Universidad Nacional Autónoma de México

Facultad de Ciencias

Aproximación al Kernel Neuronal Tangente

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Matemático

PRESENTA:

Axel Francisco Leon Paloma

TUTOR:

Dr. Miguel Arturo Ballesteros Montero

COTUTOR:

Dr. Iván Pavlovich Naumkin Kaikin

Ciudad Universitaria, Ciudad de México, 2023

A Dios, y a mis padres,
Francisco y Angela.

Agradecimientos

A mis padres, Francisco y Angela, que me han apoyado a lo largo de toda mi formación de manera incondicional, dando todo de sí mismos para ayudarme a cumplir mis metas. A mi hermano, Gustavo por el apoyo y acompañamiento en diversas situaciones a lo largo de mi vida.

A mis amigos, Humberto por estar presente en los momentos más complicados y ayudarme a creer en mi; Jairo por tantos años de amistad y apoyo; Janeth por animarme y ser un ejemplo de trabajo para mi; Rodrigo en quien encontré un gran amigo en los últimos momentos de la carrera; Daniel y Mario de quienes recibí apoyo cuando iniciaba en temas de inteligencia artificial, que hoy es el tema del presente trabajo.

A mi asesor y director de tesis, el Dr. Miguel Ballesteros, por todo el apoyo que recibí a lo largo de mi carrera, por permitirme ser parte de su equipo de trabajo, y brindarme así oportunidades de crecimiento académico, humano y profesional. A Gerardo y Fedro, con quienes trabajé de manera más cercana y me brindaron siempre su apoyo y respaldo, y a todos los del departamento de física matemática del IIMAS, por quienes siempre fui bien recibido y acogido.

A la UNAM, y en especial a la Facultad de Ciencias, por brindarme tantas oportunidades de crecimiento, por permitirme aprender de personas con un gran perfil académico y una increíble formación humana a la vez, que influyeron de manera benéfica en mi formación.

La investigación de esta tesis contó con diversos apoyos:

1. Proyecto apoyado por el CONACYT, FORDECYT-PRONACES 429825/2020 (recientemente renombrado como Proyecto CF-2019/429825).
2. Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM IN101621.

Mi sincero agradecimiento por todo el apoyo recibido.

Índice general

Introducción

En la última década la inteligencia artificial, y en particular las técnicas de deep learning, han cobrado gran relevancia gracias a la redes neuronales artificiales (ANN). El origen de las ANNs se remonta al siglo pasado con investigaciones que se inspiraban en las neuronas biológicas y su capacidad de procesamiento. Por aquellos años esta teoría progresó de manera muy lenta por varias razones, de entre las cuales podemos destacar que el poder de cómputo era mucho menor al que tenemos hoy en día disponible, esto dió paso a una época que ahora nombramos como el invierno de la IA.

Cuando se creía que las ANNs no eran más que una curiosidad, surgieron métodos que atrajeron la atención de mercado, como es el caso de los métodos kernel. La teoría matemática detrás de los métodos kernel es la de los espacios de Hilbert de kernel reproductor (RKHS), que había sido propuesta alrededor de medio siglo antes de que Vladimir Vapnik y Alexey Chervonenkis comenzaran los primeros trabajos con métodos kernel, entre los cuales destacan las máquinas de soporte vectorial (SVM) que fueron planteadas por Vladimir Vapnik e Isabelle Guyon, y desarrolladas más a fondo por Bernhard Schölkopf, quién recibió la asesoría de Vapnik durante su doctorado en ciencias de la computación.

La teoría de los métodos kernel desde el principio se fundamentaba en una sólida teoría matématica, además su implemetación era posible y mostraban un gran desempeño en la práctica, razones por las cuales las ANNs parecían innecesarias. Pero, a pesar de las críticas que llegaron a recibir, las ANNs resurgieron en años recientes gracias a diversos factores entre los que podemos enumerar las tarjetas gráficas, que impulsaron el cómputo en paralelo, y la big data, que hizo posible el almacenamiento y acceso a datos para entrenamiento.

En la actualidad, el desarrollo de la inteligencia artificial, con las ANNs como principal exponente, se da a pasos agigantados (aquello que quizá parecía muy lejano hace un año, hoy es una realidad). Sin embargo, poco se sabe acerca del porqué las redes neronales funcionan tan bien, y es necesario entenderlo si de verdad se quiere sacar el mayor provecho a estas tecnologías, ya que la ignorancia se hereda a nuestros modelos, los cuales que terminan sesgados.

Un enfoque interesante para tratar el estudio de las ANNs surgió en 2018 con la propuesta del Kernel Neuronal Tangente, el cual tiene la ventaja de relacionar directamente los métodos kernel, cuyos fundamentos teóricos se conocen bien, y las ANNs. La importancia de este kernel tan particular se ha podido constatar en años recientes con resultados tan interesantes como el que presentó P. Domingos en su artículo *Every model learned by gradient descent is approximately a kernel machine*.

Sin duda alguna, las ANNs han venido a cambiar la manera en que hacemos las cosas, optimizando muchos procesos que hasta hace pocos años representaban un alto coste económico y humano, pero aún falta mucho por entender y mejorar.

Capítulo 1

Teoría básica de Espacios de Hilbert de Kernel Reproductor (RKHS)

En este capítulo daremos la teoría básica de los RKHSs, que es el fundamento matemático de los métodos kernel. Comenzaremos con resultados básicos de análisis, para después exponer los RKHSs para funciones escalares, y finalmente presentaremos definiciones y un par de resultados que generalizan la teoría a funciones vectoriales. Cabe destacar que trabajaremos sobre el caso real, aunque la teoría puede desarrollarse perfectamente para el caso complejo, así que siempre que hablemos de un espacio vectorial supondremos que es sobre $\mathbb{R}$.

La teoría expuesta aquí se basa principalmente en [5], [8], [2], [21]. Para los resultados que pertenezcan a una fuente particular pondremos la referencia al inicio del enunciado.

1.1. Espacios de Hilbert

Definición 1.1.1.
Sea V un espacio vectorial. Definimos el ***producto interior*** *como una función $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$, que satisface lo siguiente:*

IP1 $\langle x + \lambda y, z \rangle = \langle x, z \rangle + \lambda \langle y, z \rangle, \quad \forall x, y, z \in V, \quad \forall \lambda \in \mathbb{R};$

IP2 $\langle x, y \rangle = \langle y, x \rangle, \quad \forall x, y \in V;$

IP3 $\langle x, x \rangle > 0, \quad$ *si* $x \neq \boldsymbol{0}$.

Notación. Con $(V, \langle \cdot, \cdot \rangle)$ denotaremos a un espacio vectorial V con producto interior $\langle \cdot, \cdot \rangle$.

Proposición 1.1.1 (Propiedades básicas del producto interior)**.**
Sea $(V, \langle \cdot, \cdot \rangle)$. Para toda $x, y, z \in V$ y para toda $\lambda \in \mathbb{R}$ se cumple lo siguiente:

1. $\langle x, y + \lambda z \rangle = \langle x, y \rangle + \lambda \langle x, z \rangle$;
2. $\langle x, \boldsymbol{0} \rangle = \langle \boldsymbol{0}, x \rangle = 0$;
3. $\langle x, x \rangle = 0 \iff x = \boldsymbol{0}$;
4. *Si para toda $x \in V$, $\langle x, y \rangle = \langle x, z \rangle$, entonces $y = z$.*

Demostración. Dados $x, y, z \in V$ y $\lambda \in \mathbb{R}$:

1. $$\langle x, y + \lambda z\rangle = \langle y + \lambda z, x\rangle = \langle y, x\rangle + \lambda\langle z, x\rangle = \langle x, y\rangle + \lambda\langle x, z\rangle.$$

2. $$\langle x, \mathbf{0}\rangle = \langle x, x - x\rangle = \langle x, x\rangle - \langle x, x\rangle = 0.$$
   $$\langle \mathbf{0}, x\rangle = \langle x - x, x\rangle = \langle x, x\rangle - \langle x, x\rangle = 0.$$

3. $\underline{\Longrightarrow}$
   Por contrapuesta de ***IP3***.
   $\underline{\Longleftarrow}$
   Por el inciso anterior.

