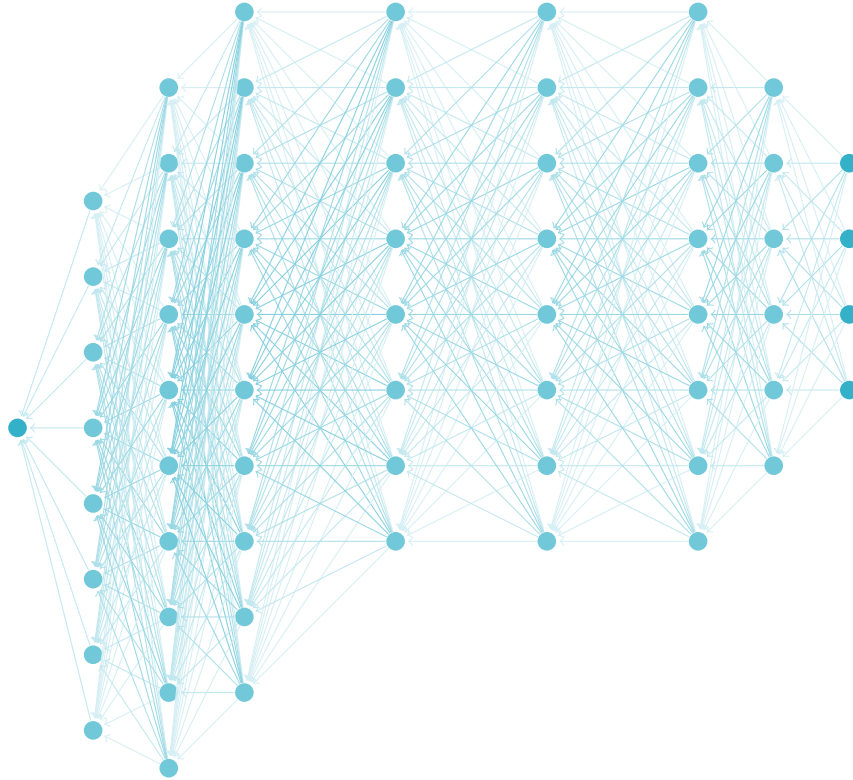


Teoría Matemática de Redes Neuronales

Notas de Clase

Miguel Arturo Ballesteros Montero



SEMINARIO DE APLICACIONES MATEMÁTICAS 2021-1, 2021-2,
UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

<https://github.com/danielrole/Notas-TMRN>

La siguiente recopilación de notas están basadas en los Seminarios de Matemáticas Aplicadas 2021-1 y 2021-2, impartidos por el Dr. Miguel Arturo Ballesteros Montero.

First release, August 2020

Índice general

Prólogo	5
1. ¿Qué es la Inteligencia Artificial?	9
1.1. Definiciones Históricas	9
1.1.1. Actuar Humanamente: El enfoque de la prueba de Turing	9
1.1.2. Pensar Humanamente: El enfoque de modelado cognitivo	10
1.1.3. Actuar Racionalmente: El enfoque del agente racional	10
1.1.4. Pensar Racionalmente: El enfoque de las leyes del pensamiento	11
1.2. ¿Qué es el Machine Learning?	12
1.2.1. Supervised Learning	12
1.2.2. Unsupervised Learning	13
1.2.3. Reinforcement Learning	14
2. Neuronas y Redes Neuronales	17
3. Resultados de Análisis Funcional	21
3.1. Espacios de Banach y Espacios de Hilbert	21
3.2. El Dual de las Funciones Continuas e Integrables	26
4. Aproximación por medio de Redes Neuronales	29
4.1. Universalidad	29
4.2. Redes Neuronales	33
4.3. Operaciones Básicas con Redes Neuronales	35
4.4. Reaproximación de Dicionarios	40

4.5. Aproximación de Funciones Suaves	43
5. Análisis de Datos con Redes Neuronales	57
5.1. Ejemplo	57
5.2. Esquema General	60
5.3. Método de Descenso por el Gradiente (Convergencia)	60

Prólogo

Durante miles de años hemos intentado comprender cómo pensamos, la historia nos ha regalado grandes pensadores que han tratado de estructurar de diversas formas el complejo organismo que somos; la inteligencia artificial intenta ir un paso más allá, intenta no solo comprender si no también construir entidades inteligentes. La inteligencia Artificial es uno de los campos más nuevos de la ciencia y la ingeniería que encontró su origen poco después de la Segunda Guerra Mundial.

Inteligencia Artificial a traves del tiempo:

- **1943-1955:** La gestación de la inteligencia artificial.
 - Modelo de neuronas artificiales de Warren McCulloch y Walter Pitts en 1943.
 - Turing dio conferencias sobre IA desde 1947.
- **1956:** El nacimiento de la inteligencia artificial.
 - John MaCarthy convenció a Marvin Minsky, Claude Shannon y Nathaniel Rochester para que lo ayudaran a reunir investigadores estadounidenses interesados en la teoría de autómatas, las redes neuronales y el estudio de la inteligencia en un taller de dos meses en Dartmouth.
- **1952-1969:** Entusiasmo inicial, grandes exspectativas.
 - Primeros solucionadores y jugadores autómatas, algunos capaces de hacer demostraciones matemáticas.
 - John McCarthy se refirió a este periodo como el “¡Mira mama, sin manos!”.
 - Creación de LISP.
 - *Perceptron* de Frank Rosenblatt en 1958.
 - Adalines (*Adaptive Linear Neuron*) de Bernie Widrow y Marcian Hoff en 1960.

- **1966-1973:** Una dosis de realidad.
 - Prueba y error: explosión combinatoria.
 - Falta de recursos computacionales.
- **1969-1979:** Sistemas basados en el conocimiento ¿La clave del poder?
 - Algoritmos que utilizan conocimientos específicos de dominio en lugar de solucionadores de propósito general.
 - Sistemas expertos para diagnóstico médico.
 - Incorporación de incertidumbre.
- **1980-presente:** La IA se convierte en industria.
 - Optimización de la logística.
 - Auge repentino, pero solo unos pocos proyectos estuvieron a la altura de las expectativas.
 - Invierno IA.
- **1986-presente:** El regreso de las redes neuronales.
 - El algoritmo de retropropagación (Backpropagation) para entrenar redes neuronales se reinventó en “Learning Representation by Back-Propagating Errors.” Por David E. Rumelhart, Geoffrey E. Hinton y Ronald J. Williams.
- **1987-presente:** La IA adopta el método científico.
 - Modelos ocultos de Markov.
 - Redes Bayesianas.
- **1995-presente.**
 - Internet impulsa el desarrollo de agentes inteligentes, por ejemplo:
 - Chatbots.
 - Sistemas de recomendación.
 - Recomendaciones de amistades.
 - Acceso a recursos de computación a suficiente velocidad.
 - La era del *Big Data*: Gran cantidad de *label training data*, por ejemplo:
 - Diccionarios.

- Wordnets.
- Wikipedia.
- Google.
- Los fundadores de la IA descontentos con su estado actual:
 - La IA debería volver a sus raíces de luchar por, en palabras de Herbert Simon, “Maquinas que piensan, que aprenden y crean”.
- La IA en la cultura popular.
 - Lucha contra el spam.
 - Reconocimiento de voz: Siri, Alexa, Cortana.
 - Reconocimiento facial: Facebook, Apple Photos, Google Photos.
 - *Deep Blue vs Garry Kasparov*.
 - Planificación y programación autónoma: Mars rover de la NASA.
 - Vehículos robóticos: EL automóvil autónomo de Tesla.
 - Traducción automática: Traductor de Google.
- **Hoy** Tú empiezas a leer estas notas.

1 | ¿Qué es la Inteligencia Artificial?

1.1. Definiciones Históricas

1.1.1. Actuar Humanamente: El enfoque de la prueba de Turing

“El arte de crear maquinas que realizen funciones que requieren inteligencia cuando las personas lo ejecuten”

Kurzweil, 1990

“El estudio de hacer a las computadoras hacer cosas en las que de momento, somos mejores.”

Rich and Knight, 1991

En 1950 Turing ideó una prueba para proporcionar una definición operativa satisfactoria de inteligencia;. Una computadora pasa la prueba si un interrogador humano, después de formular algunas preguntas escritas, no puede decir si las respuestas escritas provienen de una persona o una computadora.

La computadora debe poseer las siguientes características:

- **Procesamiento del lenguaje natural:** para comunicarse, por ejemplo en inglés.
- **Representación del conocimiento:** para almacenar información.
- **Razonamiento automatizado:** utilizar la información almacenada, responder preguntas y sacar conclusiones.
- **Aprendizaje automático:** adaptarse a nuevas circunstancias, extrapolar y detectar patrones.

La prueba de Turing sigue siendo relevante incluso hoy, pero menos desde la ingeniería y más desde la postura filosófica; algunos ejemplos parecidos son:

- La búsqueda del “vuelo artificial” tuvo éxito cuando los hermanos Wright y otros dejaron de imitar a las aves y comenzaron a usar túneles de viento y aprender sobre aerodinámica.
- Los textos de ingeniería aeronáutica no definen el objetivo de su campo como fabricar “Máquinas que vuelan tan exactamente como las palomas que pueden engañar incluso a otras palomas”.

1.1.2. Pensar Humanamente: El enfoque de modelado cognitivo

“El nuevo y emocionante esfuerzo de hacer que las computadoras piensen [...] máquinas con mentes, en el sentido completo y literal.”

Haugeland, 1985

“[La automatización de] actividades que asociamos con el pensamiento humano, actividades como la toma de decisiones, la resolución de problemas, el aprendizaje [...]”

Bellman, 1978

Para saber si un programa “piensa como un ser humano”, debemos aprender qué es el pensamiento humano:

- Observa nuestros pensamientos a medida que pasan,
- observar a las personas durante una tarea,
- realizar experimentos psicológicos/neurocientíficos.

Sin embargo, para nuestro esfuerzo, será una buena práctica mantener separados los campos como la ciencia cognitiva, la neurociencia, la psicología y la filosofía el mayor tiempo posible.

1.1.3. Actuar Racionalmente: El enfoque del agente racional

“La Inteligencia Computacional es el estudio del diseño de agentes inteligentes”

Poole et al., 1998

“IA [...] se preocupa por el comportamiento inteligente de los artefactos”

Nilsson, 1998

Crear agentes (por ejemplo, programas informáticos) que operen de forma autónoma, perciban su entorno, persistan durante un período de tiempo prolongado, se adapten al cambio, creen, persigan metas. Hacer inferencias es el caso extremo de ser un agente racional.

En muchas ocasiones, sin embargo, no es posible realizar inferencias correctas, por ejemplo:

- Cuando la comprensión del medio ambiente es insuficiente, o
- no hay suficientes datos de entrada para basar una decisión.

Un agente racional es aquel que actúa para lograr el mejor resultado o, cuando hay incertidumbre, el mejor resultado esperado.

De esta forma el estándar de racionalidad está matemáticamente bien definido y es completamente general (como mostraremos en la siguiente sección: “El enfoque de las leyes del pensamiento”); por otro lado, el comportamiento humano esta bien adaptado para un entorno específico y se define a sí mismo por la suma total de todas las cosas que hacen los humanos.

1.1.4. Pensar Racionalmente: El enfoque de las leyes del pensamiento

Un uso de razonamiento lógico y argumentación, por ejemplo:

- **Deducción:** Una regla general aplicada a un caso particular implica un resultado trivial.

Entrada	Implicación	Ejemplo
REGLA		En un planeta, su sol sale todos los días
CASO		Estamos en un planeta
	RESULTADO	El sol sale todos los días

El reino de las matemáticas.

- **Inducción:** De un resultado trivial en un caso particular esperamos inferir la regla general.

Entrada	Implicación	Ejemplo
RESULTADO		El sol salía todos los días
CASO		Estamos en un planeta
	REGLA	En un planeta, su sol sale todos los días

El reino de la ciencia.

- **Secuestro:** De una regla general y un resultado trivial esperamos inferir el caso particular.

Entrada	Implicación	Ejemplo
REGLA		En un planeta, su sol sale todos los días
RESULTADO		El sol salía todos los días
	CASO	Estamos en un planeta

Más raramente utilizado.

1.2. ¿Qué es el Machine Learning?

“[Machine Learning] brinda a las computadoras la capacidad de aprender sin estar programadas explícitamente.”

Samuel, 1959

Dicho de otra manera, el Machine Learning es un sub-campo de la Inteligencia Artificial en donde se buscan algoritmos “suaves” que, hasta cierto punto, pueden adaptarse a un cierto tipo de tarea en lugar de consistir simplemente en una lógica codificada que busca manejar grandes cantidades de datos y realizar, por ejemplo:

- Estructuración de datos.
- Encontrar correlaciones.
- Clasificar datos.
- Reconocer patrones.
- Comprender datos.
- Tomar decisiones basadas en datos.
- Adaptar tareas a datos.
- Extrapolación / predicción.

1.2.1. Supervised Learning

Algoritmos “suaves” que inferen la tarea designada mediante la inspección de los datos de entrenamiento apropiados.

Ejemplos:

- Clasificación: Predicción de clases discretas.
 - Determinar si el correo electrónico es spam o no.
 - Identificar imágenes borrosas.
- Regresión: Predicción de parámetros continuos.
 - Consumo de energía de acuerdo con el comportamiento aprendido del usuario.
 - Predicción de una tendencia según una historia determinada.

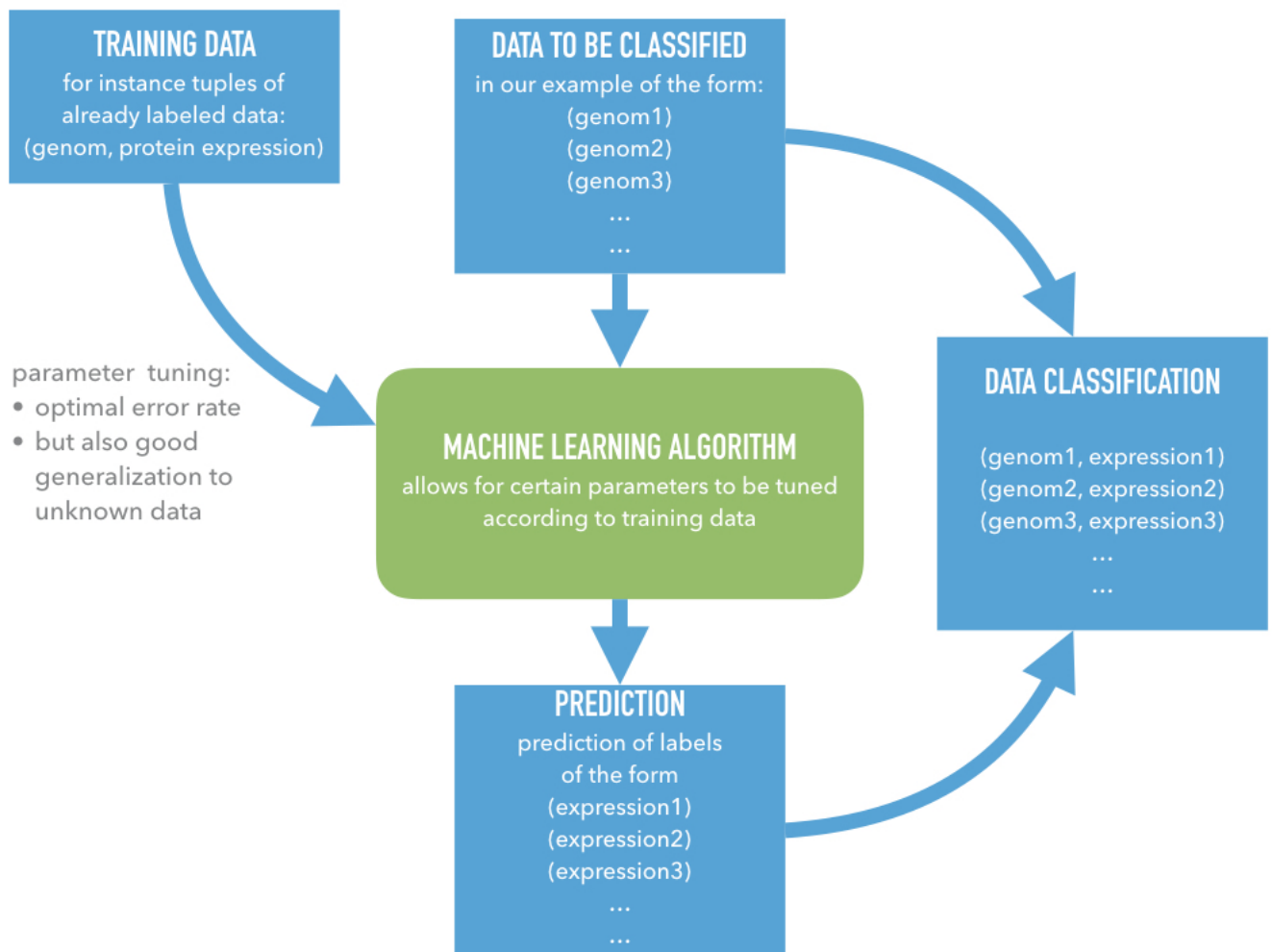


Figura 1.1: Algoritmo Machine Learning

1.2.2. Unsupervised Learning

Estructurar los datos en grupos sin conocimientos previos detallados.

