ_m y $\Psi_{\varepsilon,\Phi_m}$, i.e., $A_l = A'_l, \forall l = 1, \ldots, L(\Phi) - 1$.

Ahora definimos la siguiente red,

$$\tilde{\Psi}_{\varepsilon,\Phi} \coloneqq \left((A'_1, b'_1), (A'_2, b'_2), \ldots, (A'_{L'-1}, b'_{L'-1}), \left(\begin{bmatrix} A'_{L'} \\ a_L^{(m+1)} 0_{\mathbb{R}^{1, \sum_{k=L}^{L'} N'_k}} \end{bmatrix}, \begin{bmatrix} b'_{L'} \\ b_L^{(m+1)} \end{bmatrix} \right) \right)$$

se sigue manteniendo que las primeras $L - 1$ capas de $\tilde{\Psi}_{\varepsilon,\Phi}$ y Φ coinciden y $\tilde{\Psi}_{\varepsilon,\Phi}$ es una red neuronal con dimensión de salida 2.

Notar que usando la ecuación (3.25), y de la ecuación (3.22) que tenemos por hipótesis del Lema 3.1.3, y que $\varepsilon < 1$ obtenemos

$$\begin{aligned} \left\| \left[R_\rho(\tilde{\Psi}_{\varepsilon,\Phi}) \right]_1 \right\|_{L^\infty((0,1)^d)} &= \|R_\rho(\Psi_{\varepsilon,\Phi_m})\|_{L^\infty((0,1)^d)} \leq \\ &\leq \left\| R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^{m} [R_\rho(\Phi_m)]_l \right\|_{L^\infty((0,1)^d)} + \left\| \prod_{i=1}^{m} [R_\rho(\Phi_m)]_l \right\|_{L^\infty((0,1)^d)} \\ &\leq m\varepsilon + 1 < m + 1 \end{aligned} \tag{3.27}$$

Además, usando la ecuación (3.22) que tenemos por hipótesis del Lema 3.1.3, resulta

$$\left\| \left[R_\rho(\tilde{\Psi}_{\varepsilon,\Phi}) \right]_2 \right\|_{L^\infty((0,1)^d)} = \left\| [R_\rho(\Phi_m)]_{m+1} \right\|_{L^\infty((0,1)^d)} \leq 1 < m + 1 \tag{3.28}$$

Ahora notamos $\tilde{\times}$ es la red de la Proposición 3.1.4 con $M \coloneqq m + 1$ y el mismo ε. Y definimos,

$$\Psi_{\varepsilon,\Phi} \coloneqq \tilde{\times} \odot \tilde{\Psi}_{\varepsilon,\Phi}$$

Consecuentemente $\Psi_{\varepsilon,\Phi}$ tiene dimensión de entrada d y dimensión de salida 1 y combinando la hipótesis de inducción con el estamento (IV) de la proposición 3.1.4 con constantes $c' > 0, c'' = c''(m + 1) > 0$, y observación 2.2.1, y la hipótesis del

Lema 3.1.3, usando el hecho de que $K, \log_2\left(\frac{1}{\varepsilon}\right) \geq 1$, tiene a lo sumo

$$\begin{aligned} L(\tilde{\times} \odot \tilde{\Psi}_{\varepsilon,\Phi}) &= L(\tilde{\times}) + L(\tilde{\Psi}_{\varepsilon,\Phi}) \\ &\leq \left(c' \log_2\left(\frac{1}{\varepsilon}\right) + c''\right) + KC_0 \log_2\left(\frac{1}{\varepsilon}\right) \\ &\leq K \log_2\left(\frac{1}{\varepsilon}\right) 2\left(c' + c'' + C_0\right) = \\ &= KC \log_2\left(\frac{1}{\varepsilon}\right) \end{aligned}$$

$$\begin{aligned} N(\tilde{\times} \odot \tilde{\Psi}_{\varepsilon,\Phi}) &\leq N(\tilde{\times}) + 2N(\tilde{\Psi}_{\varepsilon,\Phi}) \\ &\leq 2\left(c' \log_2\left(\frac{1}{\varepsilon}\right) + c''\right) + 2KC_0 \log_2\left(\frac{1}{\varepsilon}\right) \\ &\leq K \log_2\left(\frac{1}{\varepsilon}\right) 2\left(c' + c'' + C_0\right) = \\ &= KC \log_2\left(\frac{1}{\varepsilon}\right) \end{aligned}$$

$$\begin{aligned} M(\tilde{\times} \odot \tilde{\Psi}_{\varepsilon,\Phi}) &\leq 2M(\tilde{\times}) + 2M(\tilde{\Psi}_{\varepsilon,\Phi}) \\ &\leq 2\left(c' \log_2\left(\frac{1}{\varepsilon}\right) + c''\right) + 2KC_0 \log_2\left(\frac{1}{\varepsilon}\right) \\ &\leq K \log_2\left(\frac{1}{\varepsilon}\right) 2\left(c' + c'' + C_0\right) = \\ &= KC \log_2\left(\frac{1}{\varepsilon}\right) \end{aligned}$$

capas, neuronas y pesos respectivamente donde $C = C(m+1) = 2(c' + c'' + C_0) > 0$. Claramente las primeras $L' - 1$ capas de $\Psi_{\varepsilon,\Phi}$ y $\tilde{\Psi}_{\varepsilon,\Phi}$, y en particular las $L - 1$ capas. Por lo cual, las primeras $L - 1$ capas de $\Psi_{\varepsilon,\Phi}$ y Φ coinciden. Por propiedades de aproximación, se tiene que

$$\begin{aligned} &\left| R_\rho(\Psi_{\varepsilon,\Phi}) - \prod_{i=1}^{m+1} [R_\rho(\Phi)]_l \right|_{W^{k,\infty}((0,1)^d)} \\ &= \left| R_\rho(\tilde{\times}) \circ R_\rho(\tilde{\Psi}_{\varepsilon,\Phi}) - [R_\rho(\Phi)]_{m+1} \prod_{i=1}^{m} [R_\rho(\Phi)]_l \right|_{W^{k,\infty}((0,1)^d)} \leq \\ &\leq \left| R_\rho(\tilde{\times}) \left(R_\rho(\Psi_{\varepsilon,\Phi_m}), [R_\rho(\Phi)]_{m+1}\right) - [R_\rho(\Phi)]_{m+1} R_\rho(\Psi_{\varepsilon,\Phi_m}) \right|_{W^{k,\infty}((0,1)^d)} \\ &\quad + \left| [R_\rho(\Phi)]_{m+1} \left(R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^{m} [R_\rho(\Phi)]_l \right) \right|_{W^{k,\infty}((0,1)^d)} \end{aligned} \tag{3.29}$$

para $k \in \{0, 1\}$.

Empecemos con $k = 0$.

Para el primer término de la desigualdad (3.29), como las dos componentes de salida de $\tilde{\Psi}_{\varepsilon,\Phi}$ por $m + 1$ por las ecuaciones (3.27) y (3.28), pues son unos x, y particulares en $(-(m+1), m+1)^2$

$$\left\|R_\rho(\tilde{\times})\left(R_\rho(\Psi_{\varepsilon,\Phi_m}),[R_\rho(\Phi)]_{m+1}\right)-[R_\rho(\Phi)]_{m+1}\,R_\rho(\Psi_{\varepsilon,\Phi_m})\right\|_{L^\infty((0,1)^d)}$$
$$\leq \left\|R_\rho(\tilde{\times})(x,y)-x\cdot y\right\|_{L^\infty((-(m+1),m+1)^2,dxdy)}$$

Por 3.1.4 ítem (II) resulta

$$\left\|R_\rho(\tilde{\times})\left(R_\rho(\Psi_{\varepsilon,\Phi_m}),[R_\rho(\Phi)]_{m+1}\right)-[R_\rho(\Phi)]_{m+1}\,R_\rho(\Psi_{\varepsilon,\Phi_m})\right\|_{L^\infty((0,1)^d)}$$
$$\leq \left\|R_\rho(\tilde{\times})(x,y)-x\cdot y\right\|_{L^\infty((-(m+1),m+1)^2,dxdy)}\leq\varepsilon \quad (3.30)$$

Para el segundo término de la desigualdad (3.29), usamos que $\|f\cdot g\|_{L^\infty((0,1)^d)}\leq \|f\|_{L^\infty((0,1)^d)}\|g\|_{L^\infty((0,1)^d)}$ para cualquier par de funciones $f,g\in L^\infty((0,1)^d)$ y la ecuación 3.25 con $k=0$ que viene de la hipótesis inductiva, y que $\|\,[R_\rho(\Phi)]_{m+1}\,\|_{L^\infty((0,1)^d)}\leq$ $N^0=1$ por hipótesis del Lema 3.1.3,

$$\left\|[R_\rho(\Phi)]_{m+1}\left(R_\rho(\Psi_{\varepsilon,\Phi_m})-\prod_{i=1}^{m}[R_\rho(\Phi)]_l\right)\right\|_{L^\infty((0,1)^d)}\leq$$
$$\leq \left\|[R_\rho(\Phi)]_{m+1}\right\|_{L^\infty((0,1)^d)}\left\|R_\rho(\Psi_{\varepsilon,\Phi_m})-\prod_{i=1}^{m}[R_\rho(\Phi)]_l\right\|_{L^\infty((0,1)^d)}\leq 1\cdot m\cdot\varepsilon \quad (3.31)$$

Luego, combinando la ecuación (3.29) con las ecuaciones (3.30) obtenemos que

$$\left\|R_\rho(\Psi_{\varepsilon,\Phi})-\prod_{i=1}^{m+1}[R_\rho(\Phi)]_l\right\|_{L^\infty((0,1)^d)}\leq \varepsilon+m\cdot\varepsilon=(m+1)\cdot\varepsilon. \quad (3.32)$$

Ahora acotemos los dos términos de (3.29) para $k=1$.