4. $$\langle x, y\rangle = \langle x, z\rangle, \quad \forall x \in V \quad \Longrightarrow \quad \langle x, y - z\rangle = 0, \quad \forall x \in V.$$
   En particular $\langle y - z, y - z\rangle = 0$, entonces por el inciso anterior $y = z$.

□

Definición 1.1.2.
*Si en la Definición 1.1.1 sustituimos la propiedad **IP3** por* $\boxed{\langle x, x\rangle \geq 0, \forall x \in V}$*, decimos que la función* $\langle \cdot, \cdot\rangle$ *es un* ***semi-producto interior****.*

Para distinguir el producto interior del semi-producto interior usaremos el subíndice s para este último.

Lema 1.1.1.
Sea V un espacio vectorial con semi-producto interior $\langle \cdot, \cdot\rangle_s$, entonces $\forall x, y \in V$, tal que $\langle x, x\rangle_s \neq 0$:

$$\langle z, x\rangle_s = 0,$$

con $z = y - \frac{\langle y, x\rangle_s}{\langle x, x\rangle_s} x$.

Demostración.

$$\langle z, x\rangle_s = \langle y - \frac{\langle y, x\rangle_s}{\langle x, x\rangle_s} x, x\rangle_s = \langle y, x\rangle_s - \frac{\langle y, x\rangle_s}{\langle x, x\rangle_s}\langle x, x\rangle_s = 0.$$

□

Proposición 1.1.2 (Desigualdad de Cauchy, para semi-producto interior)**.**
Sea V un espacio vectorial con semi-producto interior $\langle \cdot, \cdot\rangle_s$, entonces:

$$|\langle x, y\rangle_s| \leq \sqrt{\langle x, x\rangle_s}\sqrt{\langle y, y\rangle_s}, \quad \forall x, y \in V.$$

Demostración. Sean $x, y \in V$.

1. **Si** $\langle x, x\rangle_s > 0$.
   Sea $z = y - \lambda x$, con $\lambda = \frac{\langle y,x\rangle_s}{\langle x,x\rangle_s}$. Por el lema anterior $\langle z, x\rangle_s = 0$, y además:

$$\langle z, \lambda x\rangle_s = \lambda\langle z, x\rangle_s = 0.$$

   Entonces:

$$\begin{aligned}\langle y, y\rangle_s &= \langle z + \lambda x, z + \lambda x\rangle_s \\ &= \langle z, z\rangle_s + \langle \lambda x, z\rangle_s + \langle z, \lambda x\rangle_s + \langle \lambda x, \lambda x\rangle_s \\ &= \langle z, z\rangle_s + \langle \lambda x, \lambda x\rangle_s \\ &\geq \lambda^2\langle x, x\rangle_s = \frac{\langle y, x\rangle_s^2}{\langle x, x\rangle_s}.\end{aligned}$$

   $\therefore |\langle y, x\rangle_s| \leq \sqrt{\langle x, x\rangle_s}\sqrt{\langle y, y\rangle_s}$.
2. **Si** $\langle x, x\rangle_s = 0$.
   Notemos que para toda $r > 0$:

$$\begin{aligned}0 &\leq \langle ry - \langle y, x\rangle_s x, ry - \langle y, x\rangle_s x\rangle_s \\ &= r^2\langle y, y\rangle_s - r\langle y, x\rangle_s\langle y, x\rangle_s - r\langle y, x\rangle_s\langle x, y\rangle_s + \langle y, x\rangle_s^2\langle x, x\rangle_s \\ &= r^2\langle y, y\rangle_s - 2r\langle y, x\rangle_s^2,\end{aligned}$$

   de modo que

$$2\langle y, x\rangle_s^2 \leq r\langle y, y\rangle_s, \text{ para toda } r > 0.$$

   $\therefore |\langle y, x\rangle_s| = 0 = \sqrt{\langle x, x\rangle_s}\sqrt{\langle y, y\rangle_s}$.

□

Proposición 1.1.3 (Desigualdad de Cauchy)**.**
Sea $(V, \langle\cdot,\cdot\rangle)$*, entonces*

$$|\langle x, y\rangle| \leq \sqrt{\langle x, x\rangle}\sqrt{\langle y, y\rangle}, \quad \forall x, y \in V.$$

Demostración. Se sigue de la proposición anterior, pues todo producto interior es un semi-producto interior. □

Definición 1.1.3 (Norma de un espacio vectorial)**.**
Sea V *un espacio vectorial. Definimos una* ***norma*** *en* V *como una función* $\|\cdot\| : V \to \mathbb{R}$*, que satisface lo siguiente:*

N1 $\|x\| = 0 \iff x = \mathbf{0}$;

N2 $\|\lambda x\| = |\lambda|\,\|x\|, \quad \forall x \in V, \quad \forall \lambda \in \mathbb{R}$;

N3 $\|x + y\| \leq \|x\| + \|y\|, \quad \forall x, y \in V$.

Notación. Con $(V, \|\cdot\|)$ denotaremos a un espacio vectorial normado V, con norma $\|\cdot\|$.

Proposición 1.1.4 (Propiedades básicas de la norma)**.**
Sea $(V, \|\cdot\|)$*, entonces se cumple lo siguiente:*

1. $|\|x\| - \|y\|| \leq \|x - y\|, \quad \forall x, y \in V.$
2. *La norma es continua.*

Demostración.

1. $$\|x\| \leq \|x - y\| + \|y\| \Longrightarrow \|x\| - \|y\| \leq \|x - y\|$$
$$\|y\| \leq \|y - x\| + \|x\| \Longrightarrow -\|y - x\| \leq \|x\| - \|y\|.$$

   $\therefore |\|x\| - \|y\|| \leq \|x - y\|$
2. Se sigue del inciso anterior.

□

Proposición 1.1.5.
*Sea $(V, \langle \cdot, \cdot \rangle)$, entonces la función $\| \cdot \| : V \to \mathbb{R}$, definida como $\|x\| := \sqrt{\langle x, x \rangle}$, $\forall x \in V$, es una norma en V (**norma inducida** por el producto interior).*

Demostración.
Por ***IP3*** tenemos que $\sqrt{\langle x, x \rangle} \in \mathbb{R}, \quad \forall x \in V.$

- ***N1.*** Del inciso 3 de la Proposición 1.1.1 se sigue que:
$$\sqrt{\langle x, x \rangle} = 0 \Longleftrightarrow \langle x, x \rangle = 0 \Longleftrightarrow x = \mathbf{0}.$$
- ***N2.***
$$\sqrt{\langle \lambda x, \lambda x \rangle} = \sqrt{\lambda^2 \langle x, x \rangle} = |\lambda| \sqrt{\langle x, x \rangle}, \quad \forall x \in V, \quad \forall \lambda \in \mathbb{R}.$$
- ***N3.***
$$\begin{aligned} \langle x + y, x + y \rangle &= \langle x, x \rangle + \langle x, y \rangle + \langle x, y \rangle + \langle y, y \rangle \\ &= \langle x, x \rangle + 2\langle x, y \rangle + \langle y, y \rangle \\ &\leq \langle x, x \rangle + 2\sqrt{\langle x, x \rangle}\sqrt{\langle y, y \rangle} + \langle y, y \rangle \\ &= \left(\sqrt{\langle x, x \rangle} + \sqrt{\langle y, y \rangle} \right)^2. \end{aligned}$$

  $\therefore \sqrt{\langle x + y, x + y \rangle} \leq \sqrt{\langle x, x \rangle} + \sqrt{\langle y, y \rangle}, \quad \forall x, y \in V.$

$\therefore \| \cdot \| = \sqrt{\langle \cdot, \cdot \rangle}$ es una norma en V. □

Cuando hablemos de la norma de un espacio vectorial con producto interior nos referiremos a la norma inducida. De este modo podemos reescribir la desigualdad de Cauchy como:

$$\textit{Sea } (V, \langle \cdot, \cdot \rangle), \textit{ entonces } |\langle x, y \rangle| \leq \|x\|\|y\|, \quad \forall x, y \in V.$$

Proposición 1.1.6.
El producto interior es continuo.