Ejemplos:

- Wordnets: relaciones entre palabras de un lenguaje natural.
- Referencias cruzadas entre documentos.
- Comprensión de datos y reducción de dimensionalidad.

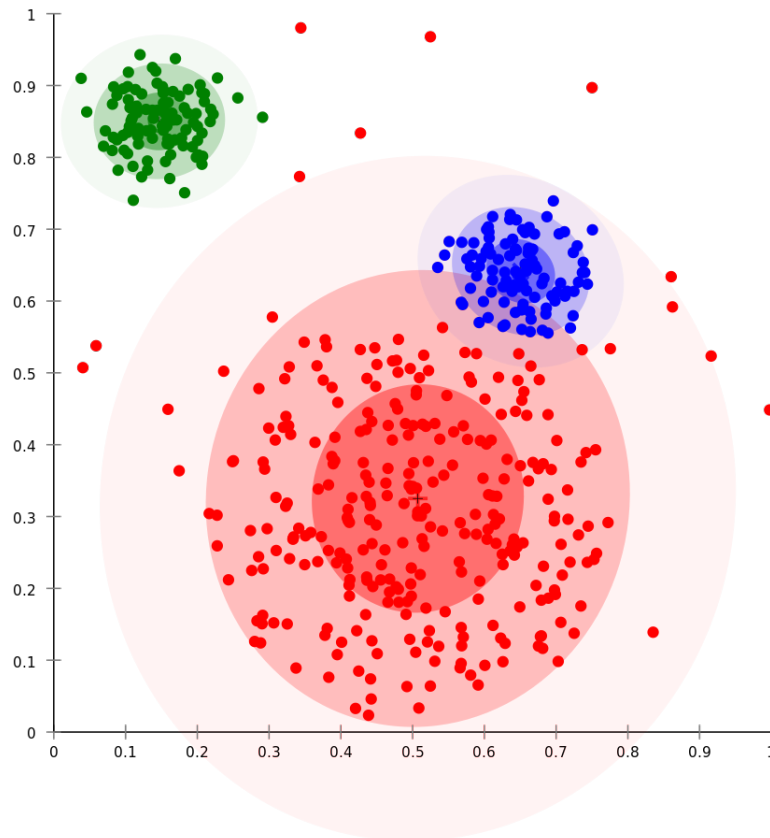


Figura 1.2: Ejemplo ficticio de puntos de datos 2D. El color indica una relación entre los puntos de datos. A partir de estas relaciones, las regiones sombreadas pueden inferirse mediante un algoritmo de aprendizaje automático no supervisado. Esto puede ser útil cuando se buscan propiedades estructuradas burdas de un conjunto de datos.

1.2.3. Reinforcement Learning

Un agente (Machine Learning program + artifact) aprende a realizar una determinada tarea mediante, por ejemplo, prueba y error. El aprendizaje se ve facilitado por la capacidad de observar el entorno y recibir retroalimentación en función de las acciones.

Ejemplos:

- Movimiento de un robot en terreno desconocido o bajo condiciones variables.
- Obteniendo puntuaciones altas en juegos de Atari como *Google deepmind*.

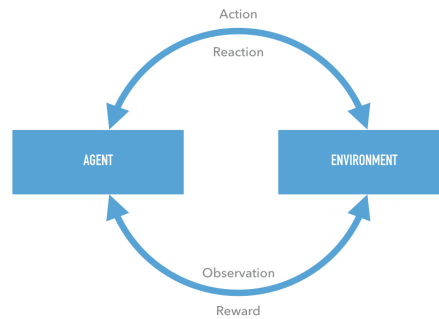


Figura 1.3

2 | Neuronas y Redes Neuronales

En este curso estudiaremos objetos matemáticos llamados redes neuronales. Estos son elementos centrales en IA (Inteligencia Artificial).

Históricamente, la primera neurona en el contexto del cerebro fue propuesta por McCulloch y Pitts en 1943.

$$\hat{x} \rightarrow \chi_{\mathbb{R}^+} \left(\sum w_i x_i - \theta \right),$$

en donde:

- $\hat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$ con $d \in \mathbb{N}$,
- para todo i , los *pesos* $w_i \in \mathbb{R}$,
- el *umbral* $\theta \in \mathbb{R}$,
- $\mathbb{R}^+ = [0, \infty]$
- y la función de activación $\chi_{\mathbb{R}^+}$ se define, para cada conjunto A , como

$$\chi_A(y) = \begin{cases} 1 & \text{si } y \in A, \\ 0 & \text{si } y \notin A. \end{cases}$$

Si la combinación lineal $\sum x_i w_i$ es mayor que θ , entonces la neurona se activa, de lo contrario no lo hace.

La función:

$$\hat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} \rightarrow \sum x_i w_i - \theta$$

es una transformación afín (una transformación lineal compuesta con una traslación) y la función $\chi_{\mathbb{R}^+}$ es una función no lineal.

De forma más general, definimos una neurona como una transformación afín compuesta con una función no lineal.

Definición 2.1: Neurona

Sean $T : \mathbb{R}^m \rightarrow \mathbb{R}^n$ una transformación afín y $\rho : \mathbb{R} \rightarrow \mathbb{R}$ una función no lineal. Definimos una neurona como la composición:

$$\hat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \Rightarrow \rho(T(\hat{x})) := \begin{pmatrix} \rho(T(\hat{x})_1) \\ \vdots \\ \rho(T(\hat{x})_n) \end{pmatrix}$$

en donde:

- $\hat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \in \mathbb{R}^m$,
- $T(\hat{x}) = \begin{pmatrix} T(\hat{x})_1 \\ \vdots \\ T(\hat{x})_n \end{pmatrix} \in \mathbb{R}^n$.

La función ρ es la función de activación y la transformación afín T contiene a los pesos y al umbral (de forma abstracta).

Una red neuronal es una composición de neuronas. En nuestro caso consideramos una única función de activación para todas las neuronas de la red.

Definición 2.2: Red Neuronal

Una red neuronal de L niveles consiste de L transformaciones afines, T_1, T_2, \dots, T_L y una función de activación $\rho : \mathbb{R} \rightarrow \mathbb{R}$. Definimos la red neuronal como la composición

$$F = T_L \circ \rho \circ T_{L-1} \circ \rho \circ \dots \circ \rho \circ T_1$$

en donde estamos identificando a la función ρ con funciones $\rho : \mathbb{R}^n \rightarrow \mathbb{R}^n$ de tal forma que

$$\rho \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} := \begin{pmatrix} \rho(Y_1) \\ \vdots \\ \rho(Y_n) \end{pmatrix}.$$

(Esto es un abuso de notación que siempre se usa)

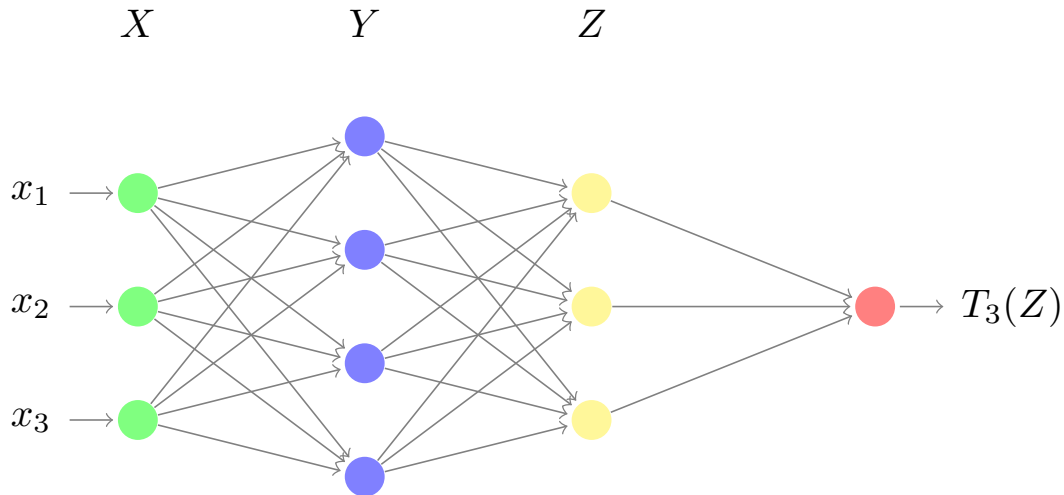


Figura 2.1: Ejemplo de Red Neuronal.

- $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$

- $Y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \rho(T_1(X))$

- $Z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} = \rho(T_2(Y))$

A la función F se le nombra un “Multilayer Perceptron”(MLP) de dimensión d , L niveles y función de activación ρ ; donde el dominio de T_1 es \mathbb{R}^d .

3 | Resultados de Análisis Funcional

3.1. Espacios de Banach y Espacios de Hilbert

Definición 3.1: Norma

Sea V un espacio vectorial. Una norma en V es una asignación $\hat{x} \rightarrow \|\hat{x}\|$ para todo \hat{x} en V , tal que:

1. $\|\hat{x}\| \geq 0$,
2. $\|\hat{x}\| = 0 \Leftrightarrow \hat{x} = \hat{0}$,
3. $\|\alpha \cdot \hat{x}\| = |\alpha| \|\hat{x}\| \quad \forall \alpha \in \mathbb{R}$,
4. $\|\hat{x} + \hat{y}\| \leq \|\hat{x}\| + \|\hat{y}\|$.

Definición 3.2: Producto Interior

Sea V un espacio vectorial. Un producto interior en el espacio de los números reales es una asignación $\hat{x}, \hat{y} \rightarrow \langle \hat{x}, \hat{y} \rangle$ para cualesquiera \hat{x}, \hat{y} en V , tal que:

1. $\langle \hat{u} + \hat{v}, \hat{w} \rangle = \langle \hat{u}, \hat{w} \rangle + \langle \hat{v}, \hat{w} \rangle$,
2. $\langle \lambda \hat{v}, \hat{w} \rangle = \lambda \langle \hat{v}, \hat{w} \rangle \quad \forall \lambda \in \mathbb{R}$,
3. $\langle \hat{u}, \hat{v} \rangle = \langle \hat{v}, \hat{u} \rangle$,
4. $\langle \hat{v}, \hat{v} \rangle \geq 0$, $\langle \hat{v}, \hat{v} \rangle = 0 \Leftrightarrow \hat{v} = \hat{0}$.

Notación 3.1

Todo producto interior induce una norma de la forma

$$\|\hat{x}\| = \sqrt{\langle \hat{x}, \hat{x} \rangle}.$$

Definición 3.3: Completitud

Decimos que un espacio vectorial normado V es completo si toda sucesión de Cauchy converge, es decir:

Sea $(\hat{v}_n)_{n \in \mathbb{N}}$ una sucesión de elementos en V tal que

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}, \text{ tal que } \forall n, m \geq N, \|\hat{v}_n - \hat{v}_m\| < \varepsilon,$$

entonces existe

$$\hat{v} = \lim_{n \rightarrow \infty} \hat{v}_n.$$

Definición 3.4: Espacios de Banach y de Hilbert

- Un espacio de Banach es un espacio vectorial normado y completo.
- Un espacio de Hilbert es un espacio vectorial con producto interior y completo.

Notación 3.2

Sea V un espacio vectorial normado. Se usará la siguiente notación:

- Bola:

Dado $\hat{x} \in V$ y $r > 0$, denotamos a la bola con centro en \hat{x} y radio r por

$$B(\hat{x}, r) := \{y \in V \mid \|\hat{y} - \hat{x}\| < r\}.$$

- Conjunto Abierto:

Dado $A \subset V$, decimos que A es abierto si para todo $\hat{a} \in A$, existe $r > 0$ tal que $B(\hat{a}, r) \subset A$.

- Conjunto Cerrado:

Dado $B \subset V$, decimos que B es cerrado si B^c es abierto o equivalentemente si la siguiente proposición es verdadera:

Si $\hat{y} \in V$ es tal que, para todo $\epsilon > 0$ tal que $B(\hat{y}, \epsilon) \cap B \neq \emptyset$, entonces $\hat{y} \in B$.

- Cerradura:

La cerradura de un conjunto C es el conjunto cerrado mas chico que contiene a C . Denotamos con \bar{C} a la cerradura de C .

El conjunto \bar{C} se caracteriza con la siguiente propiedad:

$$\hat{y} \in \bar{C} \Leftrightarrow \forall \epsilon > 0, B(\hat{y}, \epsilon) \cap C \neq \emptyset.$$

Notemos que C es cerrado $\Leftrightarrow C = \bar{C}$.

- Interior:

El interior de un conjunto A es el conjunto abierto mas grande contenido en A . Denotamos con A° al interior de A .

Definición 3.5: Densidad

Sean V un espacio vectorial normado y $M \subset V$. Decimos que M es denso si se cumple alguno de los siguientes puntos:

1. $\bar{M} = V$.
2. $\forall \hat{v} \in V, \forall \epsilon > 0$ tal que $B(\hat{v}, \epsilon) \cap M \neq \emptyset$.

Definición 3.6: Norma de una transformación lineal continua

Sean V y W espacios vectoriales normados y $\Lambda : V \rightarrow W$ una transformación lineal. Definimos:

$$\|\Lambda\| := \sup\{\|\Lambda(\hat{v})\| \mid \hat{v} \in V \text{ y } \|\hat{v}\| \leq 1\}.$$

Note que Λ es continua si y sólo si $\sup\{\|\Lambda(\hat{v})\| \mid \hat{v} \in V \text{ y } \|\hat{v}\| \leq 1\} < \infty$.

Teorema 3.1: Hahn-Banach

Sean V un espacio vectorial normado y $M \subset V$ un sub-espacio vectorial. Supongamos $h : M \rightarrow \mathbb{R}$ es continua y lineal, entonces existe una extensión $g : V \rightarrow \mathbb{R}$ (lineal y continua) tal que:

1. $g(m) = h(m) \ \forall m \in M$,
2. $\|h\| = \|g\|$.

Corolario 3.1

Sean V un espacio vectorial normado y $M \subset V$ un sub-espacio vectorial. Suponemos que M no es denso (i.e. $\overline{M} \neq V$), entonces existe una función lineal $g : V \rightarrow \mathbb{R}$ continua tal que:

1. $g|_{\overline{M}} = 0$,
2. $g \neq 0$.

Demostración. Como $\overline{M} \neq V$, entonces existe $\hat{x} \in V \setminus \overline{M}$. Sea W el espacio vectorial generado por \overline{M} y \hat{x} , entonces todo elemento de W se puede escribir de manera única de la siguiente manera:

$$\hat{w} = \alpha \hat{x} + \hat{m}$$

en donde $\alpha \in \mathbb{R}$ y $m \in \overline{M}$; es decir

$$W = \{\alpha \hat{x} + \hat{m} \mid \alpha \in \mathbb{R}, \hat{m} \in \overline{M}\}.$$

Veamos que los elementos de W se pueden escribir de forma única como se describe antes.

Supongamos que existen α_1, α_2 en \mathbb{R} y \hat{m}_1, \hat{m}_2 en \overline{M} tales que:

$$\begin{aligned} \hat{w} &= \alpha_1 \hat{x} + \hat{m}_1 = \alpha_2 \hat{x} + \hat{m}_2, \\ \Rightarrow (\alpha_1 - \alpha_2) \hat{x} &= \hat{m}_2 - \hat{m}_1 \in \overline{M}. \end{aligned}$$

(Pues como M es espacio vectorial, entonces \overline{M} también lo es).

Concluimos que $(\alpha_1, \alpha_2)\hat{x} \in \overline{M}$ pues si $\alpha_1 - \alpha_2 \neq 0$, entonces tendríamos que existe $\hat{y} \in \overline{M}$ tal que $\hat{x} = (\frac{1}{\alpha_1 - \alpha_2})\hat{y} \in \overline{M}$, lo que contradice que $\hat{x} \notin \overline{M}$.

Entonces $\alpha_1 - \alpha_1 = 0$, por lo que $\alpha_1 = \alpha_2$. Como $(\alpha_1 - \alpha_2)\hat{x} = m_2 - m_1$, concluimos que $m_1 = m_2$.