Para el primer término de la desigualdad (3.29), sea $h(x,y)=xy,\forall(x,y)\in(-(m+1),m+1)^2$ $h\in W^{1,\infty}((-(m+1),m+1)^2)$, y así $R_\rho(\tilde{\times})-h\in W^{1,\infty}((-(m+1),m+1)^2)$ pues las dos componentes de salida de $\tilde{\Psi}_{\varepsilon,\Phi}$ por $m+1$ por las ecuaciones (3.27) y (3.28), pues son un x,y particulares en $(-(m+1),m+1)^2$, y además $R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})\in W^{1,\infty}((0,1)^d)$, por Corolario A.2.1, existe $\hat{C}=\hat{C}(d)>0$ tal que

$$\left|R_\rho(\tilde{\times})\left(R_\rho(\Psi_{\varepsilon,\Phi_m}),[R_\rho(\Phi)]_{m+1}\right)-[R_\rho(\Phi)]_{m+1}\cdot R_\rho(\Psi_{\varepsilon,\Phi_m})\right|_{W^{1,\infty}((0,1)^d)}=$$
$$=\left|(R_\rho(\tilde{\times})-h)\circ\left(R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})\right)\right|_{W^{1,\infty}((0,1)^d)}$$
$$\leq \hat{C}|R_\rho(\tilde{\times})-h|_{W^{1,\infty}((-(m+1),m+1)^2)}\cdot|R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})|_{W^{1,\infty}((0,1)^d)}$$
$$=\hat{C}\left|R_\rho(\tilde{\times})(x,y)-x\cdot y\right|_{W^{1,\infty}((-(m+1),m+1)^2,dxdy)}\cdot|R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})|_{W^{1,\infty}((0,1)^d)}$$

Luego, por 3.1.4 ítem (II) para $k=1$, y por estar las componentes de $\tilde{\Psi}_{\varepsilon,\Phi}$ acotadas en la norma $W^{1,\infty}$ por C_1N y N respectivamente, se tiene que

$$\left|R_\rho(\tilde{\times})\left(R_\rho(\Psi_{\varepsilon,\Phi_m}),[R_\rho(\Phi)]_{m+1}\right)-[R_\rho(\Phi)]_{m+1}\cdot R_\rho(\Psi_{\varepsilon,\Phi_m})\right|_{W^{1,\infty}((0,1)^d)}$$
$$\leq \hat{C}\left|R_\rho(\tilde{\times})(x,y)-x\cdot y\right|_{W^{1,\infty}((-(m+1),m+1)^2,dxdy)}\cdot|R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})|_{W^{1,\infty}((0,1)^d)}\leq$$
$$\leq \hat{C}\varepsilon\,\text{máx}\{C_1N,N\}\leq\hat{C}\,\text{máx}\{C_1,1\}N\varepsilon=C_1'N\varepsilon \quad (3.33)$$

Para el segundo término de la desigualdad 3.29, usando Lema A.2.2 con $p = \infty$ para $[R_\rho(\Phi)]_{m+1} \in W^{1,\infty}(\Omega_{m,N})$, $R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l \in W^{1,\infty}(\Omega_{m,N})$, entonces $[R_\rho(\Phi)]_{m+1} \left(R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right) inW^{1,\infty}(\Omega_{m,N})$ y

$$\begin{aligned}
&\left|[R_\rho(\Phi)]_{m+1}\left(R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right)\right|_{W^{1,p}((0,1)^d)} \\
&\leq \left|[R_\rho(\Phi)]_{m+1}\right|_{W^{1,\infty}((0,1)^d)} \cdot \left\|R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right\|_{L^p((0,1)^d)} \\
&+\| [R_\rho(\Phi)]_{m+1} \|_{L^\infty((0,1)^d)} \cdot \left|R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right|_{W^{1,p}((0,1)^d)}.
\end{aligned}$$

Por la ecuación 3.25 con $k = 0, \text{y} \quad k = 1$ que viene de la hipótesis inductiva, y que $\|[R_\rho(\Phi)]_l\|_{W^{k,\infty}((0,1)^d)} \leq N^k \quad \text{para} \quad k \in \{0,1\}$ por hipótesis del Lema 3.1.3, se tiene que

$$\begin{aligned}
&\left|[R_\rho(\Phi)]_{m+1}\left(R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right)\right|_{W^{1,p}((0,1)^d)} \\
&\leq \left|[R_\rho(\Phi)]_{m+1}\right|_{W^{1,\infty}((0,1)^d)} \cdot \left\|R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right\|_{L^p((0,1)^d)} \\
&+\| [R_\rho(\Phi)]_{m+1} \|_{L^\infty((0,1)^d)} \cdot \left|R_\rho(\Psi_{\varepsilon,\Phi_m}) - \prod_{i=1}^m [R_\rho(\Phi)]_l\right|_{W^{1,p}((0,1)^d)} \\
&\leq N \cdot m\varepsilon + 1 \cdot c_1 N\varepsilon = c_1' N\varepsilon, \qquad (3.34)
\end{aligned}$$

donde $c_1' = c_1'(d, m+1) = m + c_1$.

Luego, combinando la ecuación (3.29) con las ecuaciones (3.33) y (3.34) obtenemos que

$$\left|R_\rho(\Psi_{\varepsilon,\Phi}) - \prod_{i=1}^{m+1} [R_\rho(\Phi)]_l\right|_{W^{1,\infty}((0,1)^d)} \leq C_1' N\varepsilon + c_1' N\varepsilon = (C_1' + c_1')N\varepsilon = c_1'' N\varepsilon \quad (3.35)$$

donde $c_1'' = c_1''(d+1, m) = C_1' + c_1' > 0$

Luego, por las ecuaciones (3.32) y (3.35), y usando que $N \geq 1$, se tiene que

$$\left\|R_\rho(\Psi_{\varepsilon,\Phi}) - \prod_{i=1}^{m+1} [R_\rho(\Phi)]_l\right\|_{W^{1,\infty}((0,1)^d)} \leq (m+1)\varepsilon + c_1'' N\varepsilon = (m+1+c_1'')N\varepsilon = c_1''' N\varepsilon. \quad (3.36)$$

donde $c_1''' = c_1'''(d+1, m) = m + 1 + c_1'' > 0$.

Ahora probemos (3.24) para $m+1$. Supongamos que $[R_\rho(\Phi)]_l(x) = 0$ para algún $l \in \{1, \ldots, m+1\}$ y $x \in (0,1)^d$. Si $l \leq m$, entonces por la ecuación (3.26) implica que

$$\left[R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})\right]_1 (x) = R_\rho(\Psi_{\varepsilon,\Phi_m})(x) = 0.$$

Si $l = m + 1$

$$\left[R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})\right]_2 (x) = \left[R_\rho(\Phi)\right]_{m+1} (x) = 0$$

Por lo tanto, por ítem (I) de la Proposición 3.1.4, se tiene que

$$R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})(x) = R_\rho(\tilde{\times}) \left(R_\rho(\Psi_{\varepsilon,\Phi_m})(x), \left[R_\rho(\Phi)\right]_{m+1} (x)\right) = 0.$$

Finalmente, tenemos que mostrar que existe una constante $C_1'' = C_1''(d, m+1) > 0$ tal que

$$|R_\rho(\Psi_{\varepsilon,\Phi})| \leq C_1'' N \tag{3.37}$$

Por lo que como $R_\rho(\tilde{\times}) \in W^{1,\infty}((-(m+1), m+1)^2)$ pues las dos componentes de salida de $\tilde{\Psi}_{\varepsilon,\Phi}$ por $m+1$ por las ecuaciones (3.27) y (3.28), pues son un x, y particulares en $(-(m+1), m+1)^2$, y además $R_\rho(\tilde{\Psi}_{\varepsilon,\Phi}) \in W^{1,\infty}((0,1)^d)$, por Corolario A.2.1, existe $\hat{C} = \hat{C}(d) > 0$ tal que

$$\begin{aligned} |R_\rho(\Psi_{\varepsilon,\Phi})|_{W^{1,\infty}((0,1)^d)} &= \left|R_\rho(\tilde{\times}) \circ R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})\right|_{W^{1,\infty}((0,1)^d)} = \\ &\leq \hat{C} |R_\rho(\tilde{\times})|_{W^{1,\infty}((-(m+1),m+1)^2)} \cdot |R_\rho(\tilde{\Psi}_{\varepsilon,\Phi})|_{W^{1,\infty}((0,1)^d)} \end{aligned}$$

Por la proposición 3.1.4 ítem (III) con constante $\hat{c} > 0$ y con un argumento como en (3.33) se tiene que

$$|R_\rho(\Psi_{\varepsilon,\Phi})|_{W^{1,\infty}((0,1)^d)} \leq \hat{C} \cdot \hat{c} \cdot (m+1) \cdot \max\{C_1 N, N\} = C_1'' N \tag{3.38}$$

donde $C_1'' = C_1''(d, m+1) = \hat{C} \cdot \hat{c} \cdot (m+1) \cdot \max\{C_1, 1\} > 0$.