Demostración. Sean $(V, \langle\cdot,\cdot\rangle)$ y $((x_n, y_n))_{n\in\mathbb{N}} \subseteq V \times V$ una sucesión tal que $(x_n, y_n) \to (x, y)$ en $V \times V$, para alguna $(x, y) \in V \times V$. En particular $x_n \to x$, $y_n \to y$.
Por la desigualdad de Cauchy:

$$\begin{aligned} 0 \leq |\langle x, y\rangle - \langle x_n, y_n\rangle| &= |\langle x, y\rangle - \langle x, y_n\rangle + \langle x, y_n\rangle - \langle x_n, y_n\rangle| \\ &= |\langle x, y - y_n\rangle + \langle x - x_n, y_n\rangle| \\ &\leq |\langle x, y - y_n\rangle| + |\langle x - x_n, y_n\rangle| \\ &\leq \|x\|\|y - y_n\| + \|x - x_n\|\|y_n\|. \end{aligned}$$

Finalmente aplicando límite a la desigualdad anterior, y por la continuidad de la norma:

$$\begin{aligned} 0 &\leq \lim_{n\to\infty} |\langle x, y\rangle - \langle x_n, y_n\rangle| \\ &\leq \lim_{n\to\infty} (\|x\|\|y - y_n\|) + \lim_{n\to\infty} (\|x - x_n\|\|y_n\|) \\ &= \|x\| \lim_{n\to\infty} \|y - y_n\| + \lim_{n\to\infty} \|x - x_n\| \lim_{n\to\infty} \|y_n\| \\ &= \|x\|(0) + (0)\|y\| = 0 \end{aligned}$$

Así podemos concluir que $\langle x_n, y_n\rangle \to \langle x, y\rangle$ en $\mathbb{R}$, cuando $(x_n, y_n) \to (x, y)$ en $V \times V$, es decir, el producto interior es continuo.
En particular, el producto interior es continuo en cada una de sus entradas. □

Definición 1.1.4.
Dada $(V, \|\cdot\|)$ y $(x_n)_{n\in\mathbb{N}}$ una sucesión de elementos de V. Decimos que $(x_n)_{n\in\mathbb{N}}$ es una ***sucesión de Cauchy*** *si:*

$$\forall \epsilon > 0, \quad \exists N \in \mathbb{N}, \text{ tal que } \|x_n - x_m\| < \epsilon, \quad \forall n, m \geq N.$$

Definición 1.1.5.
Decimos que $(V, \langle\cdot,\cdot\rangle)$ es un ***espacio de Hilbert*** *si es un espacio vectorial completo con la norma inducida por el producto interior, es decir, si toda sucesión de Cauchy es convergente.*

Cuando hablemos de espacios de Hilbert los denotaremos como $(H, \langle\cdot,\cdot\rangle)$ o solamente H.

Proposición 1.1.7 (Producto cartesiano de espacios de Hilbert)**.**
Sean $(H_1, \langle\cdot,\cdot\rangle_1), ..., (H_n, \langle\cdot,\cdot\rangle_n)$ espacios de Hilbert. Entonces $H := H_1 \times \cdots \times H_n$ es un espacio de Hilbert con el producto interior:

$$\langle (x_1, ..., x_n), (y_1, ..., y_n)\rangle = \sum_{i=1}^{n} \langle x_i, y_i\rangle_i.$$

Demostración. Es claro que el producto cartesiano es un espacio vectorial.
Primero probaremos que $\langle\cdot,\cdot\rangle$ es un producto interior para H. Sean $x = (x_1, ..., x_n), y = (y_1, ..., y_n), z = (z_1, ..., z_n) \in H$, $\lambda \in \mathbb{R}$, entonces:

- ***IP1.***

$$\begin{aligned}\langle x+\lambda y, z\rangle &= \langle (x_1,...,x_n)+\lambda\,(y_1,...,y_n)\,,(z_1,...,z_n)\rangle \\ &= \langle (x_1+\lambda y_1,...,x_n+\lambda y_n)\,,(z_1,...,z_n)\rangle \\ &= \sum_{i=1}^{n}\langle x_i+\lambda y_i, z_i\rangle_i \\ &= \sum_{i=1}^{n}\langle x_i, z_i\rangle_i + \lambda\langle y_i, z_i\rangle_i \\ &= \sum_{i=1}^{n}\langle x_i, z_i\rangle_i + \lambda\sum_{i=1}^{n}\langle y_i, z_i\rangle_i \\ &= \langle x, z\rangle + \lambda\langle y, z\rangle.\end{aligned}$$

- ***IP2.***

$$\langle x, y\rangle = \sum_{i=1}^{n}\langle x_i, y_i\rangle_i = \sum_{i=1}^{n}\langle y_i, x_i\rangle_i = \langle y, x\rangle.$$

- ***IP3.***

$$\langle x, x\rangle = \sum_{i=1}^{n}\langle x_i, x_i\rangle_i = \sum_{i=1}^{n}\|x_i\|_i^2 > 0, \qquad \text{si } (x_1,...,x_n)\neq(0,...,0)\,.$$

Resta probar que sea completo, para ello notemos que $\forall x=(x_1,...,x_n)\in H$:

$$\|x_i\|_i^2 = \langle x_i, x_i\rangle_i \leq \sum_{j=1}^{n}\langle x_j, x_j\rangle_i = \|x\|^2.$$

Ahora consideremos una sucesión de Cauchy $(\mathbf{x}_k)_{k\in\mathbb{N}} = ((x_{1k},...,x_{nk}))_{k\in\mathbb{N}} \subseteq H$, entonces la sucesión $\{x_{ik}\}_{k\in\mathbb{N}} \subseteq H_i$ es de Cauchy, $\forall i\in\{1,...,n\}$, y por ser H_i espacio de Hilbert entonces existe $x_i\in H_i$ tal que:

$$\|x_{ik}-x_i\|_i \to 0.$$

Sea $\mathbf{x}=(x_1,...,x_n)$, entonces:

$$\|\mathbf{x}_k-\mathbf{x}\|^2 = \sum_{i=1}^{n}\langle x_{ik}-x_i, x_{ik}-x_i\rangle = \sum_{i=1}^{n}\|x_{ik}-x_i\|^2 \to 0,$$

de donde se concluye que toda sucesión de Cauchy es convergente, y por lo tanto H es completo. □

Proposición 1.1.8.
Sea $(V,\|\cdot\|)$. $\|\cdot\|$ *es inducida por un producto interior si y solo si se satisface que*

$$\|x+y\|^2+\|x-y\|^2 = 2\left(\|x\|^2+\|y\|^2\right), \quad \forall x,y\in V.$$

Demostración.

$\Longrightarrow$

Si $\|\cdot\|$ está inducido por un producto interior, entonces, $\forall x, y \in V$:

$$\begin{aligned}\|x+y\|^2 + \|x-y\|^2 &= \langle x+y, x+y\rangle + \langle x-y, x-y\rangle \\ &= \langle x,x\rangle + 2\langle x,y\rangle + \langle y,y\rangle + \langle x,x\rangle - 2\langle x,y\rangle + \langle y,y\rangle \\ &= 2\left(\langle x,x\rangle + \langle y,y\rangle\right) = 2\left(\|x\|^2 + \|y\|^2\right).\end{aligned}$$

$\Longleftarrow$

Definamos la función $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$, dada por $\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$.
P.d. $\langle\cdot,\cdot\rangle$ es un producto interior.
Notemos que $\langle\cdot,\cdot\rangle$ es continua por continuidad de la norma.