Definimos a $h : W \rightarrow \mathbb{R}$ como $h(\alpha\hat{x} + \widehat{m}) = \alpha\|\hat{x}\|$, que es lineal, pues:

$$\begin{aligned} h(\lambda(\alpha_1\hat{x} + \widehat{m}_1) + \alpha_2\hat{x} + \widehat{m}_2) &= h(\lambda\alpha_1 + \alpha_2)\hat{x} + \lambda\widehat{m}_1 + \widehat{m}_2 \\ &= (\lambda\alpha_1 + \alpha_2)\|\hat{x}\| = \lambda h(\alpha_1\hat{x} + \widehat{m}_1) + h(\alpha_2\hat{x} + \widehat{m}_2), \end{aligned}$$

en donde $\lambda\widehat{m}_1 + \widehat{m}_2 \in \overline{M}$.

Para toda $\lambda \in \mathbb{R}$, $\alpha_1, \alpha_2 \in \mathbb{R}$, $\widehat{m}_1, \widehat{m}_2 \in \overline{M}$.

Además $h(m) = 0$, $\forall m \in \overline{M}$.

Veamos que $h : W \rightarrow \mathbb{R}$ es continua para lo cual verificamos que

$$\|h\| = \sup\{|h(\hat{w})| \mid \hat{w} \in W, \|\hat{w}\| \leq 1\} \leq \infty.$$

Como $\hat{x} \in V \setminus \overline{M}$, entonces

$$\exists \varepsilon > 0 \text{ tal que } B(\hat{x}, \varepsilon) \cap \overline{M} = \emptyset.$$

Sea $\hat{w} = \alpha\hat{x} + \widehat{m} \in W$ tal que $\|\hat{w}\| \leq 1$, tenemos que

$$\|\alpha\hat{x} + \widehat{m}\| = |\alpha|\|\hat{x} + \frac{1}{\alpha}\widehat{m}\| \leq 1$$

y además como $\frac{1}{\alpha}\widehat{m} \in \overline{M}$, se obtiene que

$$\|\hat{x} - (-\frac{1}{\alpha}\widehat{m})\| = \|\hat{x} + \frac{1}{\alpha}\widehat{m}\| \geq \varepsilon.$$

Concluimos que

$$|\alpha|\|\hat{x} + \frac{1}{\alpha}\widehat{m}\| \leq 1 \Rightarrow |\alpha| \leq \frac{1}{\|\hat{x} + \frac{1}{\alpha}\widehat{m}\|} \leq \frac{1}{\varepsilon}.$$

Se tiene que

$$|\alpha| \leq \frac{1}{\varepsilon}, \forall \hat{w} = \alpha\hat{x} + \widehat{m} \in W \text{ tal que } \|\hat{w}\| \leq 1,$$

de lo que obtenemos

$$|h(\hat{w})| = |\alpha|\|\hat{x}\| \leq \frac{1}{\varepsilon}\|\hat{x}\| \quad \forall \hat{w}, \text{ con } \|\hat{w}\| \leq 1, \text{ con } w = \alpha\hat{x} + \widehat{m}.$$

Concluimos que

$$\sup\{|h(\hat{w})| \mid \hat{w} \in W, \|\hat{w}\| \leq 1\} = |h| \leq \frac{1}{\varepsilon}\|\hat{x}\|$$

y por lo tanto es continua.

Así, h es una función continua tal que $h \neq 0$, puesto que $h(\hat{x}) = \|\hat{x}\|$ y además $h(\widehat{m}) = 0, \forall \widehat{m} \in \overline{M}$.

Por el Teorema de Hahn-Banach, existe g lineal y continua tal que:

- $g(\widehat{w}) = h(\widehat{w}) \forall \widehat{w} \in W$,
- $\|g\| = \|h\|$ (en particular g es continua).

g es la función que buscamos, pues:

- $g \neq 0$ (ya que $h \neq 0$),
- $g(\widehat{m}) = h(\widehat{m}) = 0 \forall \widehat{m} \in \overline{M}$.

□

3.2. El Dual de las Funciones Continuas e Integrables

Definición 3.7: Dual de un espacio vectorial

Sea V un espacio vectorial normado. Denotamos por V^\bullet al conjunto de funciones lineales y continuas de V en \mathbb{R} .

Definición 3.8

Sea $K \subset \mathbb{R}$ un conjunto compacto (cerrado y acotado). Denotamos por

$$C(K) := \{f : K \rightarrow \mathbb{R} : f \text{ es continua} \}$$

al espacio vectorial de las funciones continuas en K con valores en \mathbb{R} , dotado de la norma.

$$\|f\|_\infty := \sup\{|f(\hat{x})| : \hat{x} \in K\}.$$

Recordemos que $C(K)^\bullet$ es el dual de $C(K)$ el cual es el conjunto de funciones lineales y continuas de $C(K)$ en \mathbb{R} y que la función lineal $h : C(K) \rightarrow \mathbb{R}$ es continua si y sólo si

$$\|h\| := \sup\{|h(\hat{x})| : \hat{x} \in C(K)\} < \infty.$$

Definición 3.9: Teorema de Representación de Riez

Decimos que una función f acotada, $f : K \rightarrow \mathbb{R}$, es integrable respecto a $h \in C(K)$ si existe una sucesión de funciones $\{f_n\}_{n \in \mathbb{N}}$ en $C(K)$ tal que:

1. Existe $c \in \mathbb{R}$ tal que

$$|f_n(\hat{x})| \leq c \quad \forall \hat{x}, \forall n;$$

2. $\lim_{n \rightarrow \infty} f_n(\hat{x}) = f(\hat{x}) \quad \forall \hat{x} \in K.$

Notación 3.3

Denotaremos por $L_a^1(h)$ al espacio de funciones integrables y acotadas con respecto a h .

Notemos que en esta definición $L_a^1(h)$ no depende de h , pero mantenemos h en nuestra notación ya que se utilizará posteriormente.

Teorema 3.2: Convergencia Dominada de Lebesgue 1

Sean $f \in L_a^1(h)$ para algún $h \in C(K)^\bullet$ y $\{f_n\}_{n \in \mathbb{N}}$ como en 3.9, entonces $\lim_{n \rightarrow \infty} h(f_n)$, existe y si $\{g_n\}_{n \in \mathbb{N}}$ es otra sucesión que cumple 1 y 2 de la definición 3.9, entonces

$$\lim_{n \rightarrow \infty} h(f_n) = \lim_{n \rightarrow \infty} h(g_n).$$

Es fácil ver que $L_a^1(h)$ es un espacio vectorial y que la función $f \rightarrow \int f dh$ es lineal.

Teorema 3.3: Convergencia Dominada de Lebesgue 2

Sea $h \in C(K)^\bullet$. Supongamos que $g \in L_a^1(h)$ es tal que existe una sucesión $\{g_n\}_{n \in \mathbb{N}}$ de elementos en $L_a^1(h)$, que cumple lo siguiente:

1. Existe $G \in L_a^1(h)$ tal que

$$|g_n(\hat{x})| \leq G(\hat{x}) \quad \forall \hat{x} \in K, \forall n \in \mathbb{N};$$

2. $\lim_{n \rightarrow \infty} g_n(\hat{x}) = g(\hat{x}) \quad \forall \hat{x} \in K,$

entonces

$$\lim_{n \rightarrow \infty} \int g_n dh = \int g dh.$$

Definición 3.10

Sea $K \subset \mathbb{R}^d$ compacto. Dado $x \in \mathbb{R}^d$ y $a \in \mathbb{R}^d$, denotamos por

$$\langle \hat{a}, \hat{x} \rangle = \sum_{i=1}^d \hat{a}_i \hat{x}_i$$

en donde

$$\hat{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_d \end{pmatrix}, \quad \hat{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}.$$

Lema 3.1

Dados $b_1, b_2 \in \mathbb{R}$ con $b_1 < b_2$ y $f \in C(K)$, se cumple que

$$\chi_{[b_1, b_2]} \circ f \in L_a^1(h),$$

en donde $f : K \subset \mathbb{R}^d \rightarrow \mathbb{R}$ y $\chi_{[b_1, b_2]} : \mathbb{R} \rightarrow \mathbb{R}$ es tal que

$$\chi_{[b_1, b_2]}(y) = \begin{cases} 1 & y \in [b_1, b_2], \\ 0 & y \notin [b_1, b_2]. \end{cases}$$

Teorema 3.4

Dado $\hat{a} \in \mathbb{R}^d$ definimos la función $f_{\hat{a}} : \mathbb{R}^d \rightarrow \mathbb{R}$ dada por $f_{\hat{a}}(\hat{x}) = \langle \hat{a}, \hat{x} \rangle$.

Sea $h \in C(K)^\bullet$. Si para cualesquiera $a \in \mathbb{R}^d$ y $b_1, b_2 \in \mathbb{R}$ con $b_1 < b_2$ se cumple que

$$\int \chi_{[b_1, b_2]} \circ f_{\hat{a}} dh = 0,$$

entonces $h = 0$.

4 | Aproximación por medio de Redes Neuronales

4.1. Universalidad

Recordemos que una función F de la forma

$$F = T_L \circ \rho \circ T_{L-1} \circ \rho \cdots \circ \rho \circ T_1,$$

en donde T_1, \dots, T_L son transformaciones afines y $\rho : \mathbb{R} \rightarrow \mathbb{R}$ es una función de activación, es nombrada “multilayer perceptron”(MLP) con dimensión de entrada d , L niveles y función de activación ρ ; en donde el dominio de T_1 es \mathbb{R}^d .

Definición 4.1: MLP

Sean $\rho : \mathbb{R} \rightarrow \mathbb{R}$ continua, $d, L \in \mathbb{N}$ y $K \subset \mathbb{R}^d$ compacto. Denotamos por $MLP(\rho, d, L)$ al conjunto de MLP's con función de activación ρ , L niveles y dimensión de entrada d tales que T_L toma valores en \mathbb{R} ; es decir, si $f \in MLP(\rho, d, L)$, entonces

$$f : \mathbb{R}^d \rightarrow \mathbb{R}.$$

Definición 4.2: Universalidad

Decimos que $MLP(\rho, d, L)$ es universal si las restricciones de las funciones $f \in MLP(\rho, d, L)$ son densas en $C(K)$ para cualquier compacto $K \subset \mathbb{R}^d$.

Normalmente se identifican las funciones $f : \mathbb{R}^d \rightarrow \mathbb{R}$ con sus restricciones a K , $f|_K \equiv f : K \rightarrow \mathbb{R}$.

Una manera de interpretar que el conjunto de MLP's sea denso en el conjunto de funciones continuas, es que para cada función continua, existe un MLP que lo aproxima.

Definición 4.3: Discriminante

Sean $d \in \mathbb{N}$ y $K \subset \mathbb{R}^d$ compacto. Una función $f : \mathbb{R} \rightarrow \mathbb{R}$ continua es llamada discriminante si dada $h \in C(K)^\bullet$ se cumple la siguiente proposición para toda $\hat{a} \in \mathbb{R}^d$, $b \in \mathbb{R}$:

$$h(t_{\hat{a},b}(f)) = 0 \Rightarrow h = 0.$$

En donde denotamos por $t_{\hat{a},b}(f) \in C(K)$ a la función dada por $t_{\hat{a},b}(f)(\hat{x}) = f(\langle \hat{a}, \hat{x} \rangle - b)$.

Proposición 4.1

$MLP(\rho, d, 2)$ es un espacio vectorial.

Demostración. Sean $F, G \in MLP(\rho, d, 2)$ y $\alpha \in \mathbb{R}$, veamos que

$$\alpha F + G \in MLP(\rho, d, 2).$$

Como $F, G \in MLP(\rho, d, 2)$, entonces existen transformaciones afines T_1, T_2, V_1, V_2 tales que:

- $F = T_2 \circ \rho \circ T_1$,
- $G = V_2 \circ \rho \circ V_1$;

en donde T_1 y V_1 tiene dominio \mathbb{R}^d y T_2, V_2 toman valores en \mathbb{R} .

Definimos

$$U_1(\hat{x}) = \begin{pmatrix} T_1(\hat{x}) \\ V_1(\hat{x}) \end{pmatrix}, \quad \forall \hat{x} \in \mathbb{R}^d;$$

U_1 es una transformación afín.

Dados y, z con y en el dominio de T_2 y z en el dominio de V_2 , definimos

$$U_2 \begin{pmatrix} y \\ z \end{pmatrix} = \alpha T_2(y) + V_2(z).$$

Tenemos que

$$\begin{aligned} U_2 \circ \rho \circ U_1(\hat{x}) &= U_2 \rho \begin{pmatrix} T_1(\hat{x}) \\ V_1(\hat{x}) \end{pmatrix} \\ &= \alpha T_2(\rho(T_1(\hat{x}))) + V_2(\rho(V_1(\hat{x}))) \\ &= \alpha F(\hat{x}) + G(\hat{x}). \end{aligned}$$

Entonces $\alpha F + G = U_2 \circ \rho \circ U_1 \in MLP(\rho, d, 2)$, de lo que concluimos que $MLP(\rho, d, 2)$ es un espacio vectorial. \square

Teorema 4.1: Aproximación Universal, Gybenko 1989

Sean $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compacto y $\rho : \mathbb{R} \rightarrow \mathbb{R}$ una función discriminante. Entonces $MLP(\rho, d, 2)$ es denso en $C(K)$.

Demostración. Procederemos por contradicción:

Supongamos que $MLP(\rho, d, 2)$ no es denso en $C(K)$, entonces $\overline{MLP(\rho, d, 2)} \subsetneq C(K)$ por lo que existe $f \in C(K) \setminus \overline{MLP(\rho, d, 2)}$. Por el Corolario 3.1 existe una función $h \in C(K)^\bullet$ tal que:

- $h \neq 0$,
- $h|_{\overline{MLP(\rho, d, 2)}} = 0$.

Sea $T_{\hat{a},b} : \mathbb{R}^d \rightarrow \mathbb{R}$, dada por $T_{\hat{a},b} = \langle \hat{a}, \hat{x} \rangle - b$, en donde $a \in \mathbb{R}^d$, $b \in \mathbb{R}$.

Se tiene que

$$\rho \circ T_{\hat{a},b} = Id \circ \rho \circ T_{\hat{a},b} \in MLP(\rho, d, 2),$$

en donde $Id : \mathbb{R} \rightarrow \mathbb{R}$ es la función identidad.

Como $h|_{\overline{MLP(\rho, d, 2)}} = 0$, concluimos que

$$h(\rho \circ T_{\hat{a},b}) = 0, \forall \hat{a} \in \mathbb{R}^d, \forall b \in \mathbb{R}.$$

Notemos que

$$\rho \circ T_{\hat{a},b}(\hat{x}) = \rho(\langle \hat{a}, \hat{x} \rangle - b) = t_{\hat{a},b}(\rho)(\hat{x}),$$

así,

$$h(t_{\hat{a},b}(\rho)) = 0, \forall \hat{a} \in \mathbb{R}^d, \forall b \in \mathbb{R}.$$

Como supusimos que ρ es discriminante, esto implica que $h = 0$.

Esto es una contradicción pues $h \neq 0$. La contradicción se obtiene por suponer que $MLP(\rho, d, 2)$ no es denso por lo que concluimos que $MLP(\rho, d, 2)$ es denso en $C(K)$. \square

Definición 4.4: Sigmoidal

Una función $f : \mathbb{R} \rightarrow \mathbb{R}$ continua tal que:

$$\lim_{r \rightarrow \infty} f(r) = 1, \lim_{r \rightarrow -\infty} f(r) = 0,$$

se llama sigmoidal.

Proposición 4.2

Sean $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compacto. Toda función sigmoideal $\rho : \mathbb{R} \rightarrow \mathbb{R}$ es discriminante (y por lo tanto $\overline{MLP(\rho, d, 2)}$ es denso en $C(K)$, por el Teorema 4.1).

Demostración. Sea ρ sigmoideal, entonces dados $\lambda, b, \theta \in \mathbb{R}$ y $\hat{a}, \hat{x} \in \mathbb{R}^d$, tenemos que:

$$\lim_{\lambda \rightarrow \infty} \rho(\lambda(\langle \hat{a}, \hat{x} \rangle - b) + \theta) = \begin{cases} 1 & \text{si } \langle \hat{a}, \hat{x} \rangle - b > 0, \\ \rho(\theta) & \text{si } \langle \hat{a}, \hat{x} \rangle - b = 0, \\ 0 & \text{si } \langle \hat{a}, \hat{x} \rangle - b < 0. \end{cases}$$

Sea $h \in C(K)^\bullet$ y sean:

$$H_{\hat{a}, b, >} := \{\hat{x} \in K : \langle \hat{a}, \hat{x} \rangle - b > 0\},$$

$$H_{\hat{a}, b, =} := \{\hat{x} \in K : \langle \hat{a}, \hat{x} \rangle - b = 0\},$$

$$H_{\hat{a}, b, <} := \{\hat{x} \in K : \langle \hat{a}, \hat{x} \rangle - b < 0\}.$$

Sea $g_{\lambda, \hat{a}, b, \theta}(\hat{x}) = \rho(\lambda(\langle \hat{a}, \hat{x} \rangle - b) + \theta)$, entonces

$$\lim_{\lambda \rightarrow \infty} g_{\lambda, \hat{a}, b, \theta}(\hat{x}) = X_{H_{\hat{a}, b, >}}(\hat{x}) + \rho(\theta)X_{H_{\hat{a}, b, =}}(\hat{x}).$$

Ahora veamos que ρ es discriminadora, es decir, si $h \in C(K)^\bullet$ es tal que $h(t_{\hat{a}, b}(\rho)) = 0$ para cualesquiera $\hat{a} \in \mathbb{R}^d$ y $b \in \mathbb{R}$, únicamente tenemos que ver $h = 0$.