□

En el siguiente lema, aproximaremos una suma de polinomios localizados con una red neuronal. Una de las dificultades es controlar la derivada de las funciones localizadas desde la partición de la unidad.

Lema 3.1.4. *Sea $d, N \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty$ y sea $\Psi = \Psi(d, N) = \{\phi_m : m \in \{0, \ldots, N\}^d\}$ la partición de la unidad del Lema 3.1.1. Entonces existen constantes $C_1 = C_1(n, d, p, s) > 0$, y $C_2 = C_2(n, d) > 0$, $C_3 = C_3(n, d) > 0$ con las siguientes propiedades:*

Para cada $\varepsilon \in \left(0, \frac{1}{2}\right)$, existe una arquitectura de red neuronal $\mathcal{A}_\varepsilon = \mathcal{A}_\varepsilon(d, n, N, \varepsilon)$ con dimensión de entrada d y dimensión de salida 1 con a lo sumo $C_2 \log_2(\frac{1}{\varepsilon})$ número de capas, y $C_3 \log_2(\frac{1}{\varepsilon})$ pesos y neuronas tal que lo siguiente se mantiene:

Sea $f \in W^{n,p}((0,1)^d)$ y $p_m(x) = p_{f,m}(x) = \sum_{|\alpha| \leq n-1} c_{f,m,\alpha} x^\alpha$ para $m \in \{0, \ldots, N\}^d$ los polinomios del Lema 3.1.2, entones existe una red neuronal $\Phi_{P,\varepsilon}$ que tiene arquitectura $\mathcal{A}_\varepsilon$ tal que

$$\left\| \sum_{m \in \{0,\ldots,N\}^d} \phi_m p_{f,m} - R_\rho(\Phi_{P,\varepsilon}) \right\|_{W^{k,p}((0,1)^d)} \leq C_1 \|f\|_{W^{n,p}((0,1)^d)} N^s \varepsilon \tag{3.39}$$

Demostración. **Paso 1:(Aproximación de monomios localizados $\phi_m(x) x^\alpha$)**

Sea $|\alpha| \leq n-1$ y $m \in \{0,\dots,N\}^d$. Es fácil ver que existe una red neuronal Φ_α con dimensión de entrada d y dimensión de salida $|\alpha|$ con una capa y a lo sumo $n-1$ pesos y $n-1+d$ neuronas.

$$x^\alpha = \prod_{l=1}^{|\alpha|} \left[R_\rho(\Phi_\alpha)\right]_l(x) \quad \text{para casi todo} \quad x \in (0,1)^d$$

y

$$\left\|\left[R_\rho(\Phi_\alpha)\right]_l\right\|_{W^{1,\infty}((0,1)^d)} \leq 1, \forall l = 1,\dots,|\alpha|, \text{y} \quad k \in \{0.,1\}. \tag{3.40}$$

En efecto, sea para $i = 1,\dots,d$ $\Phi^i = (A_1^i, b_1^i)$ con $A_1^i \coloneqq \left[0,\dots,1,\dots,0\right]$ el vector fila de dimensión $1 \times d$ que tiene un uno en la componente i, y $b_1^i = 0$ que es un número en $\mathbb{R}$. Así, $\Phi^i = (A_1^i, b_1^i)$ es una red neuronal con una capa, un peso no nulo y $d+1$ neuronas y es tal que para $x \in (0,1)^d$

$$R_\rho(\Phi^i)(x) = A_1^i x + b_1^i = \left[0,\dots,1,\dots,0\right] \begin{bmatrix} x_1 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{bmatrix} + 0 = x_i,$$

además

$$\left\|R_\rho(\Phi^i)\right\|_{W^{1,\infty}((0,1)^d)} \leq 1, \quad k \in \{0,1\}. \tag{3.41}$$

Pues $\|R_\rho(\Phi^i)\|_{L^\infty((0,1)^d)} = |x_1| = x_1 \leq |x| \leq 1$ y $\|R_\rho(\Phi^i)\|_{W^{1,\infty}((0,1)^d)} = \max\{|x_i|, 1\} \leq 1$ pues $\frac{\partial}{\partial x_j} x_i = \delta_{ji}$ $\forall j = 1,\dots,d$.

Luego, sea

$$\Phi_\alpha \coloneqq P(\underbrace{\Phi^1,\dots,\Phi^1}_{\alpha_1 \text{veces}}, \underbrace{\Phi^2,\dots,\Phi^2}_{\alpha_2 \text{veces}}, \dots, \underbrace{\Phi^d,\dots,\Phi^d}_{\alpha_d \text{veces}}).$$

Entonces Φ_α es una red neuronal con a lo sumo $L(\Phi_\alpha) = \max\{\Phi^1,\dots,\Phi^d\} = 1$ capas, $M(\Phi_\alpha) = \sum_{i=1}^d \alpha_i M(\Phi^i) = |\alpha| \leq n-1$ pesos no nulos, y $N(\Phi_\alpha) = \sum_{i=1}^d \alpha_i N(\Phi^i) - d(|\alpha|-1) = |\alpha| \cdot (d+1) - d \cdot (|\alpha|-1) = |\alpha| + d \leq n-1+d$ neuronas. Y además resulta

$$\prod_{l=1}^{|\alpha|} \left[R_\rho(\Phi_\alpha)\right]_l(x) = \prod_{l=1}^{\alpha_1} \left[R_\rho(\Phi_\alpha)\right]_l(x) \dots \prod_{l=\sum_{i=1}^{d-1}\alpha_i}^{|\alpha|} \left[R_\rho(\Phi_\alpha)\right]_l(x) = x_1^{\alpha_1} \cdot \dots \cdot x_d^{\alpha_d} = x^\alpha$$

para casi todo $x \in (0,1)^d$, y

$$\left\|\left[R_\rho(\Phi_\alpha)\right]_l\right\|_{W^{1,\infty}((0,1)^d)} \leq 1, \forall l = 1,\dots,|\alpha|, \text{y} \quad k \in \{0.,1\}.$$

Sea ahora, Φ_m la red neuronal $C, c \geq 1$ las constantes del ítem (V) del Lema 3.1.1, y definimos

$$\Phi_{m,\alpha} \coloneqq P(\Phi_m, \Phi_\alpha)$$

Entonces $\Phi_{m,\alpha}$ tiene a lo sumo $L(\Phi_{m,\alpha}) = \max\{3,1\} = 3 \leq K_0$ capas, $M(\Phi_{m,\alpha}) = M(\Phi_m) + M(\Phi_\alpha) \leq Cd + n - 1 \leq K_0$ pesos no nulos y $N(\Phi_{m,\alpha}) = N(\Phi_m) +$

$N(\Phi_\alpha) - d \leq Cd + n - 1 + d - d = Cd + n - 1 \leq K_0$ neuronas, para una constante $K_0 = K_0(n.d) \in \mathbb{N}$ y

$$\prod_{l=1}^{|\alpha|+d} \left[R_\rho(\Phi_{m,\alpha})\right]_l(x) = \prod_{l=1}^{d} \left[R_\rho(\Phi_{m,\alpha})\right]_l(x) \prod_{l=d+1}^{|\alpha|+d} \left[R_\rho(\Phi_{m,\alpha})\right]_l(x) = \phi_m(x)x^\alpha.$$

para casi todo $x \in (0,1)^d$. Además,

$$\left\|\left[R_\rho(\Phi_{m,\alpha})\right]_l\right\|_{W^{1,\infty}((0,1)^d)} \leq 1, \forall l = 1, \ldots, |\alpha| + d, \text{y} \quad k \in \{0,1\}.$$

pues $\left\|\left[R_\rho(\Phi_{m,\alpha})\right]_l\right\|_{L^\infty((0,1)^d)} \leq 1, \forall l = 1, \ldots, |\alpha| + d$ y $\left\|\left[R_\rho(\Phi_{m,\alpha})\right]_l\right\|_{W^{1,\infty}((0,1)^d)} \leq CN, \forall l = 1, \ldots, d$, y $\left\|\left[R_\rho(\Phi_{m,\alpha})\right]_l\right\|_{W^{1,\infty}((0,1)^d)} \leq 1 \leq CN, \forall l = d+1, \ldots, |\alpha| + d$.