1. ***IP1.***
   Sean $x, y, z \in V$ y $\lambda \in \mathbb{R}$.
   a) **P.d.**$\langle x+y, z\rangle = \langle x, z\rangle + \langle y, z\rangle$
   Sea $b \in \{-1, 1\}$, notemos que de las hipótesis se sigue que:

$$\begin{aligned}-\frac{1}{2}\|x-y+bz\|^2 - \frac{1}{2}\|-x+y+bz\|^2 &= \left(-\left(\|x+bz\|^2 + \|y\|^2\right) + \frac{1}{2}\|x+y+bz\|^2\right) \\ &+ \left(-\left(\|y+bz\|^2 + \|x\|^2\right) + \frac{1}{2}\|x+y+bz\|^2\right) \\ &= -\|x\|^2 - \|y\|^2 - \|x+bz\|^2 - \|y+bz\|^2 \\ &+ \|x+y+bz\|^2,\end{aligned}$$

   tal que:

$$\|x+y+bz\|^2 = \|x\|^2+\|y\|^2+\|x+bz\|^2+\|y+bz\|^2-\frac{1}{2}\|x-y+bz\|^2-\frac{1}{2}\|-x+y+bz\|^2.$$

   Entonces:

$$\begin{aligned}\langle x+y, z\rangle &= \frac{1}{4}\left(\|x+y+z\|^2 - \|x+y-z\|^2\right) \\ &= \frac{1}{4}\left(\|x\|^2 + \|y\|^2 + \|x+z\|^2 + \|y+z\|^2 - \frac{1}{2}\|x-y+z\|^2 - \frac{1}{2}\|-x+y+z\|^2\right) \\ &- \frac{1}{4}\left(\|x\|^2 + \|y\|^2 + \|x-z\|^2 + \|y-z\|^2 - \frac{1}{2}\|x-y-z\|^2 - \frac{1}{2}\|-x+y-z\|^2\right) \\ &= \frac{1}{4}\left(\|x+z\|^2 + \|y+z\|^2 - \frac{1}{2}\|x-y+z\|^2 - \frac{1}{2}\|-x+y+z\|^2\right) \\ &- \frac{1}{4}\left(\|x-z\|^2 + \|y-z\|^2 - \frac{1}{2}\|-x+y+z\|^2 - \frac{1}{2}\|x-y+z\|^2\right) \\ &= \frac{1}{4}\left(\|x+z\|^2 + \|y+z\|^2\right) - \frac{1}{4}\left(\|x-z\|^2 + \|y-z\|^2\right) \\ &= \frac{1}{4}\left(\|x+z\|^2 - \|x-z\|^2\right) + \frac{1}{4}\left(\|y+z\|^2 - \|y-z\|^2\right) \\ &= \langle x, z\rangle + \langle y, z\rangle\end{aligned}$$

b) **P.D.**$\langle \lambda x, y\rangle = \lambda\langle x, y\rangle$

1) **Si** $\lambda = 0$.

$$\langle (0)x, y\rangle = \langle \mathbf{0}, y\rangle = \frac{1}{4}\left(\|\mathbf{0}+y\|^2 - \|\mathbf{0}-y\|^2\right) = 0 = (0)\langle x, y\rangle$$

2) **Si** $\lambda = -1$.

$$\langle -x, y\rangle = \frac{1}{4}\left(\|-x+y\|^2 - \|-x-y\|^2\right) = -\frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right) = -\langle x, y\rangle$$

3) **Si** $\lambda \in \mathbb{N}$.
Por inducción sobre λ, con la propiedad $\langle x+x, y\rangle = \langle x, y\rangle + \langle x, y\rangle$ como caso base es fácil concluir este caso.

4) **Si** $\lambda \in \mathbb{Q}^+$
Entonces existen $p, q \in \mathbb{N}$ tales que $\lambda = \frac{p}{q}$, de modo que:

$$q\langle \lambda x, y\rangle = q\langle \frac{p}{q}x, y\rangle = \langle q\frac{p}{q}x, y\rangle = \langle px, y\rangle = p\langle x, y\rangle.$$

$\therefore \langle \lambda x, y\rangle = \frac{p}{q}\langle x, y\rangle = \lambda\langle x, y\rangle$.

5) **Si** $\lambda \in \mathbb{R}^+$.
Entonces existe $(\lambda_n)_{n\in\mathbb{N}} \subseteq \mathbb{Q}^+$ tal que $\lambda_n \to \lambda$. Así, por continuidad:

$$\langle \lambda x, y\rangle = \langle \lim_{n\to\infty} \lambda_n x, y\rangle = \lim_{n\to\infty} \langle \lambda_n x, y\rangle = \lim_{n\to\infty} \lambda_n\langle x, y\rangle = \lambda\langle x, y\rangle.$$

$\therefore \langle \lambda x, y\rangle = \lambda\langle x, y\rangle$.

$\therefore$ Por los casos $\lambda = 0$, $\lambda = -1$ y $\lambda \in \mathbb{R}^+$ podemos concluir que $\langle \lambda x, y\rangle = \lambda\langle x, y\rangle, \quad \forall \lambda \in \mathbb{R}$.

2. ***IP2.***
Para toda $x, y \in V$:

$$\langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right) = \frac{1}{4}\left(\|y+x\|^2 - \|y-x\|^2\right) = \langle y, x\rangle.$$

3. ***IP3.***
Si $x \in V$, con $x \neq \mathbf{0}$:

$$\langle x, x\rangle = \frac{1}{4}\left(\|x+x\|^2 - \|x-x\|^2\right) = \frac{1}{4}\|2x\|^2 = \|x\|^2 > 0.$$

$\therefore \langle x, y\rangle = \frac{1}{4}\left(\|x+y\|^2 - \|x-y\|^2\right)$ es producto interior.
Finalmente,

$$\langle x, x\rangle = \frac{1}{4}\left(\|x+x\|^2 - \|x-x\|^2\right) = \frac{1}{4}\left(4\|x\|^2 - \|\mathbf{0}\|^2\right) = \|x\|^2, \quad \forall x \in V.$$

□

1.1.1. Complementos ortogonales y proyecciones

Definición 1.1.6.
Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial de H. Definimos el ***complemento ortogonal de*** *W* ***en*** *H como:*

$$W^{\perp} := \{x \in H : \forall w \in W, \langle w, x\rangle = 0\}.$$

Proposición 1.1.9.
Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial de H, entonces $W^{\perp}$ es un subespacio vectorial cerrado.

Demostración.

1. **P.d.** $W^{\perp}$ es subespacio vectorial.
   Sean $x, y \in W^{\perp}$ y $\lambda \in \mathbb{R}$, entonces:

   $$\langle w, x + \lambda y\rangle = \langle w, x\rangle + \lambda\langle w, y\rangle = 0 + \lambda(0) = 0, \quad \forall w \in W.$$

   $\therefore x + \lambda y \in W^{\perp}$.
   Por otro lado es claro que $\mathbf{0} \in W^{\perp}$, pues $\langle v, \mathbf{0}\rangle = 0, \quad \forall v \in W$.
2. **P.d.** $W^{\perp}$ es cerrado.
   Sea $(x_n)_{n\in\mathbb{N}} \subseteq W^{\perp}$ una sucesión tal que $x_n \to x$, para alguna $x \in H$. Entonces:

   $$\langle w, x\rangle = \langle w, \lim_{n\to\infty} x_n\rangle = \lim_{n\to\infty}\langle w, x_n\rangle = \lim_{n\to\infty} 0 = 0, \quad \forall w \in W.$$

   $\therefore x \in W^{\perp}$, y por lo tanto $W^{\perp}$ es cerrado.

□

Proposición 1.1.10.
([8], capítulo 15) Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial de H. Entonces

$$W^{\perp} = \{x \in H : \|x\| = \inf_{w\in W} \|x - w\|\}$$

Demostración.