Notemos que:

$$g_{\lambda, \hat{a}, b, \theta} = t_{\lambda \hat{a}, \lambda b - \theta}(\rho),$$

entonces

$$h(g_{\lambda, \hat{a}, b, \theta}) = 0 \quad \forall \lambda, \forall \hat{a}, \forall b, \forall \theta.$$

Por el teorema de convergencia dominada de Lebesgue 2, obtenemos que

$$\begin{aligned} 0 &= \lim_{\lambda \rightarrow \infty} h(g_{\lambda, \hat{a}, b, \theta}) = \int \lim_{\lambda \rightarrow \infty} g_{\lambda, \hat{a}, b, \theta} dh \\ &= \int (\chi_{H_{\hat{a}, b, >}} + \rho(\theta)\chi_{H_{\hat{a}, b, =}}) dh \\ &= \int \chi_{H_{\hat{a}, b, >}} dh + \rho(\theta) \int \chi_{H_{\hat{a}, b, =}} dh. \end{aligned}$$

Así,

$$\int \chi_{H_{\hat{a}, b, >}} dh + \rho(\theta) \int \chi_{H_{\hat{a}, b, =}} dh = 0 \quad \forall \hat{a}, \forall b, \forall \theta.$$

1. Haciendo θ tender a ∞ obtenemos

$$\int \chi_{H_{\hat{a}, b, >}} dh + \int \chi_{H_{\hat{a}, b, =}} dh = 0.$$

2. Haciendo θ tender a $-\infty$ obtenemos

$$\int \chi_{H\hat{a},b,>} dh = 0.$$

Sea $f_{\hat{a}}(\hat{x}) := \langle \hat{a}, \hat{x} \rangle \forall \hat{x} \in K$.

$$\begin{aligned} H_{\hat{a},b,>} &= \{\hat{x} \in K : \langle \hat{a}, \hat{x} \rangle - b > 0\} = f_{\hat{a}}^{-1}((b, \infty)) \\ &\Rightarrow H_{\hat{a},b,=} = f_{\hat{a}}^{-1}(b) \\ &\Rightarrow \chi_{H\hat{a},b,>} + \chi_{H\hat{a},b,=} = \chi_{f_{\hat{a}}^{-1}([b, \infty))}. \end{aligned}$$

Dados $b_1, b_2 \in \mathbb{R}$ con $b_1 < b_2$, obtenemos que

$$\chi_{H\hat{a},b_1,>} + \chi_{H\hat{a},b,=} - \chi_{H\hat{a},b_2,>} = \chi_{f_{\hat{a}}^{-1}([b_1, b_2])}.$$

De (1) y (2) obtenemos que

$$\int \chi_{f_{\hat{a}}^{-1}([b_1, b_2])} dh = 0.$$

Notemos que

$$\chi_{f_{\hat{a}}^{-1}([b_1, b_2])} = \chi_{[b_1, b_2]} \circ f_{\hat{a}},$$

así, para cualesquiera $b_1, b_2 \in \mathbb{R}$ con $b_1 < b_2$ y $\hat{a} \in \mathbb{R}^d$,

$$\int \chi_{[b_1, b_2]} \circ f_{\hat{a}} dh = 0.$$

Usamos el Teorema 4.3 para concluir que $h=0$. Por lo tanto ρ es discriminante. \square

4.2. Redes Neuronales

Recordemos que las transformaciones afines consisten de transformaciones lineales compuestas con traslaciones. Dada una transformación afín T , existe una matriz A y vector \hat{b} tales que

$$T(\hat{x}) = A\hat{x} + \hat{b};$$

de manera que identificamos

$$T \equiv (A, \hat{b}).$$

Dada un red neuronal

$$F = T_L \circ \rho \circ T_{L-1} \circ \rho \circ \cdots \circ \rho \circ T_1$$

en donde T_1, T_2, \dots, T_L son transformaciones afines y ρ una función de activación, identificamos

$$F_{\rho} \equiv (T_1, T_2, \dots, T_L);$$

en caso de no ser necesario resaltar la función de activación, simplemente identificamos

$$F \equiv (T_1, T_2, \dots, T_L).$$

Convengamos de ahora en adelante que a (T_1, T_2, \dots, T_L) se le nombra red neuronal o neural network (NN) y F es su realización. Así, dada una red neuronal

$$\phi = (T_1, T_2, \dots, T_L),$$

para denotar su realización utilizamos

$$F_\rho = R_\rho(\phi);$$

en caso de no ser necesario resaltar la función de activación, simplemente denotamos

$$F = R(\phi).$$

Otra notación que se usa es:

- $x_0 = \hat{x},$
- $x_1 = \rho \circ T_1(x_0),$
- \vdots
- $x_p = \rho \circ T_p \circ x_{p-1},$
- \vdots
- $x_L = T_L \circ x_{L-1}.$

Definición 4.5

Sea $\phi = (T_1, T_2, \dots, T_L)$ una red neuronal donde $T_p : \mathbb{R}^{N_{p-1}} \rightarrow \mathbb{R}^{N_p}$ y $N_0 = d$.

Definimos los siguientes conceptos:

- La dimensión de entrada de ϕ es d .
- El número de niveles o capas de ϕ es L , también denotado por $L(\phi) = L$.
- El número de neuronas está dado por $d + \sum_{j=1}^L N_j$.
- Dada una matriz A y un vector \hat{b} , denotamos por:

$$\begin{aligned} \|A\|_0 &\text{ al número de entradas no cero de } A, \\ \|\hat{b}\|_0 &\text{ al número de entradas no cero de } \hat{b}. \end{aligned}$$

- Dada una transformación afín $T \equiv (A, \hat{b})$, definimos $\|T\|_0 := \|A\|_0 + \|\hat{b}\|_0$.
- Denotamos por:

$$M(\phi) = \sum_{j=1}^L M_j(\phi)$$

donde

$$M_j(\phi) = \|T_j\|_0.$$

4.3. Operaciones Básicas con Redes Neuronales

Definición 4.6: Concatenación

Sean ϕ_1, ϕ_2 redes neuronales tales que

$$\begin{aligned} \phi_1 &= (T_1, T_2, \dots, T_L), \\ \phi_2 &= (V_1, V_2, \dots, V_M). \end{aligned}$$

Definimos la concatenación de ϕ_1 y ϕ_2 como

$$\phi_1 \bullet \phi_2 := (V_1, V_2, \dots, V_{M-1}, T_1 \circ V_M, T_2, \dots, T_L).$$

La cual es una red neuronal de $L + M - 1$ niveles.

Es directo ver que

$$R(\phi_1 \bullet \phi_2) = \phi_1 \circ \phi_2.$$

Definición 4.7

Dados T y V transformaciones afines, definimos las siguientes transformaciones afines:

- $\begin{pmatrix} T \\ V \end{pmatrix}(\hat{x}) = \begin{pmatrix} T(\hat{x}) \\ V(\hat{x}) \end{pmatrix}$

donde T y V tienen el mismo dominio.

- $T \oplus V \begin{pmatrix} \hat{x} \\ \hat{y} \end{pmatrix} = \begin{pmatrix} T(\hat{x}) \\ V(\hat{y}) \end{pmatrix}$

donde T y V no necesariamente tienen el mismo dominio.

Definición 4.8

Sean ϕ_1, ϕ_2 redes neuronales tales que

$$\phi_1 = (T_1, T_2, \dots, T_L),$$

$$\phi_2 = (V_1, V_2, \dots, V_L).$$

- Definimos la paralelización con entrada compartida de ϕ^1 y ϕ^2 como

$$P(\phi_1, \phi_2) = \left(\begin{pmatrix} T_1 \\ V_1 \end{pmatrix}, T_2 \oplus V_2, \dots, T_L \oplus V_L \right)$$

donde las dimensiones de entrada de ϕ^1 y ϕ^2 son la misma.

- También definimos a la paralelización con entradas no compartidas de ϕ^1 y ϕ^2 como

$$FP(\phi^1, \phi^2) = (T_1 \oplus V_1, T_2 \oplus V_2, \dots, T_L \oplus V_L)$$

donde las dimensiones de entrada de ϕ^1 y ϕ^2 no necesariamente son la misma.

Proposición 4.3

1. $M(P(\phi_1, \phi_2)) = M(FP(\phi_1, \phi_2)) = M(\phi_1) + M(\phi_2).$

2. $R(P(\phi_1, \phi_2))(\hat{x}) = \begin{pmatrix} R(\phi_1(\hat{x})) \\ R(\phi_2(\hat{x})) \end{pmatrix}.$

Proposición 4.4

Sean $d \in \mathbb{N}$ y $K \subseteq \mathbb{R}^d$ compacto. Supongamos que $\rho : \mathbb{R} \rightarrow \mathbb{R}$ es no constante y diferenciable en un abierto, entonces, para todo $\varepsilon > 0$, existe una red neuronal ϕ tal que

$$\phi = (T_1, T_2)$$

donde $T_1, T_2 : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $M(\phi) \leq 4d$ y $|R(\phi)(\hat{x}) - \hat{x}| < \varepsilon$ para todo $\hat{x} \in K$.

Demostración. Sean $d = 1$, $x^\bullet \in \mathbb{R}$ tal que ρ es diferenciable en una vecindad alrededor de x^\bullet y supongamos que

$$\rho'(x^\bullet) = \theta \neq 0.$$

Dada $\lambda > 0$ definimos:

$$\begin{aligned} T_1(\hat{x}) &= \frac{1}{\lambda} \hat{x} + x^\bullet, \\ T_2(\hat{x}) &= \frac{\lambda}{\theta} \hat{x} - \lambda \rho\left(\frac{\hat{x}}{\lambda} + x^\bullet\right). \end{aligned}$$

Se tiene que

$$\underbrace{|R(\phi)(\hat{x}) - \hat{x}|}_{T_2(\rho \circ T_1(\hat{x}))} = \left| \frac{\lambda}{\theta} \left(\overbrace{\rho\left(\frac{\hat{x}}{\lambda} + x^\bullet\right)}^{T_1(\hat{x})} - \rho(x^\bullet) \right) - \hat{x} \right|.$$

Es claro que $|R(\phi)(0) - 0| = 0$ y para $\hat{x} \neq 0$, se tiene que

$$\begin{aligned} |R(\phi)(\hat{x}) - \hat{x}| &= \frac{|\hat{x}|}{|\theta|} \left| \frac{\rho\left(\frac{\hat{x}}{\lambda} + x^\bullet\right) - \rho(x^\bullet) - \theta}{\frac{\hat{x}}{\lambda}} \right| \\ &= \frac{|\hat{x}|}{\theta} \left| \frac{\rho(x^\bullet + h_\lambda) - \rho(x^\bullet)}{h_\lambda} - \rho'(x^\bullet) \right|. \end{aligned}$$

Así,

$$|R(\phi)(\hat{x}) - \hat{x}| = \frac{|\hat{x}|}{|\theta|} \left| \frac{\rho(x^\bullet + h_\lambda) - \rho(x^\bullet)}{h_\lambda} - \rho'(x^\bullet) \right|.$$

Sea $\alpha = \max\{|\hat{x}| : \hat{x} \in K\}$. Como ρ es diferenciable en un intervalo y que contiene a x^\bullet entonces dada $\varepsilon > 0$ existe $\delta > 0$ tal que si $|y| < \delta$, entonces

$$\left| \frac{\rho(x^\bullet + y) - \rho(x^\bullet)}{y} - \rho'(x^\bullet) \right| < \frac{|\theta|}{M} \varepsilon.$$

Como $h_\lambda = \frac{\hat{x}}{\lambda}$ tomamos λ tal que $\frac{\alpha}{\lambda} < \delta$, así,

$$|h_\lambda| = \frac{|\hat{x}|}{\lambda} < \delta.$$

En este caso obtenemos que, para todo $\hat{x} \in K$,

$$|R(\phi)(\hat{x}) - \hat{x}| = \frac{|\hat{x}|}{|\theta|} \left| \frac{\rho(x^\bullet + y) - \rho(x^\bullet)}{y} - \rho'(x^\bullet) \right| < \frac{|\theta|}{M} \frac{|\hat{x}|}{|\theta|} \varepsilon \leq \varepsilon;$$

de lo que concluimos que

$$|R(\phi)(\hat{x}) - \hat{x}| \leq \varepsilon \quad \forall \hat{x} \in K.$$

□

Teorema 4.2: Universalidad, Caso $L \geq 2$

Sean $d \in \mathbb{N}$, $K \subset \mathbb{R}^d$ compacto y $\rho : \mathbb{R} \rightarrow \mathbb{R}$ una función discriminante tal que es diferenciable y no constante en un abierto. Entonces $MLP(\rho, d, L)$ es universal para $L \in \mathbb{N}$ con $L \geq 2$, es decir, $MLP(\rho, d, L)$ es denso en $C(K)$.

Demostración. Considere $f \in C(K)$ y supongamos que $(\phi_n)_{n \in \mathbb{N}}$ es una sucesión de redes neuronales tal que

$$\lim_{n \rightarrow \infty} \phi_n = f$$

donde $\phi_n \in MLP(\rho, d, L)$ para todo $n \in \mathbb{N}$.

Por el teorema de universalidad para el caso $L = 2$, existe una red neuronal φ tal que $R(\varphi^n) \in MLP(\rho, d, 2)$ satisface que

$$\|R(\varphi^n) - f\| < \frac{1}{n} \quad \forall n \in \mathbb{N}. \quad (1)$$

Como f es continua, entonces $f(K)$ es compacto (cerrado y acotado).

Sea $\tilde{K} = \overline{B(0, r)} \subseteq \mathbb{R}$ tal que $r > L$ y $f(K) \subseteq B(0, r - L)$.

Usamos la proposición 4.4 para encontrar redes neuronales ψ^n tales que la profundidad de $R(\psi^n)$ es 2 y

$$|R(\psi^n)(y) - y| < \frac{1}{n} \quad \forall y \in \tilde{K}. \quad (2)$$

Definimos

$$\phi_n = \underbrace{\psi^n \bullet \psi^n \bullet \dots \bullet \psi^n}_{L-2 \text{ veces}} \bullet \varphi^n$$

donde ϕ_n tiene profundidad L .

Recordemos que

$$R(\phi_n) = R(\psi^n) \circ R(\psi^n) \circ \cdots \circ R(\psi^n) \circ R(\varphi^n).$$

Así,

$$\begin{aligned} |R(\phi_n)(\hat{x}) - f(\hat{x})| &\leq |R(\psi^n)[R(\psi^n) \circ \cdots \circ R(\varphi^n)(\hat{x})] - \overbrace{R(\psi^n) \circ \cdots \circ R(\varphi^n)(\hat{x})}^{n-1 \text{ veces}}| \\ &\quad + |R(\psi^n)[\overbrace{R(\psi^n) \circ \cdots \circ R(\varphi^n)(\hat{x})}^{n-2 \text{ veces}}] - \overbrace{R(\psi^n) \circ \cdots \circ R(\varphi^n)(\hat{x})}^{n-2 \text{ veces}}| \\ &\quad + \cdots + |R(\phi^n)[R(\varphi^n)(\hat{x})] - R(\varphi^n)(\hat{x})| + |R(\varphi^n)(\hat{x}) - f(\hat{x})|. \end{aligned}$$

Luego, por la ecuación 2,

$$|R(\phi_n)(\hat{x}) - f(\hat{x})| < \overbrace{\frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n}}^{L-2 \text{ veces}} + |R(\varphi^n)(\hat{x}) - f(\hat{x})|$$

y por la ecuación 1

$$|R(\phi_n)(\hat{x}) - f(\hat{x})| < (L-1) \frac{1}{n}.$$

De lo que concluimos que

$$\lim_{n \rightarrow \infty} R(\phi_n) = f.$$

□

Esto último demuestra el teorema, sin embargo, para poder emplear la ecuación 2 usamos que

$$\underbrace{R(\psi^n) \circ \cdots \circ R(\varphi^n)(\hat{x})}_{m-\text{veces}} \in \tilde{K}, \quad \forall \hat{x} \in K, \quad \forall m \leq L-2.$$

Veamos porqué pasa esto:

Notemos que por la ecuación 1

$$|R(\varphi^n)(\hat{x}) - f(\hat{x})| < \frac{1}{n},$$

por lo que la distancia

$$\text{dist}(R(\varphi^n)(x), F(K)) \leq \frac{1}{n}.$$

Luego, por la ecuación 2,

$$|R(\psi^n) \circ R(\varphi^n)(\hat{x}) - R(\varphi^n)(\hat{x})| < \frac{1}{n},$$

por lo que

$$\text{dist}(R(\psi^n) \circ \varphi^n(\hat{x}), (R \circ \varphi^n)(K)) \leq \frac{1}{n}.$$

Así,

$$\text{dist}(R(\psi^n) \circ R(\varphi^n)(\hat{x}), F(K)) \leq \frac{2}{n}.$$

De esta forma obtenemos que

$$\text{dist}[R(\psi^n) \circ \dots \circ R(\varphi^n)(\hat{x}), f(K)] \leq \frac{L}{n} \quad \forall \hat{x} \in K.$$

Esto implica que $R(\psi^n) \circ \dots \circ R(\varphi^n)(\hat{x}) \in \tilde{K}$ para todo $\hat{x} \in K$.