Para construir una aproximación del monomio localizado $\phi_m(x)x^\alpha$, sea $\Psi_{\varepsilon,(m,\alpha)}$ la red neuronal del lema 3.1.3(con $\Phi_{m,\alpha}$ en lugar de Φ, $m = n-1+d \in \mathbb{N}$, $K = K_0 \in \mathbb{N}$ y CN en lugar de N) para $m \in \{0, \ldots, N\}^d$ y $\alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1$. Entonces existe una constante $C_1 = C_1(n,d) \geq 1$ tal que $\Psi_{\varepsilon,(m,\alpha)}$ tiene a lo sumo $C_1 \log_2(\frac{1}{\varepsilon})$ capas, número de neuronas y pesos, pues K_0 está fijo. Más aún,

$$\begin{aligned}\left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{1,\infty}|((0,1)^d)} &= \\ = \left\|\prod_{l=1}^{|\alpha|+d} \left[R_\rho(\Phi_{m,\alpha})\right]_l(x) - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{1,\infty}((0,1)^d)} &\leq c' N^k \varepsilon. \quad (3.42)\end{aligned}$$

para alguna constante $c' = c'(n,d,k) > 0$ y $k \in \{0,1\}$ y

$$R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x) = 0 \quad \text{si} \quad \phi_m(x)x^\alpha = \prod_{l=1}^{|\alpha|+d} \left[R_\rho(\Phi_{m,\alpha})\right]_l(x) = 0, \forall x \in (0,1)^d. \quad (3.43)$$

Paso 2 (Construimos una arquitectura aceptable de aproximando sumas de polinomios localizados) Ponemos

$$M \coloneqq |\{(m,\alpha) : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1\}|$$

y definimos la matriz $A_{\text{sum}} \in \mathbb{R}^{1\times M}$ por $A_{\text{sum}} \coloneqq [c_{m,\alpha} : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1]$ y red neuronal $\Phi_{\text{sum}} \coloneqq ((A_{\text{sum}}, 0))$ con dimensión de entrada M y dimensión de salida 1, con a lo sumo 1 capa, M pesos no nulos y $M+1$ neuronas.

Finalmente, ponemos

$$\Phi_{P,\varepsilon} \coloneqq \Phi_{\text{sum}} \odot P\left(\Psi_{\varepsilon,(m,\alpha)} : m \in \{0, \ldots, N\}^d, \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1\right)$$

Entonces existen constantes $C_2 = C_2(n,d), C_3 = C_3(n,p) > 0$ tal que $\Phi_{P,\varepsilon}$ con dimensión de entrada d y con dimensión de salida 1, usando el hecho de que $C_1, \log_2\left(\frac{1}{\varepsilon}\right) \geq 1$ y $M \leq (N+1)^d \cdot (n-1)^d$, tiene a lo sumo

$$\begin{aligned}L(\Phi_{P,\varepsilon}) &= L(\Phi_{\text{sum}}) + \max_{m\in\{0,\ldots,N\}^d, \alpha\in\mathbb{N}_0^d, |\alpha|\leq n-1} L(\Psi_{\varepsilon,(m,\alpha)}) \\ &\leq 1 + C_1 \log_2\left(\frac{1}{\varepsilon}\right) \leq (C_1+1)\log_2\left(\frac{1}{\varepsilon}\right) \\ &= C_2 \log_1\left(\frac{1}{\varepsilon}\right),\end{aligned}$$

$$\begin{aligned} M(\Phi_{P,\varepsilon}) &\leq 2M(\Phi_{\text{sum}}) + 2 \sum_{m\in\{0,\dots,N\}^d,\alpha\in\mathbb{N}_0^d,|\alpha|\leq n-1} M(\Psi_{\varepsilon,(m,\alpha)}) \\ &\leq 2M + 2MC_1 \log_2\left(\frac{1}{\varepsilon}\right) \leq 4MC_1 \log_2\left(\frac{1}{\varepsilon}\right) \\ &\leq 4(N+1)^d(n-1)^d C_1 \log_2\left(\frac{1}{\varepsilon}\right) \leq C_3 \log_2\left(\frac{1}{\varepsilon}\right), \end{aligned}$$

$$\begin{aligned} N(\Phi_{P,\varepsilon}) &\leq 2N(\Phi_{\text{sum}}) + 2 \sum_{m\in\{0,\dots,N\}^d,\alpha\in\mathbb{N}_0^d,|\alpha|\leq n-1} N(\Psi_{\varepsilon,(m,\alpha)}) \\ &\leq 2(M+1) + 2MC_1 \log_2\left(\frac{1}{\varepsilon}\right) \leq 4M + 2MC_1 \log_2\left(\frac{1}{\varepsilon}\right) \\ &\leq 6MC_1 \log_2\left(\frac{1}{\varepsilon}\right) \leq 6(N+1)^d(n-1)^d C_1 \log_2\left(\frac{1}{\varepsilon}\right) = C_3 \log_2\left(\frac{1}{\varepsilon}\right), \end{aligned}$$

capas, pesos no nulos y neuronas respectivamente, donde $C_2 = C_2(n,d) = C_1+1 > 0$ y $C_3 = C_3(n,d) = 6(N+1)^d(n-1)^d C_1 \log_2\left(\frac{1}{\varepsilon}\right) > 0$, y

$$R_\rho(\Phi_{P,\varepsilon}) = \sum_{m\in\{0,\dots,N\}^d,\alpha\in\mathbb{N}_0^d,|\alpha|\leq n-1} c_{m,\alpha} R_\rho(\Psi_{\varepsilon,(m,\alpha)}).$$

$\Phi_{P,\varepsilon}$ solo depende de $p_{f,m}$(y así de f) por sus coeficientes $c_{m,\alpha}$. Ahora, es fácil ver que existe una arquitectura $\mathcal{A}_\varepsilon = \mathcal{A}_\varepsilon(d,n,N,\varepsilon)$ con $L(\mathcal{A}_\varepsilon) \leq C_2 \log_2\left(\frac{1}{\varepsilon}\right)$ capas, y número de neuronas y pesos acotados por $C_3 \log_2\left(\frac{1}{\varepsilon}\right)$ tal que $\Phi_{P,\varepsilon}$ tiene arquitectura $\mathcal{A}_\varepsilon$ para cualquier elección de coeficientes $c_{m,\alpha}$ y por lo tanto para cualquier elección de f.

Paso 3(Estimación del error de aproximación en $\|\cdot\|_{W^{k,p}}, k \in \{0,1\}$). Para cada $m \in \{0,\dots,N\}^d$ ponemos,

$$\Omega_{m,N} \coloneqq B_{\frac{1}{N},\|\cdot\|_\infty}\left(\frac{m}{N}\right).$$

Y obtenemos para $k \in \{0,1\}$

$$\begin{aligned} &\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x) \right\|^p_{W^{k,p}((0,1)^d;dx)} = \\ &= \left\| \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\leq n-1} c_{m,\alpha}\left(\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right) \right\|^p_{W^{k,p}((0,1)^d);dx} \leq \\ &\leq \sum_{\tilde{m}\in\{0,\dots,N\}^d} \left\| \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\leq n-1} c_{m,\alpha}\left(\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right) \right\|^p_{W^{k,p}(\Omega_{\tilde{m},N}\cap(0,1)^d;dx)} \end{aligned} \tag{3.44}$$

Donde el último paso se sigue de $(0,1)^d \subset \bigcup_{\tilde{m}\in\{0,\dots,N\}^d} \Omega_{\tilde{m},N}$. Para $\tilde{m} \in \{0,\dots,N\}^d$, tenemos que

$$\left\| \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\le n-1} c_{m,\alpha}\left(\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right)\right\|_{W^{k,p}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} \le$$
$$\le \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\le n-1} |c_{m,\alpha}| \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,p}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} \le$$
$$\le c_1 \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\le n-1} \|\tilde f\|_{W^{n-1,p}(\Omega_{m,N})} N^{d/p} \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,p}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} \tag{3.45}$$

por desigualdad triangular y $c_1 = c_1(n,d) > 0$ la constante de la acotación de las constantes $c_{\alpha,m}$ en el Lema 3.1.2.

Ahora, tenemos que

$$\left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,p}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} \le$$
$$\le \lambda\left(\Omega_{\tilde m,N}\cap(0,1)^d\right)^{\frac{1}{p}} (d+1)^{\frac{1}{p}} \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,\infty}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} \le$$
$$\le \left(\frac{2}{N}\right)^{\frac{d}{p}} (d+1)^{\frac{1}{p}} \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,\infty}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} =$$
$$= c_3\left(\frac{1}{N}\right)^{\frac{d}{p}} \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,\infty}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)}$$

donde λ denota la medida de Lebesgue y usamos que $\lambda\left(\Omega_{\tilde m,N}\cap(0,1)^d\right) \le \left(\frac{2}{N}\right)^d$ $c_3 = c_3(d,p) = (d+1)^{\frac{1}{p}}\cdot 2^{\frac{d}{p}} > 0$ es una constante.