1. **P.d.** $W^{\perp} \subseteq \{x \in H : \|x\| = \text{ínf}_{w\in W} \|x - w\|\}$.
   Sea $x \in W^{\perp}$, entonces:

   $$\|x-w\|^2 = \langle x-w, x-w\rangle = \langle x, x\rangle - \langle x, w\rangle - \langle w, x\rangle + \langle w, w\rangle = \|x\|^2 - 0 - 0 + \|w\|^2, \qquad \forall w \in W,$$

   tal que

   $$\|x-w\|^2 \geq \|x\|^2, \ \forall w \in W \quad \Longrightarrow \quad \|x-w\| \geq \|x\|, \ \forall w \in W \quad \Longrightarrow \quad \inf_{w\in W} \|x-w\| \geq \|x\|.$$

   Por otro lado,

   $$\|x\| = \|x - \mathbf{0}\| \geq \inf_{w\in W} \|x - w\|.$$

   $\therefore \quad \|x\| = \text{ínf}_{w\in W} \|x - w\|$.

2. **P.d.** $\{x \in H : \|x\| = \text{ínf}_{w\in W} \|x - w\|\} \subseteq W^{\perp}$.
Sea $x \in H$, tal que $\|x\| = \text{ínf}_{w\in W} \|x - w\|$. Fijemos $w_0 \in W$, con $\|w_0\| = 1$, y definamos la función $f : \mathbb{R} \to \mathbb{R}$, dada por

$$\begin{aligned} f(\lambda) = \|x - \lambda w_0\|^2 &= \langle x - \lambda w_0, x - \lambda w_0\rangle \\ &= \|x\|^2 - 2\lambda\langle x, w_0\rangle + \lambda^2\|w_0\|^2 = \|x\|^2 - 2\lambda\langle x, w_0\rangle + \lambda^2. \end{aligned}$$

Claramente f es diferenciable y convexa, con un único punto crítico (que es un mínimo global) en $\lambda = \langle x, w_0\rangle$. Por otro lado,

$$f(0) = \|x\|^2 = \inf_{w\in W} \|x - w\|^2 \leq \inf_{\lambda\in\mathbb{R}} \|x - \lambda w_0\|^2 = \inf_{\lambda\in\mathbb{R}} f(\lambda)$$

de manera que en 0 se encuentra el mínimo global de la función; entonces $\langle x, w_0\rangle = 0$. Lo que hemos probado lo hemos hecho para todo $w_0 \in W$, con $\|w_0\| = 1$, sin embargo, $\forall w \in W$ existen $w_0 \in W$, con $\|w_0\| = 1$, y $\lambda_0 \in \mathbb{R}$, tales que $w = \lambda_0 w_0$. Entonces:

$$\langle x, w\rangle = \lambda_0\langle x, w_0\rangle = 0 \quad \forall w \in W.$$

$\therefore x \in W^{\perp}$

□

Teorema 1.1.2 (Proyección Ortogonal sobre un subespacio vectorial cerrado)**.**
([8], capítulo 15) Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial cerrado de H. Entonces para todo $x \in H$ existe un único elemento $u \in W$, tal que:

$$\min_{w\in W} \|x - w\| = \|x - u\|$$

Demostración. Sea $x \in H$. Definimos la función $f : W \to \mathbb{R}$, dada por $f(w) = \|x - w\|^2$, que claramente es continua y no negativa, tal que el ínfimo $\alpha := \text{ínf}_{w\in W} f(w)$ existe (como número real no negativo).

Por definición de ínfimo, para toda $n \in \mathbb{N}$ existe $w_n \in W$ tal que $|\alpha - f(w_n)| < \frac{1}{n}$ y $\lim_{n\to\infty} f(w_n) = \alpha$. Aplicando la Proposición 1.1.8 para los vectores $(x - w_n)$, $(x - w_m)$ tenemos que:

$$\begin{aligned} \|w_m - w_n\|^2 &= \|(x - w_n) - (x - w_m)\|^2 \\ &= 2\left(\|(x - w_n)\|^2 + \|(x - w_m)\|^2\right) - \|(x - w_n) + (x - w_m)\|^2 \\ &= 2\left(\|(x - w_n)\|^2 + \|(x - w_m)\|^2\right) - 4\|x - \left(\frac{w_n + w_m}{2}\right)\|^2 \\ &\leq 2\left(f(w_n) + f(w_m)\right) - 4\alpha \\ &= 2\left(f(w_n) - \alpha\right) + 2\left(f(w_m) - \alpha\right) \\ &\leq \frac{2}{n} + \frac{2}{m}. \end{aligned}$$

De modo que la sucesión $(w_n)_{n\in\mathbb{N}} \subseteq W$ es de Cauchy, y, por ser W cerrado, entonces existe $u \in H$ tal que $w_n \to u$ en W. Más aún:

$$\|x-u\| = \sqrt{f\left(\lim_{n\to\infty} w_n\right)} = \sqrt{\lim_{n\to\infty} f(w_n)} = \sqrt{\alpha} = \inf_{w\in W} \|x-w\|$$

Así el *ínfimo* es en realidad un *mínimo*.
Para probar unicidad consideremos $\tilde{u} \in W$ tal que $\|x-\tilde{u}\| = \min_{w\in W} \|x-w\|$, por la Proposición 1.1.8, con los vectores $(x-\tilde{u})$, $(x-u)$, tenemos:

$$\begin{aligned}\|u-\tilde{u}\|^2 &= \|(x-u)-(x-\tilde{u})\|^2 \\ &= 2\left(\|x-u\|^2 + \|x-\tilde{u}\|^2\right) - 4\|x-\frac{\tilde{u}+u}{2}\|^2 \\ &\leq 2\left(\alpha+\alpha\right) - 4\alpha = 0.\end{aligned}$$

$\therefore \quad u = \tilde{u}.$ □

Definición 1.1.7 (Proyección Ortogonal)**.**
Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial cerrado de H. Definimos la ***proyección ortogonal de*** *H* ***sobre*** *W como la función $P_W : H \to W$, dada por:*

$$P_W(x) = \arg\min_{w\in W} \|x-w\|, \quad \forall x \in H.$$

Además, si $x \in W$ es claro que:

$$0 \leq \min_{w\in W} \|x-w\| \leq \|x-x\| = 0,$$

por lo que

$$P_W(x) = \arg\min_{w\in W} \|x-w\| = x, \quad \forall x \in W.$$

Corolario 1.1.2.1.
([8], capítulo 15) Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial cerrado de H. Entonces para toda $x \in H$, $P_W(x)$ es el único elemento de H que satisface $x - P_W(x) \in W^\perp$.

Demostración. Por la Proposición 1.1.10:

$$x - P_W(x) \in W^\perp \quad \Longleftrightarrow \quad \|x-P_W(x)\| = \inf_{w\in W} \|x-P_W(x)-w\| = \min_{w\in W} \|x-w\|$$

La última igualdad la tenemos porque $P_W(x) \in W$, y además tenemos la unicidad por el Teorema 1.1.2. □

Definición 1.1.8.
Sean $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$. Definimos $\mathcal{L}(V,W)$ como el ***espacio de funciones lineales y continuas en*** *V* ***con codominio en*** *W, dotado con la suma y el producto por escalar puntuales.*

Definición 1.1.9 (Norma de operadores)**.**
Sean $(V, \|\cdot\|_V)$, $(W, \|\cdot\|_W)$ y $\mathcal{L}(V,W)$. Entonces para el espacio vectorial $\mathcal{L}(V,W)$ definimos la norma como:

$$\|\phi\|_{\mathcal{L}(V,W)} = \sup_{\|x\|_V=1} \|\phi(x)\|_W, \quad \forall \phi \in \mathcal{L}(V,W).$$

Proposición 1.1.11.
Sea $f : V \to W$ una función lineal entre espacios vectoriales normados, entonces:

$$f \text{ es continua} \iff \exists M \in \mathbb{R}^+ \text{ tal que } \|f(x)\|_W \leq M, \quad \forall x \in V, \quad \|x\|_V \leq 1. \text{ (acotada)}$$

Demostración.
$\Longrightarrow$
Para $x = \mathbf{0}_V$ es claro que se cumple, entonces consideraremos $x \neq \mathbf{0}_V$.