4.4. Reaproximación de Dicionarios

Definición 4.9

Sean H un espacio normado y $(A_n)_{n \in \mathbb{N}}$ una colección anidada de subconjuntos, es decir, $A_n \subseteq A_{n+1}$ para todo $n \in \mathbb{N}$. Dado $C \subseteq H$, definimos

$$\sigma(A_n, C) := \sup_{f \in C} \inf_{g \in A_n} \|f \cdot g\|.$$

Sea $h : \mathbb{N} \rightarrow \mathbb{R}^+$. Si se cumple que, cuando $n \rightarrow \infty$

$$\sigma(A_n, C) = \mathcal{O}(h(n))$$

entonces decimos que $(A_n)_{n \in \mathbb{N}}$ tiene una tasa de aproximación h para C .

Recordemos que la expresión $\sigma(A_n, C) = \mathcal{O}(h(n))$ significa que existe una constante α tal que

$$\sigma(A_n, C) \leq \alpha \cdot h(n) \quad \forall n.$$

Definición 4.10

Dadas funciones f y g , decimos que

$$f \lesssim g$$

si existe una constante α tal que $f \leq \alpha \cdot g$ puntualmente.

De lo anterior tenemos que

$$\sigma(A_n, C) = \mathcal{O}(h(n)) \Leftrightarrow \sigma(A_n, C) \lesssim h(n).$$

Definición 4.11: Diccionario

Sea $D = (f_i)_{i=1}^{\infty} \subseteq H$ una sucesión (a la cual llamaremos Diccionario).

Definimos los espacios

$$A_N := \left\{ \sum_{i=1}^{\infty} C_i f_i : C_i \in \mathbb{R} \text{ y } \|C\|_0 \leq N \right\},$$

en donde $\|C\|_0 = \#\{i \in \mathbb{N} : C_i \neq 0\}$.

Es claro que $A_N \subseteq A_{N+1}$ para todo N .

Teorema 4.3

Sean $d \in \mathbb{N}$, $H \subseteq \{f : K \subseteq \mathbb{R}^d \rightarrow \mathbb{R}\}$ un espacio normado, $\rho : \mathbb{R} \rightarrow \mathbb{R}$ y $D := (f_i)_{i=1}^{\infty} \subseteq H$ un diccionario.

Supongamos que existen $L, C \in \mathbb{N}$ tales que, para todo $i \in \mathbb{N}$ y para todo $\varepsilon > 0$, existe una red neuronal ϕ_i^ε tal que:

$$L(\phi_i^\varepsilon) = L,$$

$$M(\phi_i^\varepsilon) \leq C,$$

$$\|R(\phi_i^\varepsilon) - f_i\|_H \leq \varepsilon.$$

Para todo $C \subseteq H$, definimos A_N como en la definición 4.11 y definimos

$$B_N := \{R(\phi) : \phi \text{ es una red neuronal, } L(\phi) = L, M(\phi) \leq N\},$$

entonces, para todo $\zeta \subseteq H$,

$$\sigma(B_N, \zeta) \leq \sigma(A_N, \zeta).$$

Demostración. Sea $a \in A_N$, entonces

$$a = \sum_{j=1}^n C_{i(j)} f_{i(j)}.$$

Sea $\varepsilon > 0$, entonces por hipótesis existen redes neuronales $(\phi)_{j=1}^n$ tales que:

- $L(\phi_j) = L$,
- $M(\phi_j) \leq C$,
- $\|R(\phi_j) - f_{i(j)}\| \leq \frac{\varepsilon}{N \cdot \max(\{|C_{i(j)}|\})}$.

Definimos

$$\phi^{(a, \varepsilon)} = \phi^c \cdot P[\phi_1, \phi_2, \dots, \phi_N],$$

en donde $\phi^c = ([C_{i(1)}, C_{i(2)}, \dots, C_{i(N)}], 0)$, entonces

$$R(\phi^{(a, \varepsilon)}) = \sum_{j=1}^n C_{i(j)} R(\phi_j).$$

Por lo que

$$\begin{aligned}
 \|R(\phi^{(a,\varepsilon)}) - a\| &= \left\| \sum_{j=1}^n C_{i(j)} R(\phi_j) - \sum_{j=1}^n C_{i(j)} f_{i(j)} \right\| \\
 &\leq \sum_{j=1}^n |C_{i(j)}| \|R(\phi_j) - f_{i(j)}\| \\
 &\leq N \cdot \max(\{|C_{i(j)}|\}) \frac{\varepsilon}{N \cdot \max(\{|C_{i(j)}|\})} \\
 &\leq \varepsilon.
 \end{aligned}$$

Así, para cualesquiera $\varepsilon > 0$ y $a \in A_N$, existe $\phi^{(a,\varepsilon)}$ tal que:

- $L(\phi^{(a,\varepsilon)}) = L$,
- $M(\phi^{(a,\varepsilon)}) \leq C \cdot N$,
- $\|R(\phi^{(a,\varepsilon)}) - a\| \leq \varepsilon$.

En otras palabras, para cualesquiera $\varepsilon > 0$ y $a \in A_N$, existe $R(\phi^{(a,\varepsilon)}) \in B_{CN}$ tal que

$$\|R(\phi^{(a,\varepsilon)}) - a\| \leq \varepsilon. \quad (1)$$

Recordemos que:

$$\begin{aligned}
 \sigma(A_N, \zeta) &= \sup_{f \in \zeta} \inf_{a \in A_N} \|f - a\|, \\
 \sigma(B_{CN}, \zeta) &= \sup_{f \in \zeta} \inf_{b \in B_{CN}} \|f - b\|.
 \end{aligned}$$

Sean $f \in \zeta$ y $a \in A_N$. Por la ecuación 1 encontramos $b_\varepsilon \in B_{CN}$ tal que $\|a - b_\varepsilon\| \leq \varepsilon$. Así, para todo $a \in A_N$,

$$\begin{aligned}
 \inf_{b \in B_{CN}} \|f - b\| &\leq \|f - b_\varepsilon\| = \|f - b_\varepsilon + a - a\| \\
 &\leq \|f - a\| + \|a - b_\varepsilon\| \\
 &\leq \|f - a\| + \varepsilon.
 \end{aligned}$$

Por lo que

$$\inf_{b \in B_{CN}} \|f - b\| \leq \inf_{a \in A_N} \|f - a\| + \varepsilon \quad \forall \varepsilon > 0.$$

Luego, haciendo tender $\varepsilon \rightarrow 0$, obtenemos

$$\inf_{b \in B_{CN}} \|f - b\| \leq \inf_{a \in A_N} \|f - a\| \quad \forall f \in \zeta.$$

Tomando el supremo en la desigualdad, obtenemos que

$$\sigma(B_{CN}, \zeta) = \sup_{f \in \zeta} \inf_{b \in B_{CN}} \|f - b\| \leq \sup_{f \in \zeta} \inf_{a \in A_N} \|f - a\| = \sigma(A_N, \zeta),$$

de lo que concluimos que

$$\sigma(B_{CN}, \zeta) \leq \sigma(A_N, \zeta).$$

□

4.5. Aproximación de Funciones Suaves

Definición 4.12: B-Spline

Definimos el B-Spline de orden $k \in \mathbb{N}$ de la siguiente manera:

$$\mathcal{N}_k(x) := \frac{1}{(k-1)!} \sum_{s=0}^k (-1)^s \binom{k}{s} (x-s)_+^{k-1},$$

en donde $x \in \mathbb{R}$, $0^0 = 0$, $0! = 1$ y $t_+ = \begin{cases} y & \text{si } y \geq 0, \\ 0 & \text{si } y \leq 0. \end{cases}$

Definición 4.13: B-Spline Multivariado

Dados $t \in \mathbb{R}$ y $l \in \mathbb{N}$, definimos

$$\mathcal{N}_{l,t,k}(x) := \mathcal{N}_k(2^l(x-t)),$$

donde $x \in \mathbb{R}$.

Dados $d, l \in \mathbb{N}$ y $t = (t_1, t_2, \dots, t_d) \in \mathbb{R}^d$, definimos el B-Spline multivariado como

$$\mathcal{N}_{l,t,k}^d(\hat{x}) := \prod_{j=1}^d \mathcal{N}_{l,t_j,k}(x_j),$$

donde $\hat{x} = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$.

Definición 4.14

Definimos

$$\mathcal{B}^k := \left\{ \mathcal{N}_{l,t_l,k}^d : l \in \mathbb{N}, t_l \in 2^{-l}\mathbb{Z}^d \right\}.$$

Teorema 4.4

Sean $d, k \in \mathbb{N}$ y $s \in (0, k)$. Entonces existe $c > 0$ tal que para toda $f \in C^s([0, 1]^d)$ tenemos que, para toda $\delta > 0$ y todo $N \in \mathbb{N}$, existen $C_i \in \mathbb{R}$ y $B_i \in \mathcal{B}^k$ tales que:

- $|C_i| \leq C\|f\|$,
- $f - \sum_{i=1}^N C_i B_i \lesssim N^{\frac{\delta-s}{d}} \|f\|_{C^s([0,1]^d)}$,

en donde $i \in \{1, 2, \dots, N\}$ y

$$\|f\|_{C^s([0,1]^d)} := \sum_{J_i + \dots + J_d \leq s} \left\| \frac{\partial^{J_i}}{\partial x_i} \cdots \frac{\partial^{J_d}}{\partial x_d} f \right\|.$$

Definición 4.15: Sigmoidal

Una función $\rho : \mathbb{R} \rightarrow \mathbb{R}$ es llamada sigmoidal de orden $q \in \mathbb{N}$ si cumple:

1. $\rho \in C^{q-1}(\mathbb{R})$,
2. $\frac{\rho(x)}{x^q} \rightarrow 0$ como $x \rightarrow -\infty$,
3. $\frac{\rho(x)}{x^q} \rightarrow 1$ como $x \rightarrow \infty$,
4. $|\rho(x)| \lesssim (1 + |x|)^q \quad \forall x \in \mathbb{R}$.

Definición 4.16: Función Sigmoidal Estandar

El ejemplo estándar de función sigmoidal de orden q es $\rho_s : \mathbb{R} \rightarrow \mathbb{R}$ tal que

$$\rho_s(x) = x_+^q.$$

Frecuentemente usaremos la función de activación $\rho_s = x_+^q$ para facilitar las demostraciones, aunque éstas son válidas para cualquier función sigmoidal de orden q .

Lema 4.1

Dados $\varepsilon > 0$ y $p \in \mathbb{N}$, existe una red neuronal ϕ tal que:

- $L(\phi) = \lceil \max\{\log_q(p-1), 0\} \rceil + 1$,
- $\|R(\phi) - x_+^{p-1}\|_{[-z, z]} \leq \varepsilon$,

para cualquier intervalo de la forma $[-z, z]$ donde ϕ depende de z .

Demostración. Sea $\lambda := \lceil \max(\log_q(p-1), 0) \rceil$, se tiene que

$$\underbrace{\rho_s \circ \rho_s \circ \cdots \circ \rho_s(x)}_{\lambda\text{-veces}} = x_+^{q^\lambda} = x_+^b,$$

donde $b = q^\lambda$.

Así,

$$b = q^\lambda \geq q^{\log_q(p-1)} \geq p-1.$$

Denotamos por $g(x) = x_+^b$. Luego

$$\rho_s \circ \rho_s \circ \cdots \circ \rho_s = g.$$

Tomamos, dada $\delta \in \mathbb{R}$,

- $T_1^\delta(x) = \begin{pmatrix} x + \delta \\ x \end{pmatrix} \quad \forall x \in \mathbb{R},$
- $T_l^\delta \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix} \quad \forall x, y \in \mathbb{R}, \quad \forall l \in \{2, 3, \dots, \lambda-1\},$
- $T_{\lambda+1}^\delta \begin{pmatrix} x \\ y \end{pmatrix} = \frac{x-y}{\delta} \quad \forall x, y \in \mathbb{R}.$

Sea $\phi^\delta = (T_1^\delta, T_2^\delta, \dots, T_{\lambda+1}^\delta)$. Tenemos que

$$\begin{aligned} R(\phi^\delta)(x) &= T_{\lambda+1}^\delta \circ \rho_s \circ T_{\lambda-1}^\delta \circ \rho_s \circ \cdots \circ \rho_s \circ T_1(x) \\ &= T_{\lambda+1}^\delta \circ \rho_s \circ \cdots \circ T_2 \begin{pmatrix} \rho_s(x + \delta) \\ \rho_s(x) \end{pmatrix} \\ &= T_{\lambda+1}^\delta \begin{pmatrix} \rho_s \circ \rho_s \circ \cdots \circ \rho_s(x + \delta) \\ \rho_s \circ \rho_s \circ \cdots \circ \rho_s(x) \end{pmatrix} \\ &= T_{\lambda+1}^\delta \begin{pmatrix} g(x + \delta) \\ g(x) \end{pmatrix} \\ &= \frac{g(x + \delta) - g(x)}{\delta}, \end{aligned}$$

esto implica que, cuando $\delta \rightarrow 0$, $|R(\phi^\delta)(x) - g'(x)| \rightarrow 0$ uniformemente en $[-z, z]$.

Concluimos que dado $\varepsilon > 0$ existe una red neuronal ϕ^δ tal que

$$\|R(\phi^\delta) - g'\|_{[-z, z]} < \varepsilon \quad \text{y} \quad L(\phi^\delta) = \lambda + 1.$$

El siguiente paso es aproximar $g''(x) = b(b-1)x_+^{b-2}$ usando redes neuronales.

Tomamos $h \in (0, 1)$ suficientemente chico tal que

$$\left| \frac{g(x+h) - g(x) - h \cdot g'(x) - \frac{h^2}{2} g''(x)}{h^2} \right| < \varepsilon \quad \forall x \in [-z, z],$$

(usamos el Teorema de Taylor: $g(x+h) = g(x) + h \cdot g'(x) + \frac{h^2}{2} g''(x) + \mathcal{O}(h^3)$).