Combinando lo anterior con la ecuación (3.45), usando el Lema 3.1.1 ítem (III) (razonamiento visto en la demostración del lema 3.1.2) y por la ecuación (3.42) , obtenemos

$$\left\| \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\le n-1} c_{m,\alpha}\left(\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right)\right\|_{W^{k,p}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} \le$$
$$\le c_3c_1 \sum_{m\in\{0,\dots,N\}^d} \sum_{|\alpha|\le n-1} \|\tilde f\|_{W^{n-1,p}(\Omega_{m,N})} N^{d/p}\left(\frac{1}{N}\right)^{\frac{d}{p}}$$
$$\times \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,\infty}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)} =$$
$$= c_4 \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde m\|_{\ell^\infty}\le 1}} \sum_{|\alpha|\le n-1} \|\tilde f\|_{W^{n-1,p}(\Omega_{m,N})} N^{d/p}\left(\frac{1}{N}\right)^{\frac{d}{p}}$$
$$\times \left\|\phi_m(x)x^\alpha - R_\rho(\Psi_{\varepsilon,(m,\alpha)})(x)\right\|_{W^{k,\infty}(\Omega_{\tilde m,N}\cap(0,1)^d;dx)}$$
$$\le c_4c' N^k\varepsilon \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde m\|_{\ell^\infty}\le 1}} \sum_{|\alpha|\le n-1} \|\tilde f\|_{W^{n-1,p}(\Omega_{m,N})} \tag{3.46}$$

donde $c_4 = c_4(n,d,p) = c_1c_3 > 0$.

Luego, usando que $|\{\alpha : \alpha \in \mathbb{N}_0^d, |\alpha| \leq n-1\}| \leq (n-1)^d$ y por desigualdad de Hölder con $q = 1 - \frac{1}{p}$,

$$\sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{\ell^\infty}\leq 1}} \sum_{|\alpha|\leq n-1} \|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})} \leq (n-1)^d \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{\ell^\infty}\leq 1}} \|\tilde{f}\|_{W^{n-1,p}(\Omega_{m,N})} \leq$$

$$\leq (n-1)^d 3^{d/q} \left(\sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{\ell^\infty}\leq 1}} \|\tilde{f}\|^p_{W^{n-1,p}(\Omega_{m,N})} \right)^{\frac{1}{p}} \tag{3.47}$$

Juntando las ecuaciones (3.46) y (3.47) obtenemos,

$$\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x) \right\|^p_{W^{k,p}((0,1)^d;dx)} \leq$$

$$\leq (c_4 c' 3^{d/q}(n-1)^d N^k \varepsilon)^p \sum_{\tilde{m}\in\{0,\dots,N\}^d} \sum_{\substack{m\in\{0,\dots,N\}^d \\ \|m-\tilde{m}\|_{\ell^\infty}\leq 1}} \|\tilde{f}\|^p_{W^{n-1,p}(\Omega_{m,N})}$$

$$\leq (c_4 c' 3^{d/q}(n-1)^d)^p N^{kp} \varepsilon^p \|\tilde{f}\|^p_{W^{n-1,p}(\mathbb{R}^d)} \leq$$

$$\leq C_E (c_4 c' 3^{d/q}(n-1)^d)^p N^{kp} \varepsilon^p \|f\|^p_{W^{n,p}((0,1)^d)} = c_5 N^{kp} \varepsilon^p \|f\|^p_{W^{n,p}((0,1)^d)} \tag{3.48}$$

donde hacemos lo mismo que en el Paso 3 de la demostración del Lema 3.1.2 y $c_5 = c_5(n,d,p,k) = C_E(c_4 c' 3^{d/q}(n-1)^d)^p > 0$ es una constante. Por lo tanto, obtenemos

$$\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x) \right\|_{W^{k,p}((0,1)^d;dx)} \leq c_5 N^k \varepsilon \|f\|_{W^{n,p}((0,1)^d)}. \tag{3.49}$$

Con lo cual, queda probado para los casos $s = 0$ y $s = 1$.

Paso 4(Interpolación). Para $0 \leq s \leq 1$ usando el Corolario A.1.1 con $B_0 := L^p((0,1)^d)$, $B_1 := W^{1,p}((0,1)^d)$, y $u(x) = \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x)$, y usando la ecuación (3.49) para $k = 0$ y $k = 1$ se tiene que

$$\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x) \right\|_{W^{s,p}((0,1)^d;dx)} \leq$$

$$\leq C \left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x) \right\|^{1-s}_{L^p((0,1)^d);dx)}$$

$$\times \left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m(x)p_m(x) - R_\rho(\Phi_{P,\varepsilon})(x) \right\|^{s}_{W^{1,p}((0,1)^d);dx)}$$

$$\leq C(c_5(n,p,d,0)N^0\varepsilon\|f\|_{W^{n,p}((0,1)^d)})^{1-s}(c_5(n,p,d,1)N^1\varepsilon\|f\|_{W^{n,p}((0,1)^d)})^s$$

$$= C c_5(n,p,d,0)^{1-s} c_5(n,p,d,1)^s N^s \varepsilon \|f\|_{W^{n,p}((0,1)^d)} =$$

$$= c_6 N^s \varepsilon \|f\|_{W^{n,p}((0,1)^d)}.$$

donde $C = C(s,p) > 0$ constante del Corolario A.1.1, y ponemos $c_6 = c_6(n.p.d.s) = = Cc_5(n,p,d,0)^{1-s}c_5(n,p,d,1)^s > 0$. Esto concluye la prueba del Lema. □

3.1.3. Demostración del Teorema principal

Ahora, estamos en condiciones de probar el Teorema 3.1.1.

Demostración del Teorema 3.1.1. La demostración se puede dividir en dos pasos: Primero, aproximamos la función f mediante una suma de polinomios localizados y luego aproximamos esta suma mediante una red neuronal.

Para el primer paso, definimos

$$N = \left\lceil \left(\frac{\varepsilon}{2CB}\right)^{-1/(n-s)} \right\rceil, \tag{3.50}$$

donde $C = C(n,d,p) > 0$ es la constante del Lema 3.1.2 Sin pérdida de generalidad, podemos asumir que $CB \geq 1$. El mismo lema implica que si $\Psi = \Psi(d,N) = \{\phi_m : m \in \{0,\dots,N\}^d\}$ es la partición de la unidad del Lema 3.1.1, entonces existen polinomios $p_m(x) = \sum_{|\alpha|\leq 1} c_{m,\alpha}x^\alpha$ para $m \in \{0,\dots,N\}^d$ tales que

$$\begin{aligned}\left\| f - \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m \right\|_{W^{s,p}((0,1)^d)} &\leq C\|f\|_{W^{n,p}((0,1)^d)} \left(\frac{1}{N}\right)^{n-s} \leq CB\left(\frac{1}{N}\right)^{n-s} \\ &\leq CB\frac{\varepsilon}{2CB} = \frac{\varepsilon}{2}.\end{aligned} \tag{3.51}$$

usando que $N \geq \left(\frac{\varepsilon}{2CB}\right)^{-1/(n-s)} \Rightarrow N^{n-s} \geq \frac{\varepsilon}{2CB} \Rightarrow \left(\frac{1}{N}\right)^{n-s} \leq \frac{2CB}{\varepsilon}$

Para el segundo paso, sean $C_1 = C_1(n,d,p,s) > 0$, $C_2 = C_2(n,d) > 0$ y $C_3 = C_3(n,d) > 0$ las constantes del Lema 3.1.4 y sea $\Phi_{P,\varepsilon}$ la red neuronal proporcionada por el Lema 3.1.4 con $\varepsilon^{n/(n-s)}/(8CB^2C_1)$ en lugar de ε . Usando el hecho que $\log_2(\varepsilon^{-n/(n-s)}), (2CB)/\varepsilon \geq 1$, entonces $\Phi_{\rho,\varepsilon}$ tiene a lo sumo

$$\begin{aligned}C_2\log_2(8CB^2C_1\varepsilon^{-n/(n-s)}) &= C_2(\log_2(8CB^2C_1) + \log_2(\varepsilon^{-n/(n-s)})) \\ &\leq (C_2(\log_2(8CB^2C_1)+1)\log_2(\varepsilon^{-n/(n-s)})) \\ &= C'\log_2(\varepsilon^{-n/(n-s)})\end{aligned}$$

capas para una constante $C' = C'(n,d,p,B,s) = C_2(\log_2(8CB^2C_1) + 1 > 0$ y a lo sumo

$$\begin{aligned}&C_3\left(\left(\frac{\varepsilon}{2CB}\right)^{-1/(n-s)} + 2\right)^d \log_2(8CB^2C_1\varepsilon^{-n/(n-s)}) \\ &\leq C_3 3^d \left(\frac{\varepsilon}{2CB}\right)^{-d/(n-s)} \log_2(8CB^2C_1\varepsilon^{-n/(n-s)}) \\ &= C_3 3^d (2CB)^{d/(n-s)}\varepsilon^{-d/(n-s)} \\ &= C''\varepsilon^{-d/(n-s)}\log_2(\varepsilon^{-n/(n-s)}),\end{aligned}$$

para una constante $C'' = C''(n,d,p,B,s) = C_3 3^d(2CB)^{d/(n-s)} > 0$ pesos no nulos y neuronas.