Por ser f continua existe $\delta > 0$ tal que:

$$\|f(x)\|_W < 1, \quad \forall x \in V, \|x\|_V < \delta.$$

Sea $x \in V \setminus \{\mathbf{0}_V\}$, entonces existe $s > 0$ que satisface $\|sx\|_V < \delta$. Denotando $\delta' = \frac{\delta}{2}$:

$$s\|f(x)\|_W = s\|\frac{\|x\|_V}{\delta'} f\left(\frac{\delta' x}{\|x\|_V}\right)\|_W = \frac{s\|x\|_V}{\delta'}\|f\left(\frac{\delta' x}{\|x\|_V}\right)\|_W < \frac{s\|x\|_V}{\delta'},$$

lo que implica:

$$\|f(x)\|_W < \frac{\|x\|_V}{\delta'}, \quad \forall v \in V$$

$\therefore \|f(x)\|_W < \frac{1}{\delta'}, \quad \forall x \in V, \quad \|x\|_V \leq 1.$

$\Longleftarrow$
La hipótesis equivale a:

$$\|f(x)\|_W \leq M\|x\|_V, \quad \forall x \in V.$$

$\therefore f$ es Lipschitz continua y por lo tanto continua. □

Teorema 1.1.3.
([8], capítulo 15) Sean $(H, \langle\cdot,\cdot\rangle)$ y $W \leqslant H$ un subespacio vectorial cerrado de H. Entonces:

1. *La función P_W es elemento de $\mathcal{L}(H,W)$, idempotente y $P_W^{-1}\{\boldsymbol{0}_H\} = W^\perp$.*
   Más aún, si $W \neq \{\boldsymbol{0}_H\}$, entonces $\|P_H\|_{\mathcal{L}(H,W)} = 1$.
2. *La función $S : H \to W \bigoplus W^\perp$, dada por $S(x) = P_W(x) + (x - P_W(x))$ es un isomorfismo lineal e isométrico.*

Demostración.

1.

   a) $P_W \in \mathcal{L}(H,W)$.

1) *Linealidad:*
Sean $x, y \in H$ y $\lambda \in \mathbb{R}$, entonces por ser $W^\perp$ un subespacio vectorial:

$$x + \lambda y - (P_W(x) + \lambda P_W(y)) = (x - P_W(x)) + \lambda\,(y - P_W(y)) \in W^\perp.$$

Por la unicidad del Corolario 1.1.2.1:

$$(P_W(x) + \lambda P_W(y)) = P_W(x + \lambda y).$$

2) *Continuidad:*
Sea $x \in H$.

$$\begin{aligned}\|x\|^2 &= \|\,(x - P_W(x)) + P_W(x)\|^2 \\ &= \|x - P_W(x)\|^2 + \langle x - P_W(x), P_W(x)\rangle + \langle P_W(x), x - P_W(x)\rangle + \|P_W(x)\|^2 \\ &= \|x - P_W(x)\|^2 + 0 + 0 + \|P_W(x)\|^2 \\ &\geq \|P_W(x)\|^2\end{aligned}$$

$\therefore \quad \|P_W(x)\| \leq \|x\|$
$\therefore P_W$ es Lipschitz continua, y por lo tanto es continua.

b) *Idempotente:*

$$P_W^2(x) = P_W\,(P_W(x)) = P_W(x),$$

porque $P_W(x) \in W$.

c) $P_W^{-1}\{\mathbf{0}_H\} = W^\perp$.

1) Sea $x \in P_W^{-1}\{\mathbf{0}_H\}$.

$$\begin{aligned}x \in P_W^{-1}\{\mathbf{0}_H\} &\Longrightarrow P_W(x) = \mathbf{0}_H \\ &\Longrightarrow x \in W^\perp, \quad \text{por el Corolario 1.1.2.1.}\end{aligned}$$

2) Sea $x \in W^\perp$, entonces por el Corolario 1.1.2.1 $P_W(x) = \mathbf{0}_H$.
$\therefore x \in P_W^{-1}\{\mathbf{0}_H\}$.

d) Si $W \neq \{\mathbf{0}_H\}$, entonces existe $x \in W$, con $\|x\| = 1$, de modo que $P_W(x) = x$ y por lo tanto:

$$\|P_W(x)\| = \|x\| = 1.$$

Por otro lado, sabemos que la función es Lipschitz continua, con constante de Lipschitz 1, por lo que:

$$\|P_H\|_{\mathcal{L}(H,W)} = \sup_{\|x\|=1} \|P_W(x)\| \leq 1.$$

$\therefore \quad \|P_H\|_{\mathcal{L}(H,W)} = 1.$

2. **Lineal, biyectiva, isométrica:**
Dado que $S(x) = P_W(x) + (x - P_W(x)) = x$, $\forall x \in H$, entonces S es lineal y biyectiva. Más aún, $\|S(x)\| = \|x\|$, $\forall x \in H$, de modo que S es isométrica.

□

Observación 1.1.1.
Del teorema anterior podemos concluir que:

$$H = W \bigoplus W^{\perp}, \quad \forall W \leqslant H, \quad W \text{ cerrado.}$$

Definición 1.1.10.
*Sea $(V, \|\cdot\|)$. Definimos el **espacio dual** de V como:*

$$V^* := \{f : V \to \mathbb{R} | f \text{ lineal y continua}\},$$

el cual es un espacio vectorial normado con la norma de operadores.

Teorema 1.1.4 (Teorema de Representación de Riesz)**.**
([5], capítulo 5) Sea $(H, \langle\cdot,\cdot\rangle)$, entonces para todo $\phi \in H^*$ existe un único elemento $u_\phi \in H$ tal que:

$$\phi(x) = \langle x, u_\phi\rangle, \quad \forall x \in H.$$

Más aún, $\|\phi\|_{H^*} = \|u_\phi\|_H$.

Demostración. Sea $\phi \in H^*$. Si $\phi \equiv 0$ es claro que $u_\phi = \mathbf{0}_H$ y $\|\phi\|_{H^*} = \|u_\phi\|_H$, entonces supongamos que $\phi \neq 0$.

1. **P.d.**$\exists!$ $u_\phi \in H$ tal que $\phi(x) = \langle x, u_\phi\rangle$, $\forall x \in H$.

   $K := \phi^{-1}\{0\}$ es un subespacio vectorial de H, cerrado y distinto de H. En particular $K^\perp \neq \{\mathbf{0}_H\}$.
   Sean $x \in H$, $y \in K^\perp$, con $\|y\|_H = 1$, y definamos $u = \phi(x)y - \phi(y)x$, tal que $u \in K$, y más aún:

   $$\begin{aligned} 0 = \langle u, y\rangle = \langle \phi(x)y - \phi(y)x, y\rangle &= \phi(x)\langle y, y\rangle - \phi(y)\langle x, y\rangle \\ &= \phi(x)\|y\|_H^2 - \phi(y)\langle x, y\rangle = \phi(x) - \langle x, \phi(y)y\rangle. \end{aligned}$$

   $\therefore \phi(x) = \langle x, \phi(y)y\rangle, \quad \forall x \in H$, y $u_\phi = \phi(y)y$.

   Unicidad
   Sea $\tilde{u} \in H$, tal que $\phi(x) = \langle x, \tilde{u}\rangle$, $\forall x \in H$, entonces:

   $$0 = \phi(x) - \phi(x) = \langle x, u_\phi\rangle - \langle x, \tilde{u}\rangle = \langle x, u_\phi - \tilde{u}\rangle, \quad \forall x \in H.$$

   $\therefore u_\phi = \tilde{u}$.

2. **P.d.** $\|\phi\|_{H^*} = \|u_\phi\|_H$.

   $$\|\phi\|_{H_*} = \sup_{\|x\|_H=1} |\phi(x)| = \sup_{\|x\|_H=1} |\langle x, u_\phi\rangle| \leq \|u_\phi\|_H.$$

   Por otro lado, sabemos que $u_\phi = \phi(y)y$, con $\|y\|_H = 1$, tal que:

   $$\|u_\phi\|_H = \|\phi(y)y\|_H = |\phi(y)| \, \|y\|_H = |\phi(y)| \, .$$

   $\therefore \quad \|\phi\|_{H^*} = \|u_\phi\|_H$.