Encontramos $\delta > 0$ tal que, de manera uniforme,

$$|R(\phi^\delta)(x) - g'(x)| < h^2 \cdot \varepsilon.$$

Así, de manera uniforme,

$$\begin{aligned} & \left| \frac{g(x+h) - g(x) - h \cdot R(\phi^d)(x) - \frac{h^2}{2} g''(x)}{h^2} \right| \\ & \leq \left| \frac{g'(x) - R(\phi^d)(x)}{h} \right| + \left| \frac{g(x+h) - g(x) - h \cdot g'(x) - \frac{h^2}{2} g''(x)}{h^2} \right| \\ & < 2 \cdot \varepsilon. \end{aligned} \tag{1}$$

Recordemos que

$$\phi^d = (T_1^\delta, \dots, T_{\lambda+1}^\delta),$$

tomemos ahora

$$\phi^{\varepsilon, \delta} = (T_1^{\varepsilon, \delta}, T_2^{\varepsilon, \delta}, \dots, T_{\lambda+1}^{\varepsilon, \delta}),$$

en donde:

- $T_1^{\varepsilon, \delta}(x) = \begin{pmatrix} x+h \\ x \\ T_1^\delta(x) \end{pmatrix},$
- $T_l^{\varepsilon, \delta} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ T_l^\delta(y_3) \end{pmatrix} \quad \forall l \in \{1, 2, \dots, \lambda\},$
- $T_{\lambda+1}^{\varepsilon, \delta} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \frac{y_1 - y_2 - h \cdot T_{\lambda+1}^\delta(y_3)}{h^2}.$

Se tiene que

$$\begin{aligned}
 R(\phi^{\varepsilon, \delta})(x) &= T_{\lambda+1} \circ \rho_s \circ \cdots \circ \rho_s \circ T_1(x) \\
 &= T_{\lambda+1}^{\varepsilon, \delta} \begin{pmatrix} \rho_s \circ \cdots \circ \rho_s(x+h) \\ \rho_s \circ \cdots \circ \rho_s(x) \\ \rho_s \circ T_1^\delta \circ \cdots \circ \rho_s \circ T_1^\delta(x) \end{pmatrix} \\
 &= T_{\lambda+1}^{\varepsilon, \delta} \begin{pmatrix} g(x+h) \\ g(x) \\ \rho_s \circ T_1^\delta \circ \cdots \circ \rho_s \circ T_1^\delta(x) \end{pmatrix} \\
 &= \frac{g(x+h) - g(x) - T_{\lambda+1}^\delta \circ \cdots \circ T_1^\delta(x)}{h^2} \\
 &= \frac{g(x+h) - g(x) - h \cdot R(\phi^\delta)}{h^2},
 \end{aligned}$$

así,

$$R(\phi^{\varepsilon, \delta})(x) = \frac{g(x+h) - g(x) - h \cdot R(\phi^\delta)}{h^2}$$

y de la ecuación (1) obtenemos que

$$|R(\phi^{\varepsilon, \delta})(x) - \frac{1}{2}g''(x)| < 2 \cdot \varepsilon$$

Por lo tanto converge uniformemente.

De la misma forma se demuestra que para toda n y para toda $\epsilon > 0$ existe una constante $c \neq 0$ y una red neuronal ϕ con $L(\phi) = \lambda + 1$ tal que, de manera uniforme,

$$|R(\phi)(x) - c \cdot \frac{d^n}{dx^n}g(x)| < \epsilon;$$

siempre y cuando se pueda aplicar el teorema de Taylor.

Sea m tal que

$$\frac{d^m}{dx^m}g(x) = \alpha \cdot x_+^{p-1}$$

en donde α es una constante. Por lo anterior, existe una red neuronal ψ con $L(\psi) = \lambda + 1$ tal que, de manera uniforme,

$$|R(\psi)(x) - c \cdot x_+^{p-1}| < |c| \cdot \epsilon$$

en donde c es una constante no cero.

Sean

$$\begin{aligned}
 \psi &= (\tilde{T}_1, \tilde{T}_2, \dots, \tilde{T}_{\lambda+1}), \\
 \phi &= (\tilde{T}_1, \tilde{T}_2, \dots, \tilde{T}_\lambda, \frac{1}{c} \cdot \tilde{T}_{\lambda+1}),
 \end{aligned}$$

se tiene que

$$R(\phi) = \frac{1}{c} R(\psi).$$

Así, de manera uniforme,

$$\begin{aligned} |R(\phi)(x) - x_+^{p-1}| &= \left| \frac{1}{c} R(\psi)(x) - x_+^{p-1} \right| \\ &= \frac{1}{|c|} |R(\psi)(x) - c \cdot x_+^{p-1}| \\ &< \frac{1}{|c|} \cdot |c| \cdot \epsilon = \epsilon. \end{aligned}$$

Hemos demostrado que, para toda $\epsilon > 0$, existe ϕ tal que, de manera uniforme,

$$|R(\phi)(x) - x_+^{p-1}| < \epsilon.$$

Lo que implica el resultado del lema 4.1. □

Corolario 4.1

Sea $t \in \mathbb{R}$, $p \in \mathbb{N}$, $z > 0$. Si que ρ es una función sigmoideal de orden q , entonces existe una red neuronal ϕ^ϵ tal que:

$$\begin{aligned} \|R(\phi) - \mathcal{N}_{l,t,\rho}\|_{[-z,z]} &< \epsilon, \\ L(\phi) &= \lceil \max\{\log_q(p-1), 0\} \rceil + 1. \end{aligned}$$

Demostración. Por el lema 4.1, dada $s > 0$ y $A > 0$, existe una red neuronal ϕ^s tal que:

$$\begin{aligned} \|R(\phi^s) - x_+^{p-1}\|_{[-A,A]} &< s, \\ L(\phi^s) &= \lceil \max\{\log_q(p-1), 0\} \rceil + 1 = L. \end{aligned}$$

Supongamos que

$$\phi^s = (T_1, T_2, \dots, T_L)$$

y sean, para toda $l \in \{0, 1, \dots, p\}$,

$$T_{1,l}(x) = T_1(2^l(x-t) - l).$$

Definimos las transformaciones afines:

$$\bullet \quad V_1(x) = \begin{pmatrix} T_{1,0}(x) \\ T_{1,1}(x) \\ \vdots \\ T_{1,p}(x) \end{pmatrix},$$

$$\begin{aligned}
\bullet \quad V_j \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} &= \begin{pmatrix} T_j(x_0) \\ T_j(x_1) \\ \vdots \\ T_j(x_p) \end{pmatrix} \quad \forall j \in \{2, 3, \dots, L-1\}, \\
\bullet \quad V_j \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} &= \frac{1}{(p-1)!} \sum_{l=0}^p (-1)^l \binom{p}{l} T_L(x_l).
\end{aligned}$$

Definimos también

$$\psi^s = (V_1, V_2, \dots, V_L),$$

de manera que

$$R(\psi^s) = \frac{1}{(p-1)!} \sum_{i=0}^p (-1)^i \binom{p}{i} R(\psi^s)(2^i(x-t) - i).$$

Obtenemos que

$$|R(\psi^s)(x) - \mathcal{N}_{i,t,p}(x)| = \left| \frac{1}{(p-1)!} \sum_{i=0}^p (-1)^i \binom{p}{i} \left[R(\psi^s)(2^i(x-t) - i) - (2^i(x-t) - i)_+^{p-1} \right] \right|,$$

entonces

$$\begin{aligned}
|R(\psi^s)(x) - \mathcal{N}_{i,t,p}(x)| &\leq \frac{1}{(p-1)!} \sum_i \binom{p}{i} |R(\psi^s)(2^i(x-t) - i) - (2^i(x-t) - i)_+^{p-1}| \\
&\leq \left[\frac{1}{(p-1)!} \sum_i \binom{p}{i} \right] \cdot s.
\end{aligned}$$

Tomamos

$$s = \frac{1}{\frac{1}{(p-1)!} \sum_i \binom{p}{i}} \cdot \epsilon$$

y elegimos A adecuadamente para que finalmente obtengamos

$$\|R(\psi^s) - \mathcal{N}_{i,t,p}(x)\|_{[-z,z]} < \epsilon.$$

□

Corolario 4.2: De la demostración del Lema 4.1

Sean $K \subset \mathbb{R}^2$ compacto y $\epsilon > 0$. Si ρ es una función sigmoideal de orden $q \geq 2$, entonces existe una red neuronal ϕ_ϵ con $L(\phi_\epsilon) = 2$ tal que

$$|\phi_\epsilon(x) - (x_1 + x_2)_+^2| < \epsilon \quad \forall x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in K$$

y existe otra red neuronal ψ_ϵ con $L(\psi_\epsilon) = 2$ tal que

$$|\psi_\epsilon(x) - x_1 x_2| < \epsilon \quad \forall x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in K.$$

Corolario 4.3

Dado $K \subset \mathbb{R}^d$ y $\epsilon > 0$, existe una red neuronal $\phi_\epsilon^{mult,d}$ tal que

$$|\phi_\epsilon^{mult,d}(x) - x_1 x_2 \cdots x_d| < \epsilon \quad \forall x \in K,$$

donde

$$L(\phi_\epsilon^{mult,d}) = \lceil \log_2(d) \rceil + 1.$$

Demostración. Haremos la demostración para el caso en el que $d = 2^\alpha$ para alguna $\alpha \in \mathbb{N}$. Por el corolario 4.2, para cada ϵ existe ψ_ϵ tal que, de manera uniforme,

$$\left| \psi_\epsilon \left(\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right) - y_1 y_2 \right| < \epsilon.$$

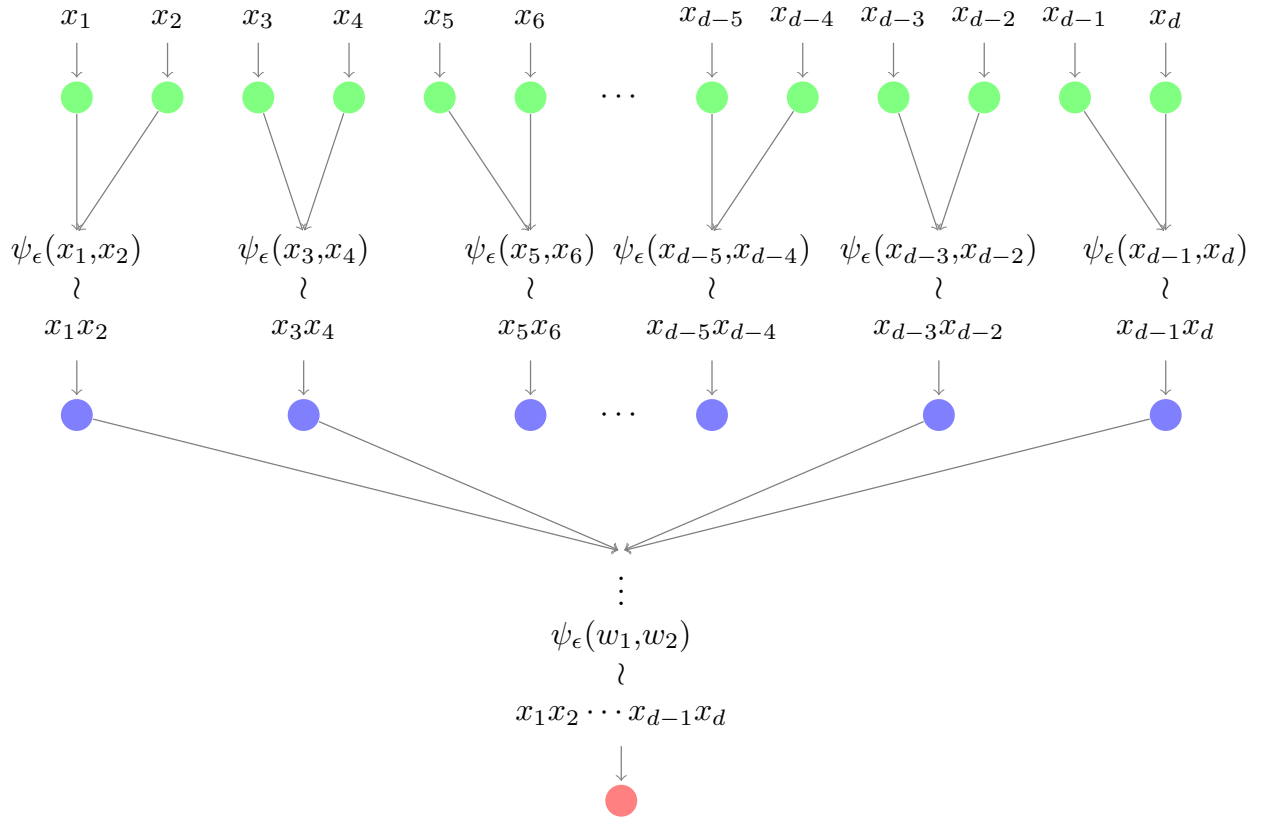
Luego, definimos

$$\phi_\epsilon^{mult,d} = \underbrace{FP(\psi_\epsilon, \psi_\epsilon)}_{2^{\alpha-(\alpha-1)} \text{ - veces}} \bullet FP(\psi_\epsilon, \psi_\epsilon, \psi_\epsilon, \psi_\epsilon) \bullet \cdots \bullet \underbrace{FP(\psi_\epsilon, \dots, \psi_\epsilon)}_{2^{\alpha-2} \text{ - veces}} \bullet \underbrace{FP(\psi_\epsilon, \psi_\epsilon, \dots, \psi_\epsilon)}_{2^{\alpha-1} \text{ - veces}}$$

en donde

$$FP(\psi_\epsilon, \psi_\epsilon, \dots, \psi_\epsilon) \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} = \begin{pmatrix} \psi_\epsilon(z_1) \\ \psi_\epsilon(z_2) \\ \vdots \\ \psi_\epsilon(z_n) \end{pmatrix}.$$

Visualicemos de forma gráfica lo que está sucediendo:



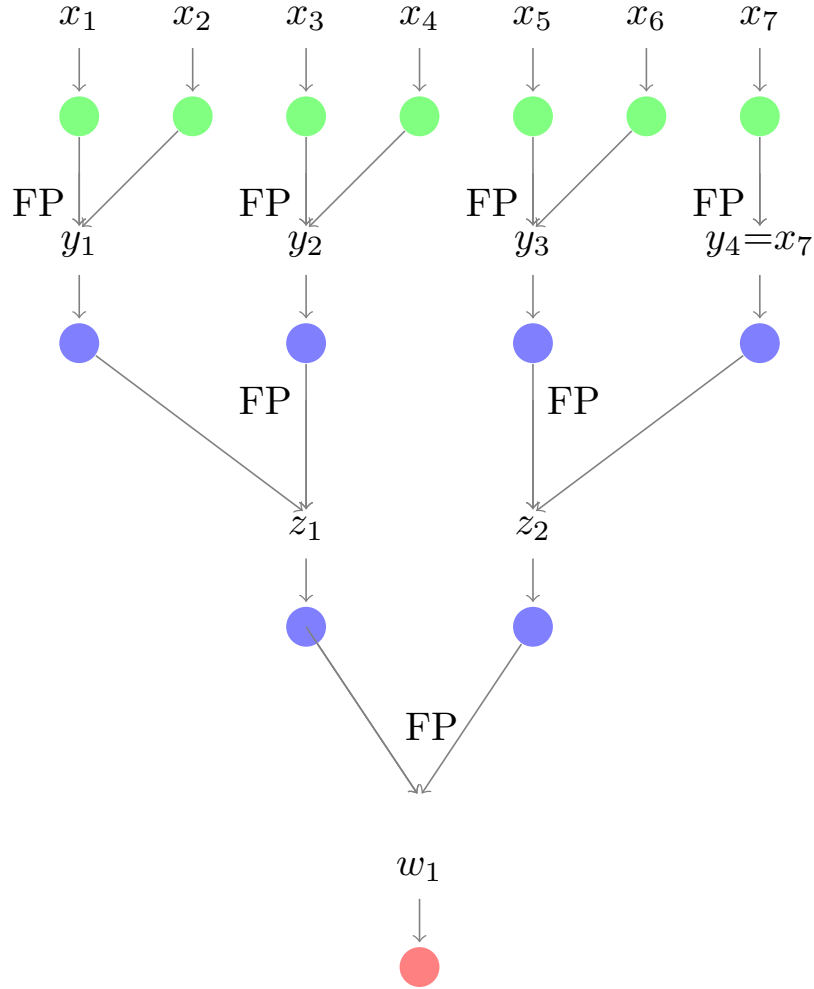
en donde $w_1 \sim x_1, \dots, x_{\frac{d}{2}}$ y $w_2 \sim x_{\frac{d}{2}+1}, \dots, x_d$.

Es fácil ver que, de manera uniforme,

$$\lim_{\epsilon \rightarrow 0} |\Phi_\epsilon^{mult,d}(x) - x_1x_2, \dots, x_d| = 0.$$

Claramente la profundidad de $\Phi_\epsilon^{mult,d}$ es α .

En el caso de que d no sea de la forma $d = 2^\alpha$, hacemos algo similar a lo que se muestra en el siguiente ejemplo con $d = 7$:



Esto demuestra el resultado. □

Teorema 4.5

Sean $d, p \in \mathbb{N}$, $q \geq 2$, $z > 0$ y $\varepsilon > 0$. Dada $f \in B^p$, existe una red neuronal ϕ_ε tal que

$$\|f - R(\phi_\varepsilon)\|_{[-z, z]^d} \leq \varepsilon \quad (0)$$

en donde la función de activación es ρ_s de orden $q \geq 2$ y

$$L(\phi_\varepsilon) = \lceil \log_2(d) \rceil + \max \{ \lceil \log_q(p-1) \rceil, 0 \} + 1.$$

Ademas existe una constante C tal que

$$M(\phi_\varepsilon) \leq C;$$

donde C no depende de f .

Demostración. Sea $f(x) = \prod_{i=1}^d N_{i,t_i,p}(x_i)$, con $x = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$.