$$\begin{aligned}\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m - R_\rho(\Phi_{P,\varepsilon}) \right\|_{W^{s,p}((0,1)^d)} &\leq C_1 \|f\|_{W^{n,p}((0,1)^d)} N^s \frac{\varepsilon^{n/(n-s)}}{8CB^2C_1} \\ &\leq CB\left(\left(\frac{\varepsilon}{2CB}\right)^{-1/(n-s)} + 1\right)^s \frac{\varepsilon^{n/(n-s)}}{8CB^2} \\ &\leq CB\left(\left(\frac{\varepsilon}{2CB}\right)^{-s/(n-s)} + 1\right) \frac{\varepsilon^{n/(n-s)}}{8CB^2},\end{aligned}$$

donde en el último paso usamos la desigualdad $(x+y)^\theta \leq x^\theta + y^\theta$ para $x, y \geq 0$ y $0 \leq \theta \leq 1$. Continuamos lo anterior usando el hecho de que $(2CB)^{s/(n-s)} \leq 2CB$ pues $2CB \geq 1$ y $s/(n-s) \leq 1$, y obtenemos

$$\begin{aligned}CB\left(\left(\frac{\varepsilon}{2CB}\right)^{-s/(n-s)} + 1\right) \frac{\varepsilon^{n/(n-s)}}{8CB^2} &\leq \frac{1}{8CB^2C_1}(2CB\varepsilon + \varepsilon^{n/(n-s)}) \\ &\leq \frac{1}{8CB^2C_1}(2CB\varepsilon + 2CB\varepsilon) = \frac{\varepsilon}{2}\end{aligned}$$

Por lo que, de lo anterior,

$$\left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m - R_\rho(\Phi_{P,\varepsilon}) \right\|_{W^{s,p}((0,1)^d)} \leq \frac{\varepsilon}{2}. \tag{3.52}$$

Por lo tanto, usando desigualdad triangular y las ecuaciones (3.51) y (3.52) obtenemos,

$$\begin{aligned}\|f - R_\rho(\Phi_{P,\varepsilon})\|_{W^{s,p}((0,1)^d)} &\leq \left\| f - \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m \right\|_{W^{s,p}((0,1)^d)} \\ &\quad + \left\| \sum_{m\in\{0,\dots,N\}^d} \phi_m p_m - R_\rho(\Phi_{P,\varepsilon}) \right\|_{W^{s,p}((0,1)^d)} \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,\end{aligned}$$

lo que concluye la prueba. □

3.2. Acotación inferior de complejidad

Ahora enunciaremos un resultado sobre cotas inferiores y mostramos que el mismo compromiso entre regularidad y complejidad, que se puede observar en las cotas superiores, también se puede demostrar para las cotas inferiores de complejidad. En Yarotsky [13] se prueba que una apropiada arquitectura puede aproximar cualquier función $f \in \mathcal{F}_{n,d,\infty,1}$ con un error en norma L^∞ menor que ε con al menos $c \cdot \varepsilon^{-d/(2n)}$ pesos. Acá se extiende este resultado de aproximación al caso de norma $W^{1,\infty}$.

Teorema 3.2.1. *Sea* $d \in \mathbb{N}, n \in \mathbb{N}_{\geq 2}, 1 \leq p \leq \infty, B > 0, 0 \leq s \leq 1$. *Entonces, existe una constante* $c = c(d, n, p, B, s) > 0$ *con la siguiente propiedad:*

Si $\forall \varepsilon \in (0, \frac{1}{2})$ *existe una arquitectura de red neuronal* $\mathcal{A}_\varepsilon = \mathcal{A}_\varepsilon(d, n, p, B, s, \varepsilon)$ *con dimensión de entrada* d *y dimensión de salida* 1 *tal que* $\forall f \in \mathcal{F}_{n,d,p,B}$ *hay una red neuronal* Φ_ε^f *con arquitectura* $\mathcal{A}_\varepsilon$ *tal que*

$$||R_\rho(\Phi_\varepsilon^f) - f||_{W^{s,p}((0,1)^d)} \leq \varepsilon.$$

Entonces $\mathcal{A}_\varepsilon$ *al menos* $M(\mathcal{A}_\varepsilon) \geq c\varepsilon^{\frac{-d}{2(n-k)}}$ *pesos.*

Las cotas superiores para el número de pesos y neuronas son del orden $\mathcal{O}(\varepsilon^{-\delta/(n-s)})$, y así depende fuertemente de la dimensión d del espacio de entrada. Más aún, las constantes en la cota superior para el número de capas, pesos y neuronas también crecen exponencialmente cuando crece d. Aunque todavía existe una brecha entre nuestras cotas superior e inferior para los pesos de una arquitectura de red, el Teorema 3.2.1 muestra que no se puede esperar eludir la maldición de la dimensión en este contexto.

Apéndice A

Definiciones y resultados necesarios

A.1. Espacios de interpolación:

Para las definiciones y el Lema A.1.1 seguimos [10].

Definición A.1.1. Si B_0, B_1 son espacios de Banach tal que B_1 esta continuamente embebido en B_0, escribimos $B_1 \subset B_0$ y que (B_0, B_1) pareja de interpolación.

Definición A.1.2. Sea (B_0, B_1) pareja de interpolación. Para cualquier $u \in B_0$ y $t > 0$, definimos

$$K(t, u, B_0, B_1) := \inf_{v \in B_1} (\|u - v\|_{B_0} + t\|v\|_{B_1})$$

Definición A.1.3. Sea (B_0, B_1) pareja de interpolación. $0 \leq \theta \leq 1$ y $1 \leq p \leq \infty$. Ponemos

$$\|u\|_{(B_0,B_1)_{\theta,p}} := \begin{cases} \left(\int_0^\infty t^{-\theta p} K(t, u, B_0, B_1)^p \frac{dt}{t}\right)^{\frac{1}{p}} & \text{si} \quad 1 \leq p \leq \infty \\ \sup_{0 \leq t \leq \infty} t^{-\theta} K(t, u, B_0, B_1) & \text{si} \qquad p = \infty \end{cases}$$

Más aún, definimos el conjunto

$$(B_0, B_1)_{\theta,p} := \left\{ u \in B_0 : \|u\|_{(B_0,B_1)_{\theta,p}} < \infty \right\}. \tag{A.1}$$

La norma $u \mapsto \|u\|_{(B_0,B_1)_{\theta,p}}$ hace $(B_0, B_1)_{\theta,p}$ un espacio de Banach. que es llamado espacio de interpolación real.

Lema A.1.1. *Sea* $1 \leq p \leq \infty$.

- *si* (B_0, B_1) *es una pareja de interpolación y* $0 < \theta_1 < \theta_2 < 1$,

$$B_1 \subset B_{\theta_2,p} \subset B_{\theta,p} \subset B_0.$$

- *si* B *es un espacio de Banach y* $0 < \theta < 1$, *entonces* $(B, B)_{\theta,p} = B$.

Uno de los teoremas más importantes en la teoría de interpolación muestra cómo la norma del operador sobre una pareja de interpolación se relaciona con la norma del operador con respecto a los espacios de interpolación correspondientes.

Teorema A.1.1. *Sean (A_0, A_1), (B_0, B_1) dos parejas de interpolación. Si $T \in \mathcal{L}(A_0, B_0) \cap \mathcal{L}(A_1, B_1)$, entonces $T \in \mathcal{L}(A_{\theta,p}, B_{\theta,p})$ para cualquier $0 < \theta < 1$ y $1 \leq p \leq \infty$. Además,*

$$\|T\|_{\mathcal{L}(A_{\theta,p},B_{\theta,p})} \leq \|T\|^{1-\theta}_{\mathcal{L}(A_0,B_0)} \|T\|^{\theta}_{\mathcal{L}(A_1,B_1)}.$$

Como corolario del teorema anterior, el siguiente resultado es obtenido.

Corolario A.1.1. *Sea (B_0, B_1) una pareja de interpolación. Más aún, sean $0 < \theta < 1$ y $1 \leq p \leq \infty$, entonces existe una constante $c = c(\theta, p) > 0$ tal que $\forall u \in B_1$*

$$\|u\|_{B_{\theta,p}} \leq c\|u\|^{1-\theta}_{B_0} \|u\|^{\theta}_{B_1}.$$

Para el caso $p = \infty$, $c(\theta, \infty) = 1$.