□

1.2. Espacios de Hilbert de Kernel Reproductor

Definición 1.2.1 (Kernel Definido Positivo, en el sentido de Moore)**.**
Sean X un conjunto distinto del vacío y $K : X \times X \to \mathbb{R}$ una función. Decimos que K es un ***kernel definido positivo*** *si para todo subconjunto finito de X, digamos $\{x_1, ...x_n\} \subseteq X$, la matriz $M \in \mathbb{R}^{n\times n}$ dada por $(M_{ij}) = K(x_i, x_j)$ es semidefinida positiva.*

A la matriz M de la definición anterior la llamamos **Matriz de Gram** para $\{x_1, ...x_n\} \subseteq X$.

Cada que hablemos de X nos referiremos a un conjunto distinto del vacío.

Definición 1.2.2 (Espacio de Hilbert de Kernel Reproductor (RKHS))**.**
Sean X un conjunto y $\mathbb{R}^X$ el espacio vectorial de funciones reales sobre X, equipado con la suma y el producto escalar puntuales. Un subespacio vectorial $H \leqslant \mathbb{R}^X$ es un ***RKHS*** *si satisface lo siguiente:*

1. *Esta equipado con un producto interior, $\langle \cdot, \cdot \rangle_H$, que lo hace un espacio de Hilbert.*
2. *Para cada $x \in X$, la función $Ev_x : H \to \mathbb{R}$, dada por $Ev_x(f) = f(x), \quad \forall f \in H$, es acotada.*

Denotaremos por Ev_X al espacio vectorial de las funciones cuyos elementos son $Ev_x, \forall x \in X$, y siempre que hablemos de un RKHS, H, denotaremos con X al dominio de las funciones en H.

Notemos que la linealidad de los elementos de Ev_X se da por la definición de la suma y el producto por escalar puntuales.

Definición 1.2.3 (Kernel Reproductor)**.**
Sea $(H, \langle \cdot, \cdot \rangle)$ un RKHS, entonces, por el Teorema de Representación de Riesz, para toda $Ev_x \in Ev_X$ existe una única $K_x \in H$ tal que:

$$Ev_x(f) = \langle f, K_x \rangle, \quad \forall f \in H.$$

Definimos al ***kernel reproductor de*** *H como la función $K : X \times X \to \mathbb{R}$, dada por $K(x, y) = \langle K_x, K_y \rangle$. Más aún,*

$$K_x(y) = Ev_y(K_x) = \langle K_x, K_y \rangle = K(x, y), \quad \forall x \in X.$$

El adjetivo *reproductor* viene del hecho de que para toda $f \in H$ se cumple que $f(x) = \langle f, K_x \rangle = \langle f(\cdot), K(x, \cdot) \rangle$, para toda $x \in X$. Llamamos a esta cualidad *propiedad reproductora.*

Proposición 1.2.1.
Para todo RKHS, su Kernel Reproductor es un kernel definido positivo, simétrico y único.

Demostración. Sea $(H, \langle \cdot, \cdot \rangle)$ un RKHS con kernel reproductor K. Es claro que K es simétrico por ser el producto interior simétrico. Veamos que K es un kernel definido positivo.

Sean $\{x_1, ...x_n\} \subseteq X$, $v = (a_1, ..., a_n) \in \mathbb{R}^n$ y $M \in \mathbb{R}^{n\times n}$ dada por $(M_{ij}) = K(x_i, x_j)$, entonces:

$$\begin{aligned} v^T M v &= \sum_{i=0}^{n}\sum_{j=0}^{n} a_i a_j K(x_i, x_j) = \sum_{i=0}^{n}\sum_{j=0}^{n} a_i a_j \langle K_{x_i}, K_{x_j}\rangle \\ &= \sum_{i=0}^{n}\sum_{j=0}^{n} \langle a_i K_{x_i}, a_j K_{x_j}\rangle = \langle \sum_{i=0}^{n} a_i K_{x_i}, \sum_{j=0}^{n} a_j K_{x_j}\rangle \\ &= \| \sum_{j=0}^{n} a_j K_{x_j}\|^2 \geq 0. \end{aligned}$$

$\therefore K$ es un kernel definido positivo y simétrico.
Por otra parte, suponiendo que existe otro kernel reproductor de H, digamos $\tilde{K}$, tenemos que para toda $x_1, x_2 \in X$:

$$\begin{aligned} K(x_1, x_2) = Ev_{x_2}(K_{x_1}) &= \langle K_{x_1}, \tilde{K}_{x_2}\rangle = \langle \tilde{K}_{x_2}, K_{x_1}\rangle \\ &= Ev_{x_1}(\tilde{K}_{x_2}) = \langle \tilde{K}_{x_2}, \tilde{K}_{x_1}\rangle = \langle \tilde{K}_{x_1}, \tilde{K}_{x_2}\rangle = \tilde{K}(x_1, x_2). \end{aligned}$$

$\therefore K$ es único. □

1.2.1. Completación de espacios pre-Hilbert y RKHS

Un espacio pre-Hilbert es un espacio vectorial que cumple todo para ser un espacio de Hilbert, excepto quizá la completez. Ahora veremos que es posible pasar de un espacio pre-hilbert de funciones a un RKHS bajo ciertas condiciones.

Sabemos que el producto interior induce una norma, que a su vez induce una métrica, por lo que un espacio pre-Hilbert incompleto es un espacio métrico incompleto, y en análisis tenemos un resultado para completar espacios métricos considerando a nuestro espacio métrico incompleto V como un subespacio vectorial W de un espacio vectorial Z, mediante un isomorfismo isométrico, en donde W es un espacio cociente y sus elementos son clases de equivalencia.

Sin embargo, recordemos que buscamos construir RKHSs, para los cuales una propiedad importante es la continuidad de las funciones en Ev_X, lo cual no se puede garantizar cuando tenemos clases de equivalencia en lugar de funciones.

Para nuestros fines lo ideal sería poder completar un espacio vectorial pre-Hilbert incompleto, digamos V, de funciones reales sobre un conjunto X, dentro de $\mathbb{R}^X$, es decir, solamente agregando más funciones de $\mathbb{R}^X$. El siguiente teorema es útil para este fin pues nos da condiciones suficientes y necesarias para que dado un subespacio vectorial pre-Hilbert $V \leqslant \mathbb{R}^X$, exista una completación de V en $\mathbb{R}^X$ que sea un RKHS.

Teorema 1.2.1.
([2]) Sea $V \leqslant \mathbb{R}^X$ un subespacio vectorial pre-Hilbert. Entonces existe una completación $\tilde{V}$ de V, $V \leqslant \tilde{V} \leqslant \mathbb{R}^X$, tal que $\tilde{V}$ es un RKHS, si y solo si V satisface:

1. *$\forall x \in X$, la función $Ev_x : V \to \mathbb{R}$, dada por $Ev_x(f) = f(x)$, es acotada.*
2. *Dada $(f_n)_{n\in\mathbb{N}} \subseteq V$ una sucesión de Cauchy, si $(f_n)_{n\in\mathbb{N}}$ converge puntualmente a cero, entonces $\|f_n\| \to 0$.*

Y además, si tal completación existe, entonces es única.

Demostración.
$\Longrightarrow$
Si la completación existe es claro que la condición 1 se satisface, pues $V \leqslant \tilde{V}$.
Por otro lado, si tenemos una sucesión $(f_n)_{n\in\mathbb{N}} \subseteq V$ de Cauchy tal que $(f_n)_{n\in\mathbb{N}}$ converge puntualmente a cero, entonces por ser de Cauchy existe $f \in \tilde{V}$ tal que $f_n \to f$ en $\tilde{V}$, y más aún $f \equiv 0$, pues:

$$|f_n(x) - f(x)| \leq M_x \|f_n - f\| \to 0, \quad \forall x \in X$$
$$\Longrightarrow 0 = \lim_{n\to\infty} f_n(x) = f(x), \quad \forall x \in X,$$

donde M_x es la cota que obtenemos por ser $Ev_x(\cdot)$ acotada, tal que depende directamente de la x que tomemos.