Usamos el corolario 4.1 para encontrar redes neuronales ψ_i^n tales que

$$\|R(\psi_i^n) - N_{i,t_i,p}\|_{[-z,z]} \leq \frac{1}{n}. \quad (1)$$

Sea $\phi^{mult,d}$ como en el corolario 4.3 tal que

$$\|R(\phi_{\frac{1}{n}}^{mult,d})(x) - x_1 x_2 \cdots x_d\|_{[-z,z]^d} \leq \frac{1}{n}. \quad (2)$$

Tomamos

$$\phi^n = \phi_{\frac{1}{n}}^{mult,d} \cdot FP(\psi_1^n, \psi_2^n, \dots, \psi_d^n),$$

así,

$$\begin{aligned} R(\phi^n)(x) &= R(\phi_{\frac{1}{n}}^{mult,d}) \circ R(FP(\psi_1^n, \psi_2^n, \dots, \psi_d^n))(x) \\ &= R(\phi_{\frac{1}{n}}^{mult,d}) \begin{pmatrix} R(\psi_1^n)(x_1) \\ \vdots \\ R(\psi_d^n)(x_d) \end{pmatrix}. \end{aligned}$$

Por la ecuación 2, tenemos que

$$\|R(\phi^n) - \prod_{i=1}^d \psi_i^n\| \rightarrow_{n \rightarrow \infty} 0. \quad (3)$$

Además, por la ecuación 1, tenemos que

$$\|\prod_{i=1}^d \psi_i^n - \prod_{i=1}^d N_{i,t_i,p}\| \rightarrow_{n \rightarrow \infty} 0. \quad (4)$$

De las ecuaciones 3 y 4 obtenemos que

$$\|R(\theta^n) - f\| \rightarrow_{n \rightarrow \infty} 0,$$

lo que demuestra la ecuación 0 para n suficientemente grande.

Además, de los corolarios 4.1 y 4.3, obtenemos que

$$L(\phi^n) = \lceil \log_2(d) \rceil + \max\{\lceil \log_q(p-1) \rceil, 0\} + 1$$

y también que existe una constante C tal que

$$M(\phi^n) \leq C \quad \forall f \in B^p.$$

□

Teorema 4.6

Sean $d, p, s \in \mathbb{N}$, $s > \delta > 0$ y $\rho : \mathbb{R} \rightarrow \mathbb{R}$ una función sigmoideal de orden $q \geq 0$. Entonces existe $\tilde{C} > 0$ tal que, para cualesquiera $f \in C^s([0, 1]^d)$ y $\varepsilon \in (0, \frac{1}{2})$, existe una red neuronal ϕ tal que:

- $\|f - R(\phi)\| \leq \varepsilon$,
- $M(\phi) \leq \tilde{C} \left(\frac{\varepsilon}{\|f\|_{C^s}} \right)^{\frac{d}{s-\delta}}$,
- $L(\phi) = \lceil \log_2(d) \rceil + \max \{ \lceil \log_q(p-1) \rceil, 0 \} + 1$.

Demostración. Sean $\delta \in (0, s)$ y $N \in \mathbb{N}$ tales que

$$CN^{\frac{\delta-s}{d}} \|f\|_{C^s([0,1]^d)} < \frac{\varepsilon}{2}$$

en donde C es como en el teorema 4.4, con $k = P$.

Luego

$$\|f - \sum_{i=1}^N CiBi\| < CN^{\frac{\delta-s}{d}} \|f\|_{C^s([0,1]^d)}.$$

Se tiene que

$$\left(\frac{C\|f\|_{C^s}}{\varepsilon/2} \right) < N^{\frac{s-\delta}{d}},$$

de lo que se obtiene que

$$\left(\frac{C\|f\|_{C^s}}{\varepsilon/2} \right)^{\frac{d}{s-\delta}} < N.$$

Así, por el teorema 4.4, se tiene que

$$\sigma(A_N, |f|) \leq \frac{\varepsilon}{2}$$

con

$$A_N = \left\{ \sum_{i=1}^{\infty} CiBi : Ci \in \mathbb{R}, \|C\|_0 \leq N \right\},$$

en donde $\|C\|_0 = \#\{i \in \mathbb{N} : C_i \neq 0\}$ y $B_i \in B^P$ (ver la definición 4.11 con $D = B^P$).

Se sigue de los Teorema 4.4 y 4.5 que existe una constante \tilde{C} tal que

$$\sigma(B_{\tilde{C}N}, |f|) \leq \sigma(A_N, |f|) \leq \frac{\varepsilon}{2},$$

con

$$B_{\tilde{C}N} = \{R(\phi) : L(\phi) = L \text{ y } M(\phi) \leq \tilde{C}N\}.$$

Así,

$$\sigma(B_{\tilde{C}N}, |f|) \leq \frac{\varepsilon}{2},$$

lo que implica que existe ϕ con $L(\phi) = L$ y $M(\phi) \leq \tilde{C}N$ tal que

$$\|f - R(\phi)\| < \varepsilon.$$

En este caso podemos tomar cualquier N tal que

$$\left(\frac{C\|f\|_{C^s}}{\varepsilon/2} \right)^{\frac{d}{s-\delta}} < N,$$

por ejemplo

$$N = \left\lceil \left(\frac{C\|f\|_{C^s}}{\varepsilon/2} \right)^{\frac{d}{s-\delta}} \right\rceil,$$

entonces existe otra constante \tilde{C} tal que

$$M(\phi) \leq \tilde{C} \left(\frac{\varepsilon}{\|f\|_{C^s}} \right)^{\frac{d}{s-\delta}}.$$

□

5 | Análisis de Datos con Redes Neuronales

5.1. Ejemplo

Suponga que tenemos un rectángulo $R \subset \mathbb{R}^2$ y que a cada punto $\hat{x} \in R$ se le asigna una etiqueta de color rojo o azul (en este ejemplo tendremos solamente diez puntos).

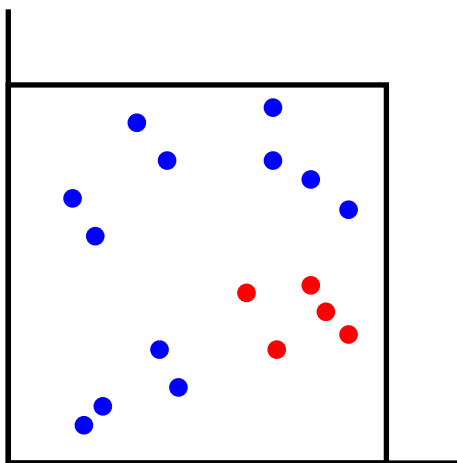


Figura 5.1

En principio no sabemos que puntos tienen asignada cual etiqueta, luego, mediante un muestreo obtenemos esa información para algunos puntos (ver figura 5.1).

Para los puntos de la imagen ya conocemos su etiqueta correspondiente.

Sea $R = [0, 1] \times [0, 1]$ y considere el conjunto de los puntos sobre los cuales se realizó el muestreo

$$\{\hat{x}^{(i)} : \hat{x}^{(i)} \in R, i = 1, \dots, 10\}.$$

Sea $y : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ tal que

$$y(\hat{x}) = \begin{cases} \begin{bmatrix} 1 \\ 0 \end{bmatrix} & \text{si la etiqueta de } x \text{ es de color rojo,} \\ \begin{bmatrix} 0 \\ 1 \end{bmatrix} & \text{si la etiqueta de } x \text{ es de color azul.} \end{cases}$$

No conocemos a la función y , pero quisiéramos encontrar una función (red neuronal) que la aproxime. Lo único que conocemos de la función y es su valor en los puntos $\hat{x}^{(i)}$; a partir de esta información queremos obtener una red neuronal que aproxime a la función y de una forma “satisfactoria”.

Planteamiento del Problema:

Queremos determinar, para cada punto del cuadrado R , si su etiqueta es de color rojo o azul usando redes neuronales.

Considere una función de activación sigmoideal $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ dada por

$$\sigma(x) = \frac{1}{1 + \exp(-x)}.$$

Podemos aproximar a la función y en los puntos conocidos $\hat{x}^{(i)}$ con una red neuronal

$$F = R(T_1, \dots, T_n),$$

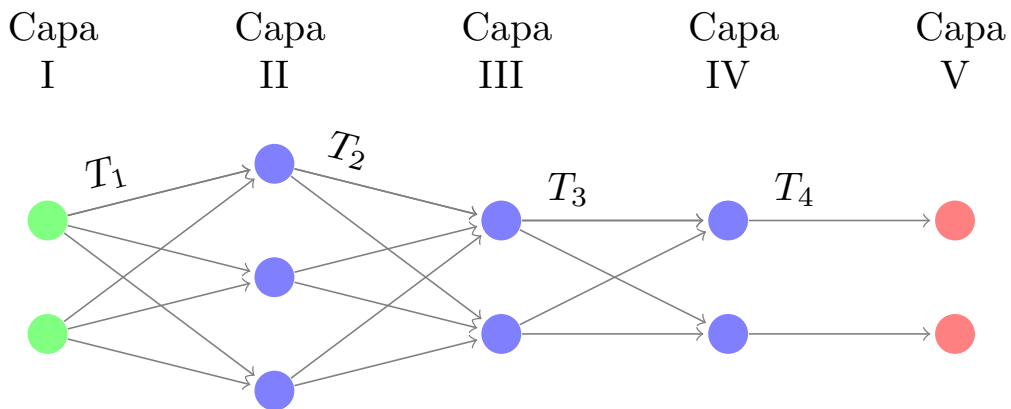
donde

$$F(\hat{x}) = T_n \circ \sigma \circ T_{n-1} \circ \sigma \circ \dots \circ \sigma \circ T_1(\hat{x})$$

y definimos una función de costo como sigue:

$$\text{cost}(T_1, \dots, T_n) := \frac{1}{10} \sum_{i=1}^{10} \frac{1}{2} \|y(\hat{x}^{(i)}) - F(\hat{x}^{(i)})\|^2.$$

Considere la siguiente red neuronal con 4 capas



donde:

- $T_1 : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$,
- $T_2 : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$,
- $T_3 : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ y
- $T_4 : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ está dada por $T_4(\hat{x}) = \hat{x}$.

Así,

$$F = \sigma \circ T_3 \circ \sigma \circ T_2 \circ \sigma \circ T_1.$$

Para cada transformación afín T_j , con $j \in \{1, 2, 3\}$, existen matrices W_j y vectores b_j tales que

$$T_j(\hat{z}) = W_j(\hat{z}) + b_j,$$

donde $W_1 \in \mathcal{M}_{3 \times 2}(\mathbb{R})$, $b_1 \in \mathbb{R}^3$, $W_2 \in \mathcal{M}_{2 \times 3}(\mathbb{R})$, $b_2 \in \mathbb{R}^2$, $W_3 \in \mathcal{M}_{2 \times 2}(\mathbb{R})$ y $b_3 \in \mathbb{R}^2$.

Tenemos que F depende W_1 , W_2 , W_3 , b_1 , b_2 y b_3 , los cuales tienen 17 parámetros que deben elegirse de tal manera que se minimice la función de costo

$$\text{cost}(T_1, \dots, T_4).$$

Para encontrar el mínimo de la función de costo normalmente se utilizan programas de cómputo; por ejemplo, en MATLAB se usa el ‘optimization toolbox’ sobre los parámetros de las transformaciones afines (específicamente se usa ‘non-linear least-squares solver is q nonlin’).

Sea $F : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ dada por

$$F(\hat{x}) = \begin{pmatrix} F_1(\hat{x}) \\ F_2(\hat{x}) \end{pmatrix} \in \mathbb{R}^2$$

la red neuronal que obtenemos después de encontrar el mínimo de la función de costo.

Si $F_1(x) > F_2(x)$ interpretamos que la etiqueta de \hat{x} es de color azul y

si $F_1(x) \leq F_2(x)$ interpretamos que la etiqueta de \hat{x} es de color rojo.

Con esta condición obtenemos la siguiente gráfica:

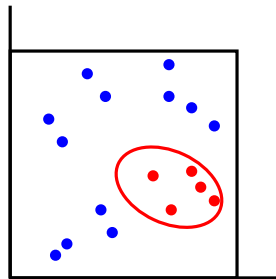


Figura 5.2

De elegir otra red, obtendríamos otro resultado; por ejemplo:

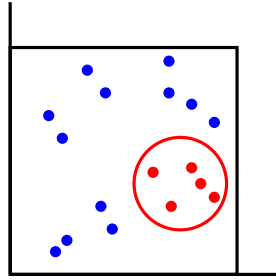


Figura 5.3

5.2. Esquema General

Supongamos que tenemos N muestras de un fenómeno. Estas muestras se obtienen en puntos $(\hat{x}_i)_{i=1}^N$ con resultados $(y(\hat{x}_i))_{i=1}^n$. Construimos una red neuronal

$$\phi = (T_1, T_2, \dots, T_n)$$

con

$$F = R(\phi).$$

La función de costo correspondiente es

$$\text{cost}(T_1, T_2, \dots, T_n) = \frac{1}{n} \sum_{i=1}^N \frac{1}{2} \|y(\hat{x}_i) - f(\hat{x}_i)\|^2.$$

La idea es encontrar los parámetros T_1, T_2, \dots, T_n que determinan el mínimo de la función de costo.

5.3. Método de Descenso por el Gradiente (Convergencia)

Dada una función continuamente diferenciable

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

queremos encontrar

$$\min_{\hat{x} \in \mathbb{R}^n} f(\hat{x}).$$

Definición 5.1

Una función de forzamiento es una función continua $\sigma : [0, \infty) \rightarrow (-\infty, 0]$ tal que:

- $\sigma(0) = 0$,
- $\sigma(t) > 0$ para todo $t > 0$ y
- si (t_n) es una sucesión en \mathbb{R} tal que, cuando $n \rightarrow \infty$,

$$\sigma(t_n) \rightarrow 0,$$

entonces

$$t_n \rightarrow 0.$$

Por ejemplo, para $c > 0$, $\sigma(t) = ct$ y $\sigma(t) = ct^2$ son funciones de forzamiento.

Definición 5.2

Denotamos por $LC_L^1(\mathbb{R}^n)$ al conjunto de funciones $f \in C^1(\mathbb{R}^n)$ tales que, para cualesquiera $\hat{x}, \hat{y} \in \mathbb{R}^n$,

$$\|\nabla f(\hat{x}) - \nabla f(\hat{y})\| \leq L\|\hat{x} - \hat{y}\|.$$

Lema 5.1

Sea $f \in LC_L^1(\mathbb{R}^n)$, entonces, para cualesquiera $\hat{x}, \hat{y} \in \mathbb{R}^n$,

$$|f(\hat{x} + \hat{y}) - f(\hat{x}) - \langle \nabla f(\hat{x}), \hat{y} \rangle| \leq \frac{L}{2} \|\hat{y}\|^2.$$

Demostración. Escribimos

$$f(\hat{x} + \hat{y}) = f(\hat{x}) + \int_0^1 \langle \nabla f(\hat{x} + z\hat{y}), \hat{y} \rangle dz.$$

Tenemos que

$$\begin{aligned} f(\hat{x} + \hat{y}) &= f(\hat{x}) + \int_0^1 \langle \nabla f(\hat{x} + z\hat{y}), \hat{y} \rangle dz \\ &= f(\hat{x}) + \int_0^1 \langle \nabla f(\hat{x}), \hat{y} \rangle dz + \int_0^1 \langle \nabla f(\hat{x} + z\hat{y}) - \nabla f(\hat{x}), \hat{y} \rangle dz \\ &= f(\hat{x}) + \langle \nabla f(\hat{x}), \hat{y} \rangle + \int_0^1 \langle \nabla f(\hat{x} + z\hat{y}) - \nabla f(\hat{x}), \hat{y} \rangle dz. \end{aligned}$$

Ahora usamos el hecho de que $f \in LC_L^1(\mathbb{R}^n)$ para obtener que, para todo $z \geq 0$,

$$\|\nabla f(\hat{x} + z\hat{y}) - \nabla f(\hat{x})\| \leq Lz\|\hat{y}\|.$$

Así,

$$|f(\hat{x} + \hat{y}) - f(\hat{x}) - \langle \nabla f(\hat{x}), \hat{y} \rangle| \leq \int_0^1 |\langle \nabla f(\hat{x} + z\hat{y}) - \nabla f(\hat{x}), \hat{y} \rangle|$$

$$\leq \int_0^1 \|\nabla f(\hat{x} + z\hat{y}) - \nabla f(\hat{x})\| \|\hat{y}\| dz \leq \int_0^1 Lz \|\hat{y}\|^2 dz = \frac{L}{2} \|\hat{y}\|^2.$$

□

Lema 5.2

Sean $(a_n)_{n \in \mathbb{N}}$ y $(\epsilon_n)_{n \in \mathbb{N}}$ sucesiones positivas tales que, para toda $n \in \mathbb{N}$,

- $0 \leq a_{n+1} \leq a_n + \epsilon_n$ y
- $\sum_{i=1}^{\omega} \epsilon_n < +\infty$,

entonces $(a_n)_{n \in \mathbb{N}}$ converge.