A.2. Espacios de Sobolev

En esta sección, damos una breve introducción a los espacios de Sobolev y presentamos conceptos relacionados que son importantes para la demostración del Teorema 3.1.1.

Para toda esta sección, sea $d \in \mathbb{N}$ y, salvo que se indique lo contrario, sea $\Omega \subset \mathbb{R}^d$ un subconjunto abierto de $\mathbb{R}^d$. La siguiente definición se puede encontrar en el [2, Capitulo 4].

Para $p \in [1, \infty]$, denotamos por $L^p(\Omega)$ el espacio estándar de Lebesgue sobre Ω.

Definición A.2.1. Sea $n \in \mathbb{N}$ y $1 \leq p \leq \infty$, definimos

$$W^{n,p}(\Omega) \coloneqq \{f \in L^p(\Omega) : D^\alpha f \in L^p(\Omega) \forall \alpha \in \mathbb{N}_0^d \text{con} |\alpha| \leq n\}.$$

Más aún, para $f \in W^{n,p}(\Omega)$ y $1 \leq p < \infty$, definimos la norma

$$\|f\|_{W^{n,p}(\Omega)} \coloneqq \left(\sum_{0 \leq |\alpha| \leq n} \|D^\alpha f\|^p_{L^p(\Omega)} \right)^{1/p}$$

y

$$\|f\|_{W^{n,\infty}(\Omega)} \coloneqq \max_{0 \leq |\alpha| \leq n} \|D^\alpha f\|_{L^\infty(\Omega)}$$

Escribimos $\nabla f := (D_1 f, \ldots, D_d f)$. Además, el espacio $W^{n,p}(\Omega)$, equipado con la norma $\|\cdot\|_{W^{n,p}(\Omega)}$, es un espacio de Banach. Para $f \in W^{n,p}(\Omega)$, usaremos la notación

$$\|f\|_{W^{n,p}(\Omega)} = \|f\|_{W^{n,p}} = \|f(x)\|_{W^{n,p}(\Omega;dx)}$$

si el dominio es claro en el contexto o si queremos enfatizar la variable x de la cual depende la función f, respectivamente.

Si $n = 0$, entonces el espacio de Sobolev es simplemente un espacio de Lebesgue, es decir, $W^{0,p}(\Omega) = L^p(\Omega)$.

Muchos resultados sobre espacios de funciones definidos en un dominio Ω requieren que Ω cumpla ciertas condiciones de regularidad.

Para nuestro propósito, es suficiente enfocarnos en la condición introducida en la siguiente definición enunciada en [5, Apéndice C.1].

Definición A.2.2. Decimos que un conjunto abierto y acotado $\Omega \subset \mathbb{R}^d$ es un dominio Lipschitz si para cada $x_0 \in \partial\Omega$, existe $r > 0$ y una función Lipschitz continua $g : \mathbb{R}^{d-1} \to \mathbb{R}$ tal que

$$\Omega \cap B_{r,|\cdot|}(x_0) = \left\{x \in B_{r,|\cdot|}(x_0) : x_d > g(x_1, \ldots, x_{d-1})\right\}$$

después de una posible reetiquetación y reordenamiento de los ejes coordenados.

En lo que sigue trabajaremos solo con dominios convexos. Más aún, todo conjunto abierto, acotado y convexo es un dominio de Lipschitz.

Enunciaremos un teorema de Extensión que usaremos en la demostración de varios lemas, este resultado está enunciado en [11, Teorema VI.3.1.5].

Teorema A.2.1. *Sea D dominio de Lipschitz. Entonces existe un operador lineal E que mapea funciones sobre D a funciones sobre $\mathbb{R}^d$ con las siguientes propiedades.*

(I) $E(f)|_D = f$, esto es E es un operador extensión.

(II) E mapea $W^{k,p}(\Omega)$ a $W^{k,p}(\mathbb{R}^d)$ de manera continua, para todo $1 \leq p \leq \infty$ y k entero no negativo.

A.2.1. Espacio de Sobolev Fraccionarios

Los espacios de Sobolev fraccionarios juegan un papel importante en el análisis de ecuaciones en derivadas parciales. En particular, caracterizan la regularidad de las funciones de los espacios de Sobolev definidos en un dominio Ω cuando se restringen a la frontera $\partial\Omega$, donde la restricción se realiza mediante un operador de traza.

A continuación, definimos los espacios de orden fraccionario en términos de una norma intrínseca.

Definición A.2.3 (**Espacios de Sobolev–Slobodeckij**)**.** Sea $0 < s < 1$ y $1 \leq p \leq \infty$. Definimos

$$W^{s,p}(\Omega) := \left\{f \in L^p(\Omega) : \|f\|_{W^{s,p}(\Omega)} < \infty\right\}$$

con para $1 \leq p < \infty$,

$$\|f\|_{W^{s,p}(\Omega)} := \left(\|f\|_{L^p(\Omega)} + \int_\Omega \int_\Omega \left(\frac{|f(x) - f(y)|}{|x-y|^{s+d/p}}\right)\right)^{1/p}$$

y

$$\|f\|_{W^{n,\infty}(\Omega)} := \text{máx}\left\{\|f\|_{L^\infty(\Omega)}, \text{ess sup}_{x,y\in\Omega} \frac{|f(x) - f(y)|}{|x-y|^s}\right\}.$$

El espacio $W^{s,p}(\Omega)$, equipado con la norma $\|\cdot\|_{W^{s,p}(\Omega)}$, es un espacio de Banach, llamado espacios de Sobolev–Slobodeckij.

Definición A.2.4. Ponemos para $n \in \mathbb{N}_0, m \in \mathbb{N}$ y $1 \leq p \leq \infty$

$$W^{n,p}(\Omega, \mathbb{R}^d) := \{(f_1, \ldots, f_m) : f_i \in W^{n,p}(\Omega), \forall i = 1, \ldots, m\}.$$

con

$$\|f\|_{W^{n,p}(\Omega,\mathbb{R}^d)} := \left(\sum_{i=1}^m \|f_i\|^p_{W^{n,p}(\Omega)}\right)^{1/p}$$

y

$$\|f\|_{W^{n,\infty}(\Omega)} := \max_{i=1,\ldots,m} \|f_i\|_{W^{n,p}(\Omega)}$$

Estos nuevamente son espacios de Banach.

Definición A.2.5. Sea $n, k \in \mathbb{N}_0$ con $k \leq n$, $m \in \mathbb{N}$ y $1 \leq p \leq \infty$. Para $f \in W^{n,p}(\Omega; \mathbb{R}^m)$, definimos la semi-norma de Sobolev como

$$|f|_{W^{k,p}(\Omega;\mathbb{R}^m)} := \left(\sum_{i=1,\dots,m,|\alpha|=k} \|D^\alpha f_i\|^p_{L^p(\Omega)} \right)^{1/p}, \quad \text{para } 1 \leq p < \infty,$$

y

$$|f|_{W^{k,\infty}(\Omega;\mathbb{R}^m)} := \max_{i=1,\dots,m,|\alpha|=k} \|D^\alpha f_i\|_{L^\infty(\Omega)}.$$

Para $k = 1$ solo escribimos $|\cdot|_{W^{k,p}}$. Notamos que $|\cdot|_{W^{0,p}}$ coincide con $\|\cdot\|_{L^p}$.

Son fáciles de probar los siguientes lemas sobre la composición y el producto de funciones en espacios de Sobolev.

Lema A.2.1. *Sean $d, m \in \mathbb{N}$ y $\Omega_1 \subset \mathbb{R}^d$,$\Omega_2 \subset \mathbb{R}^m$. Sea $f : \Omega_1 \to \Omega_2$ una función L_1- Lipschitz y $g : \Omega_2 \to \mathbb{R}$ una función L_2- Lipschitz para alguna $L_1, L_2 > 0$. Entonces la composición $g \circ f : \Omega_1 \to \mathbb{R}$ es $(L_1 \cdot L_2)-$ Lipschitz.*

Corolario A.2.1. *Sean $d, m \in \mathbb{N}$ y $\Omega_1 \subset \mathbb{R}^d$, $\Omega_2 \subset \mathbb{R}^m$ ambos abiertos acotados y convexos. Entonces, existe una constante $C = C(d, m) > 0$ con las siguientes propiedades:*

Si $f \in W^{1,\infty}(\Omega_1; \mathbb{R}^m)$ y $g \in W^{1,\infty}(\Omega_2)$ son representantes Lipschitz continuos tal que $ranf \subset \Omega_2$ entonces, para la composición $g \circ f : \Omega_1 \to \mathbb{R}$ se mantiene que $g \circ f \in W^{1,\infty}(\Omega_1)y$ tenemos

$$|g \circ f|_{W^{1,\infty}(\Omega_1)} \leq C|g|_{W^{1,\infty}(\Omega_2)}|f|_{W^{1,\infty}(\Omega_1;\mathbb{R}^d)}$$

y

$$\|g \circ f\|_{W^{1,\infty}(\Omega_1)} \leq C \max\left\{\|g\|_{L^\infty(\Omega_2)}, |g|_{W^{1,\infty}(\Omega_2)}|f|_{W^{1,\infty}(\Omega_1;\mathbb{R}^d)}\right\}.$$

Lema A.2.2. *Sea $f \in W^{1,\infty}(\Omega)$ y $g \in W^{1,p}(\Omega)$, entonces $fg \in W^{1,p}(\Omega)$ y existe una constante $C = C(d, p) > 0$ tal que*

$$|fg|_{W^{1,\infty}(\Omega)} \leq C\left(|f|_{W^{1,\infty}(\Omega)}\|g\|_{L^\infty(\Omega)} + \|f\|_{L^\infty(\Omega)}|g|_{W^{1,\infty}(\Omega)}\right).$$

Para $p = \infty$, $C = 1$.

En el siguiente resultado mostramos que interpolamos espacios de Sobolev con exponentes enteros, es equivalente al espacio de Sobolev fraccionario $W_{s,p}$ bajo cierta condición de regularidad en el dominio Ω.

Teorema A.2.2. *Sea $\Omega \subset \mathbb{R}^d$ un dominio Lipschitz. Entonces para $0 < s < 1$ y $1 \leq p \leq \infty$ tenemos que,*

$$W^{s,p}(\Omega) = (L^p(\Omega), W^{1,\infty}(\Omega))_{s,p}$$

A.2.2. Polinomio de Taylor promediado.

Introducimos una generalización de los polinomios de Taylor, que es una de las herramientas clave en la demostración del Teorema 3.1.1 Las demostraciones de los resultados que nombramos se encuentran en [1].

Definición A.2.6. Sea $n \in \mathbb{N}, 1 \leq p \leq \infty$ y $f \in W^{n-1,p}(\Omega)$, y $x_0 \in \Omega$, $r > 0$ tal que para la bola $B \coloneqq B_{r,|\cdot|}(x_0)$, y se mantiene $B \subset\subset \Omega$ ($\overline{B}$ subconjunto compacto en Ω). El correspondiente polinomio de Taylor promediado de orden n de f alrededor de B está definido para $x \in \Omega$ como

$$Q^n f(x) \coloneqq \int_B T_y^n f(x)\phi(y)dy, \tag{A.2}$$

donde

$$T_y^n f(x) \coloneqq \sum_{|\alpha|\leq n-1} \frac{1}{\alpha!} D^\alpha f(y)(x-y)^\alpha, \tag{A.3}$$

y ϕ es una función cut-off arbitraria soportada en $\overline{B}$, i.e.,

$$\phi \in C_c^\infty(\mathbb{R}^d) \quad \text{con} \quad \phi(x) \geq 0, \forall x \in \mathbb{R}^d,$$

$$\operatorname{supp}\phi = \overline{B} \quad \text{y} \quad \int_{R^n} \phi(x)dx = 1,$$

Observación A.2.1. Por la linealidad de la derivada débil, es fácil concluir que el polinomio promediado y el polinomio de Taylor promediado es lineal en f.

Recordemos que el polinomio de Taylor promediado se define mediante una integral y alguna función cut-off(A.2) que realiza un promedio de una expresión polinómica (A.3).

Además, el siguiente lema muestra que el polinomio de Taylor promediado de orden n es, de hecho, un polinomio de grado menor que n en x.

Lema A.2.3. *Sea $n \in \mathbb{N}, 1 \leq p \leq \infty$ y $f \in W^{n-1,p}(\Omega)$, y $x_0 \in \Omega$, $r > 0, R \geq 1$ tal que para la bola $B \coloneqq B_{r,|\cdot|}(x_0)$, y se mantiene $B \subset\subset \Omega$ y $B \subset B_{R,\|\cdot\|_{\ell^\infty}}(0)$. El polinomio de Taylor promediado de orden n de f alrededor de B puede ser escrito como*

$$Q^n f(x) = \sum_{|\alpha|\leq n-1} c_\alpha x^\alpha \tag{A.4}$$

para $x \in \Omega$.

Más aún, existe una constante $c = c(n,d,R) > 0$ tal que los coeficientes c_α son acotados con $|c_\alpha| \leq cr^{-d/p}\|f\|_{W^{n-1,p}(\Omega)}$ para todo α con $|\alpha| \leq n-1$.

El siguiente paso es derivar propiedades de aproximación del polinomio de Taylor promediado. Para este fin, recordemos que, en la expansión de Taylor (estándar) de una función f definida en un dominio Ω, para que la aproximación en algún punto $x_0 + h$ sea válida, todo el camino $x_0 + th$ para $0 \leq t \leq 1$ debe estar contenido en Ω,

En el caso del polinomio de Taylor promediado, el punto de expansión x_0 se reemplaza por una bola B, y requerimos que el camino entre cada $x_0 \in B$ y cada $x \in \Omega$ esté contenido en Ω.

Esta condición geométrica se precisa en la siguiente definición.

Definición A.2.7. Sean $\Omega, B \subset \mathbb{R}^d$. Entonces Ω es llamada estrellada con respecto a B si

$$\overline{\text{conv}}(\{x\} \cup B) \subset \Omega, \forall x \in \Omega.$$

La siguiente definición introduce una noción geométrica que se vuelve importante cuando se considera una familia de subdivisiones $\mathcal{T}^h$, con $0 < h \leq 1$, de un dominio Ω que se vuelve más fina a medida que h disminuye.

Usualmente, es necesario controlar la no-degeneración de $(\mathcal{T}^h)_h$, lo cual se puede lograr, por ejemplo, mediante un parámetro de grosor uniformemente acotado.

Definición A.2.8. Sea $\Omega \subset \mathbb{R}^d$ conjunto acotado. Definimos el conjunto

$$\mathcal{R} \coloneqq \left\{ r > 0 : \exists x_0 \in \Omega \text{ talque } \Omega \text{ es estrellada respecto a } B_{r,|\cdot|}(x_0) \right\}.$$

Si $\mathcal{R} \neq \emptyset$, entonces definimos,

$$r^*_{\text{máx}} \coloneqq \sup \mathcal{R} \quad \text{ y llamamos } \quad \gamma \coloneqq \frac{\text{diam}(\Omega)}{r^*_{\text{máx}}}$$

el parámetro de Chunkines de Ω.

Para enfatizar la dependencia del conjunto Ω, ocasionalmente escribiremos $r_{\text{máx}}(\Omega)$ y $\gamma(\Omega)$

Finalmente, el siguiente lema muestra las propiedades de aproximación del polinomio de Taylor promediado.

Lema A.2.4 (Bramble-Hilbert). *Sea $\Omega \in \mathbb{R}^d$ un conjunto abierto y acotado, $x_0 \in \Omega$ y $r > 0$ tal que Ω es estrellada respecto a $B \coloneqq B_{r,|\cdot|}(x_0)$, y $r > (1/2)r^*_{\text{máx}}$. Más aún, sea $n \in \mathbb{N}, 1 \leq p \leq \infty$ y denotamos γ el parámetro de Chunkines de Ω. Entonces existe una constante $C = C(n, d, \gamma) > 0$ tal que para toda $f \in W^{n,p}(\Omega)$,*

$$|f - Q^n f|_{W^{k,p}(\Omega)} \leq C h^{n-k} |f|_{W^{n,p}(\Omega)}, \quad \textit{para} \quad k = 0, 1, \ldots, n,$$

donde $Q^n f$ es el polinomio promediado de orden n de f alrededor de B y $h = diam(\Omega)$.

Una prueba puede encontrarse en Lema 4.3.8 (Brener- Scott).

Bibliografía

[1] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts Appl. Math.* New York, NY: Springer, 3rd ed. edition, 2008.

[2] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. New York, NY: Springer, 2011.

[3] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.*, 2(4):303–314, 1989.

[4] Ronald Devore, Boris Hanin, and Guergana Petrova. Neural network approximation. *Acta Numerica*, 30:327–444, May 2021. Publisher Copyright: © The Author(s), 2021.

[5] Lawrence C. Evans. *Partial differential equations*, volume 19 of *Grad. Stud. Math.* Providence, RI: American Mathematical Society, 1998.

[6] Ingo Gühring, Gitta Kutyniok, and Philipp Petersen. Error bounds for approximations with deep ReLU neural networks in $W^{s,p}$ norms. *Anal. Appl., Singap.*, 18(5):803–859, 2020.

[7] Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. ReLU deep neural networks and linear finite elements. *J. Comput. Math.*, 38(3):502–527, 2020.

[8] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4:251–257, 1991.

[9] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

[10] Alessandra Lunardi. *Interpolation theory*, volume 16 of *Appunti, Sc. Norm. Super. Pisa (N.S.)*. Pisa: Edizioni della Normale, 3rd edition edition, 2018.

[11] Elias M. Stein. *Singular integrals and differentiability properties of functions*, volume 30 of *Princeton Math. Ser.* Princeton University Press, Princeton, NJ, 1970.

[12] Matus Telgarsky. Representation benefits of deep feedforward networks, 2015.

[13] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.