Demostración.

De las hipótesis se obtiene que

$$0 \leq a_{n+m} \leq a_n + \sum_{j=n}^{n+m} \epsilon_j.$$

Luego,

$$\begin{aligned} a_{n+m} - a_n &\leq \sum_{j=n}^{n+m} \epsilon_j \\ \Rightarrow a_n - a_{n+m} &\geq - \sum_{j=n}^{n+m} \epsilon_j. \end{aligned}$$

Así, cuando $m, n \rightarrow \infty$,

$$|a_{n+m} - a_n| \leq \sum_{j=n}^{m+n} \epsilon_j \rightarrow 0;$$

es decir, $(a_n)_{n \in \mathbb{N}}$ es de Cauchy y por tanto converge.

□

Teorema 5.1: Descenso de gradiente

Sea $f \in C^1(\mathbb{R}^n)$ una función acotada inferiormente. Construimos una sucesión como sigue:

$$\hat{x}_0 \in \mathbb{R}^n \text{ y}$$

$$\hat{x}_{i+1} = \hat{x}_i - h_i d_i,$$

donde el tamaño $h_i > 0$ y la dirección del paso $d_i \in \mathbb{R}^n$.

Se cumple que $\sum_i h_i = \infty$ y existen sucesiones $\{\lambda_i\}_{i \in \mathbb{N} \setminus \{0\}}, \{\nu_i\}_{i \in \mathbb{N} \setminus \{0\}}$ en $[0, \infty)$ tales que

$$\sum \lambda_i h_i < \infty, \sum \nu_i < \infty$$

y se satisface que

$$\langle \nabla f(\hat{x}_i), d_i \rangle \geq \sigma(\|\nabla f(\hat{x}_i)\|) - \lambda_i, \quad (1)$$

$$f(\hat{x}_i) - f(\hat{x}_{i+1}) \geq h_i \langle \nabla f(\hat{x}_i), d_i \rangle - \nu_i, \quad (2)$$

en donde σ es una función de forzamiento.

En caso de que $\nabla f(\hat{x}_i) = 0$ para alguna i , truncamos la sucesión (con la primera i que satisfaga esto).

Entonces la sucesión $\{f(\hat{x}_i)\}_{i \in \mathbb{N}}$ converge y para toda $b \in \mathbb{R}$

$$\inf_{i \geq b} \|\nabla f(\hat{x}_i)\| = 0.$$

Además si $f \in LC_L^1(\mathbb{R}^n)$ y la sucesión $\{\|d_i\|\}_{i \in \mathbb{N}}$ es acotada, entonces

$$\lim_{i \rightarrow \infty} \nabla f(\hat{x}_i) = 0$$

y para cualquier punto de acumulación \hat{x} de $\{\hat{x}_i\}_{i \in \mathbb{N}}$ se cumple que $\nabla f(\hat{x}) = 0$.

Demostración. De existir un índice r tal que $\nabla f(\hat{x}_r) = 0$, entonces \hat{x}_r ya es un punto crítico y truncamos la sucesión.

Suponga que, para toda $i \in \mathbb{N}$, $\nabla f(\hat{x}_i) \neq 0$.

De las ecuaciones 1 y 2 obtenemos que

$$f(\hat{x}_i) - f(\hat{x}_{i+1}) \geq h_i \sigma(\|\nabla f(\hat{x}_i)\|) - \lambda_i h_i - \nu_i. \quad (3)$$

Sea $a = \inf_{\hat{y} \in \mathbb{R}^n} \{f(\hat{y})\}$. Usando la ecuación 3, obtenemos que

$$0 \leq f(\hat{x}_{i+1}) - a \leq f(\hat{x}_i) - a + \lambda_i h_i + \nu_i.$$

Como $\sum \lambda_i h_i < \infty$, $\sum \nu_i < \infty$, y si $\epsilon_i = h_i \lambda_i + \nu_i$, entonces $\sum_i \epsilon_i < \infty$.

Usando el lema 5.2, con $a_n = f(\hat{x}_n) - a$, se tiene que $\{a_n\}_n$ converge; por lo tanto, $\{f(\hat{x}_i)\}_{i \in \mathbb{N}}$ converge.

Aplicando 3 obtenemos que, para $i > b$,

$$\begin{aligned} f(\hat{x}_b) - f(\hat{x}_i) &= \sum_{j=b}^{i-1} (f(\hat{x}_j) - f(\hat{x}_{j+1})) \\ &\geq \sum_{j=b}^{i-1} h_j \sigma(\|\nabla f(\hat{x}_j)\|) - \sum_{j=0}^{i-1} (\lambda_j h_j + \nu_j). \end{aligned}$$

Luego,

$$f(\hat{x}_b) - f(\hat{x}_i) \geq \inf_{b \leq j \leq i-1} \sigma(\|\nabla f(\hat{x}_j)\|) \sum_{j=0}^{i-1} h_j - \sum_{j=0}^{i-1} (\lambda_j h_j + \nu_j).$$

Por lo que, para toda $i > b$,

$$f(\hat{x}_b) - a \geq \inf_{j \geq b} \sigma(\|\nabla f(\hat{x}_j)\|) \sum_{j=0}^{i-1} h_j - \sum_{j=0}^{\infty} (\lambda_j h_j + \nu_j),$$

entonces

$$f(\hat{x}_b) - a \geq \inf_{j \geq b} \sigma(\|\nabla f(\hat{x}_j)\|) \sum_{j=0}^{\infty} h_j - \sum_{j=0}^{\infty} (\lambda_j h_j + \nu_j).$$

Como por hipótesis $\sum_{j=0}^{\infty} h_j = \infty$, entonces, para todo $b \geq 0$

$$\inf_{j \geq b} \sigma(\|\nabla f(\hat{x}_j)\|) = 0.$$

Por definición de función de forzamiento tenemos que, para todo $b \geq 0$

$$\inf_{j \geq b} \|\nabla f(\hat{x}_j)\| = 0 \tag{4}$$

Ahora supongamos que $f \in LC_L^1(\mathbb{R}^n)$ y, para toda $i \in \mathbb{N}$, $\|d_i\| \leq c$.

Demostraremos que $\|\nabla f(\hat{x}_i)\| \rightarrow 0$ cuando $i \rightarrow \infty$.

Supongamos lo contrario. Entonces existe $\epsilon > 0$ y una subsucesión $\{\hat{x}_{i_j}\}_{j \in \mathbb{N}}$ tal que, para toda j ,

$$\|\nabla f(\hat{x}_{i_j})\| \geq \epsilon.$$

Además de la ecuación 4 se sigue que para toda j existe otro índice L , con $L > i_j$, tal que

$$\|\nabla f(\hat{x}_L)\| \leq \frac{\epsilon}{2}.$$

Denotaremos por $L(j)$ al número entero más pequeño que cumple que, para todo $s \in \{i_j, \dots, L(j) - 1\}$:

- $L(j) > i_j$,
- $\|\nabla f(\hat{x}_{L(j)})\| \leq \frac{\epsilon}{2}$ y
- $\|\nabla f(\hat{x}_s)\| > \frac{\epsilon}{2}$.

Usamos la desigualdad del triángulo y el hecho de que $f \in LC_L^1(\mathbb{R}^n)$ para obtener

$$\begin{aligned}
 \frac{\epsilon}{2} &\leq \|\nabla f(\hat{x}_{i_j})\| - \|\nabla f(\hat{x}_{L(j)})\| \\
 &\leq \|\nabla f(\hat{x}_{i_j}) - \nabla f(\hat{x}_{L(j)})\| \\
 &\leq L\|\hat{x}_{i_j} - \hat{x}_{L(j)}\| = L\left\|\sum_{t=i_j}^{L(j)-1} (\hat{x}_t - \hat{x}_{t+1})\right\| \\
 &\leq L\sum_{t=i_j}^{L(j)-1} \|\hat{x}_t - \hat{x}_{t+1}\| = L\sum_{t=i_j}^{L(j)-1} h_t\|d_t\| \\
 &\leq LC\sum_{t=i_j}^{L(j)-1} h_t.
 \end{aligned}$$

Entonces obtenemos

$$\frac{\epsilon}{2} \leq LC\sum_{t=i_j}^{L(j)-1} h_t,$$

por lo que

$$\sum_{t=i_j}^{L(j)-1} h_t \geq \frac{\epsilon}{2LC} := H. \quad (5)$$

Por las ecuaciones 3 y 5, obtenemos que

$$\begin{aligned}
 f(\hat{x}_{i_j}) - f(\hat{x}_{L(j)}) &= \sum_{t=i_j}^{L(j)-1} (f(\hat{x}_t) - f(\hat{x}_{t+1})) \\
 &\geq \sum_{t=i_j}^{L(j)-1} h_t \sigma(\|\nabla f(\hat{x}_t)\|) - \sum_{t=i_j}^{L(j)-1} (\lambda_t h_t + \nu_t) \\
 &\geq \inf_{i_j \leq t \leq L(j)-1} \sigma(\|\nabla f(\hat{x}_t)\|) H - \sum_{t=i_j}^{\infty} (\lambda_t h_t + \nu_t).
 \end{aligned}$$

Como la sucesión $\{f(\hat{x}_i)\}_{i \in \mathbb{N}}$ converge y $\lim_{j \rightarrow \infty} \sum_{t=i_j}^{\infty} (\lambda_t h_t + \nu_t) = 0$, obtenemos que

$$\lim_{j \rightarrow \infty} \inf_{i_j \leq t \leq L(j)-1} \sigma(\|\nabla f(\hat{x}_t)\|) = 0. \quad (6)$$

Como $\{f(\hat{x}_i)\}_{i \in \mathbb{N}}$ es una sucesión de Cauchy, entonces $\lim_{j \rightarrow \infty} |f(\hat{x}_{i_j}) - f(\hat{x}_{L(j)})| = 0$.

Note que por la elección de i_j y $L(j)$ tenemos que, para toda $t \in \{i_j, i_{j+1}, \dots, L(j) - 1\}$,

$$\|\nabla f(\hat{x}_t)\| > \frac{\epsilon}{2} \quad (7)$$

pues $L(j)$ es el primer natural mayor que i_j tal que $\|\nabla f(\hat{x}_{L(j)})\| \leq \frac{\epsilon}{2}$.

Las ecuaciones 6 y 7 son contradictorias porque σ es una función de forzamiento. Por lo tanto

$$\lim_{i \rightarrow \infty} \|\nabla f(\hat{x}_i)\| = 0.$$

Para terminar la demostración sólo falta probar que, para toda $\hat{x} \in \{\hat{x}_i\}_{i \in \mathbb{N}}$, si \hat{x} es un punto de acumulación, entonces

$$\nabla f(\hat{x}) = 0.$$

Sea \hat{x} un punto de acumulación de $\{\hat{x}_i\}_{i \in \mathbb{N}}$. Entonces existe una subsucesión $\{\hat{x}_{s_n}\}_{n \in \mathbb{N}}$ tal que

$$\lim_{n \rightarrow \infty} \hat{x}_{s_n} = \hat{x}.$$

Como $\{\|\nabla f(\hat{x}_i)\|\} \rightarrow 0$, entonces:

$$\lim_{n \rightarrow \infty} \nabla f(\hat{x}_{s_n}) = 0$$

Como ∇f es una función continua, entonces:

$$0 = \lim_{n \rightarrow \infty} \nabla f(\hat{x}_{s_n}) = \nabla f(\lim_{n \rightarrow \infty} \hat{x}_{s_n}) = \nabla f(\hat{x}).$$

□

Corolario 5.1

Sean $f \in LC_L^1(\mathbb{R}^n)$ una función acotada inferiormente y $M \in \mathbb{R}$ tal que, para todo $\hat{x} \in \mathbb{R}^n$,

$$\|\nabla f(\hat{x})\| \leq M. \quad (1)$$

Construimos una sucesión de la siguiente forma:

$$\lambda_0 \in \mathbb{R}^n \text{ arbitrario y}$$

$$\lambda_{i+1} = \lambda_i - h_i d_i, \quad (2)$$

en donde

$$d_i = \nabla f(\hat{x}_i) - \hat{e}_i \quad (3)$$

y si $\nabla f(\hat{x}_i) = 0$ paramos.

Si $\hat{e}_i \in \mathbb{R}^n$ y $h_i \in (0, \infty)$ son tales que, para toda i ,

$$\sum h_i = \infty, \quad \sum h_i^2 < \infty, \quad \sum h_i \|\hat{e}_i\| < \infty \quad \text{y para algún } \gamma \in \mathbb{R} \quad \|\hat{e}_i\| \leq \gamma,$$

entonces las conclusiones del Teorema 5.1 se cumplen, es decir, $(f(\hat{x}_i))_{i \in \mathbb{N}}$ converge,

$$\lim_{n \rightarrow \infty} \nabla f(\hat{x}_i) = 0$$

y para cualquier punto de acumulación \hat{x} de $(\hat{x}_i)_{i \in \mathbb{N}}$, se tiene que $\nabla f(\hat{x}) = 0$.

Demostración. Demostraremos las hipótesis del Teorema 5.1, es decir, veamos que:

- $\langle \nabla f(\hat{x}_i), d_i \rangle \geq \sigma(\|\nabla f(\hat{x}_i)\|) - \lambda_i$,
- $\langle \nabla f(\hat{x}_i), d_i \rangle \geq \sigma(\|\nabla f(\hat{x}_i)\|) - \lambda_i$,
- $f(\hat{x}_i) - f(\hat{x}_{i+1}) \geq h_i \langle \nabla f(\hat{x}_i), d_i \rangle - v_i$,

en donde tenemos que encontrar $\lambda_i \geq 0$, $v_i \geq 0$ tales que $\sum \lambda_i h_i < \infty$, $\sum v_i = \infty$ y además queremos ver que, para toda i , $\|d_i\| \geq c$.

Primero veamos que $\|d_i\| \leq c$, para alguna c .

De las ecuaciones 1, 2 y 3 tenemos que, para toda i ,

$$\|d_i\| \leq \|\nabla f(\hat{x}_i)\| + \|\hat{e}_i\| \leq M + \gamma := c.$$

Ahora veamos como encontramos λ_i, v_i .

Utilizamos la desigualdad de Cauchy y tomamos $\sigma(s) = s^2$. Calculamos

$$\begin{aligned}
 \langle \nabla f(\hat{x}_i), d_i \rangle &= \langle \sigma(f(\hat{x}_i)), \sigma(f(\hat{x}_i)) - \hat{e}_i \rangle \\
 &= \|\nabla f(\hat{x}_i)\|^2 - \langle \nabla f(\hat{x}_i), \hat{e}_i \rangle \\
 &\geq \|\nabla f(\hat{x}_i)\|^2 - \|\nabla f(\hat{x}_i)\| \|\hat{e}_i\| \\
 &\geq \sigma(\sigma(f(\hat{x}_i))) - M \|\hat{e}_i\| \\
 &= \sigma(\sigma(f(\hat{x}_i))) - \lambda_i.
 \end{aligned}$$

Con $\lambda_i = M \|\hat{e}_i\|$, cumple

$$\sum h_i \lambda_i < \infty, \quad \lambda_i \geq 0,$$

entonces

$$\langle \nabla f(\hat{x}_i), d_i \rangle \geq \sigma(\|f(\hat{x}_i)\|) - \lambda_i.$$

Ahora calculamos, usando el Lema 5.1,

$$|f(\hat{x}_i) - f(\hat{x}_{i+1}) - \langle \nabla f(\hat{x}_i), \hat{x}_i - \hat{x}_{i+1} \rangle| \leq \frac{L}{2} \|\hat{x}_{i+1} - \hat{x}_i\|^2.$$

Por lo que, usando la ecuación 1, obtenemos

$$\begin{aligned}
 f(\hat{x}_i) - f(\hat{x}_{i+1}) &\geq \langle \nabla f(\hat{x}_i), \hat{x}_i - \hat{x}_{i+1} \rangle - \frac{L}{2} \|\hat{x}_{i+1} - \hat{x}_i\|^2 \\
 &= \langle \nabla f(\hat{x}_i), h_i d_i \rangle - \frac{L}{2} \|h_i d_i\|^2 \\
 &\geq h_i \langle \nabla f(\hat{x}_i), d_i \rangle - \frac{L}{2} h_i^2 c^2.
 \end{aligned}$$

Ahora tomamos $v_i = \frac{L}{2} h_i^2 c^2$ con lo que tenemos

$$f(x_i) - f(x_{i+1}) \geq h_i \langle \nabla f(x_i), d_i \rangle - v_i,$$

con $v_i \geq 0$ y $\sum v_i < \infty$. Con esto terminamos de demostrar las hipótesis del Teorema 5.1 y, por lo tanto, se cumplen sus conclusiones. □