Entonces:

$$\|f_n\| = \|f_n - 0\| = \|f_n - f\| \to 0.$$

$\Longleftarrow$
Notemos que, dada $(f_n)_{n\in\mathbb{N}} \subseteq V$ una sucesión de Cauchy, por el inciso 1 tenemos que $(f_n(x))_{n\in\mathbb{N}} \subseteq \mathbb{R}$ es una sucesión de Cauchy para toda $x \in X$, de manera que existe $f \in \mathbb{R}^X$ que es el límite puntual de $(f_n)_{n\in\mathbb{N}}$, $f(x) = \lim_{n\to\infty} f_n(x)$, $\forall x \in X$.

Sea $\tilde{V} \subseteq \mathbb{R}^X$ el conjunto de funciones que son límite puntual de sucesiones de Cauchy en V; es claro que $V \subseteq \tilde{V}$. Consideremos $\tilde{V}$ con la norma $\|\cdot\|_0 : \tilde{V} \to \mathbb{R}$, dada por $\|f\|_0 = \lim_{n\to\infty} \|f_n\|$, donde $(f_n)_{n\in\mathbb{N}} \subseteq V$ es una sucesión de Cauchy que converge puntualmente a f. Es claro que $\tilde{V}$ es un espacio vectorial, pues dadas $(g_n)_{n\in\mathbb{N}}, (f_n)_{n\in\mathbb{N}} \subseteq V$ sucesiones de Cauchy, $\lambda \in \mathbb{R}$, si $f_n \to f$, $g_n \to g$ puntualmente, entonces $f_n + g_n \to f + g$, $\lambda f_n \to \lambda f$ puntualmente.

Para probar que $\|\cdot\|_0$ esta bien definida, es decir, que no depende de la sucesión de Cauchy que tomemos, sean $(g_n)_{n\in\mathbb{N}}, (f_n)_{n\in\mathbb{N}} \subseteq V$ sucesiones de Cauchy que convergen puntualmente a una misma función $f \in \mathbb{R}^X$, entonces las sucesiones $(f_n - f)_{n\in\mathbb{N}}$ y $(g_n - f)_{n\in\mathbb{N}}$ convergen puntualmente a cero, lo que a su vez implica que $(f_n - f - (g_n - f))_{n\in\mathbb{N}} = (f_n - g_n)_{n\in\mathbb{N}}$ converge puntualmente a cero. Así, por el inciso 2 tenemos que $\|f_n - g_n\| \to 0$, de modo que:

$$\left|\lim_{n\to\infty} \|f_n\| - \lim_{n\to\infty} \|g_n\|\right| = \lim_{n\to\infty} |\|f_n\| - \|g_n\|| \leq \lim_{n\to\infty} \|f_n - g_n\| \to 0.$$

$\therefore \|f\|_0 \lim_{n\to\infty} \|f_n\| = \lim_{n\to\infty} \|g_n\|$, de manera que la función $\|\cdot\|_0$ esta bien definida para toda $f \in \tilde{V}$, y más aún, $\|f\|_0 = \|f\|, \quad \forall f \in V$.

Ahora probaremos que es norma. Sean $f, g \in \tilde{V}$, $\lambda \in \mathbb{R}$ y $(f_n)_{n\in\mathbb{N}}, (g_n)_{n\in\mathbb{N}} \subseteq V$ sucesiones de Cauchy que convergen puntualmente a f y g, respectivamente. Entonces:

- ***N1.***

$$0 = \|f\|_0 = \lim_{n\to\infty} \|f_n\|.$$

  Por el inciso 1 se satisface:

$$0 \leq |f(x)| \leq \lim_{n\to\infty} |f_n(x)| \leq M_x \lim_{n\to\infty} \|f_n\| = 0, \quad \forall x \in X.$$

Entonces:

$$0 = \|f\|_0 \iff f \equiv 0$$

- ***N2.***

$$\begin{aligned}\|\lambda f\|_0 &= \lim_{n\to\infty} \|\lambda f_n\| = \lim_{n\to\infty} |\lambda| \, \|f_n\| \\ &= |\lambda| \lim_{n\to\infty} \|f_n\| = |\lambda| \, \|f\|_0.\end{aligned}$$

- ***N3.***

$$\|f+g\|_0 = \lim_{n\to\infty} \|f_n + g_n\| \leq \lim_{n\to\infty} (\|f_n\| + \|g_n\|) = \lim_{n\to\infty} \|f_n\| + \lim_{n\to\infty} \|g_n\| = \|f\|_0 + \|g\|_0.$$

$\therefore \|\cdot\|_0$ es una norma de $\tilde{V}$.

En lo que resta de la prueba seguiremos con la notación que ocupamos para probar que $\|\cdot\|_0$ es norma, a menos que se indique lo contrario.

Ahora demostraremos que $\|\cdot\|_0$ esta inducida por un producto interior, lo que equivale a probar que satisface la identidad del paralelogramo, por la Proposición 1.1.8. Entonces usando que $\|\cdot\|$ es una norma para V inducida por un producto interior (por ser V pre-Hilbert), tenemos:

$$\begin{aligned}\|f+g\|_0^2 + \|f-g\|_0^2 &= \left(\lim_{n\to\infty} \|f_n + g_n\|\right)^2 + \left(\lim_{n\to\infty} \|f_n - g_n\|\right)^2 \\ &= \lim_{n\to\infty} \left(\|f_n + g_n\|^2 + \|f_n - g_n\|^2\right) \\ &= \lim_{n\to\infty} \left(2\left(\|f_n\|^2 + \|g_n\|^2\right)\right) \\ &= 2\left(\left(\lim_{n\to\infty} \|f_n\|\right)^2 + \left(\lim_{n\to\infty} \|g_n\|\right)^2\right) \\ &= 2\left(\|f_n\|_0^2 + \|g_n\|_0^2\right)\end{aligned}$$

$\therefore \|\cdot\|_0$ es una norma inducida en $\tilde{V}$ por un único producto interior, y dado que $\|f\|_0 = \|f\|, \quad \forall f \in V$, entonces $\langle\cdot,\cdot\rangle_0|_V = \langle\cdot,\cdot\rangle$, donde $\langle\cdot,\cdot\rangle$ es el producto interior que induce a la norma $\|\cdot\|$ en V.

Por otro lado, notemos que $\|f_{m_0} - f_{m_1}\| = \|\,(f_n - f_{m_1}) - (f_n - f_{m_0})\,\|$ tal que para toda $n \in \mathbb{N}$ la sucesión $(f_n - f_m)_{m\in\mathbb{N}} \subseteq V$ es de Cauchy y $f_n - f_m \to f_n - f$ puntualmente, por lo tanto:

$$\lim_{n\to\infty} \|f_n - f\|_0 = \lim_{n\to\infty} \lim_{m\to\infty} \|f_n - f_m\| = 0,$$

tal que la convergencia puntual implica convergencia en norma (convergencia fuerte), en nuestra situación particular, tal que V es denso en $\tilde{V}$.

Para probar que $\left(\tilde{V}, \langle\cdot,\cdot\rangle_0\right)$ es completación de V, sea $\left(\tilde{f}_n\right)_{n\in\mathbb{N}} \subseteq \tilde{V}$ una sucesión de Cauchy. Por ser V denso en $\tilde{V}$, para toda $n \in \mathbb{N}$ existe $f_n \in V$ tal que $\|f_n - \tilde{f}_n\|_0 < \frac{1}{n}$.

Sea $\epsilon > 0$, dado que $\left(\tilde{f}_n\right)_{n\in\mathbb{N}}$ es de Cauchy $\exists N_0 \in \mathbb{N}$ tal que $\|\tilde{f}_n - \tilde{f}_m\|_0 < \frac{\epsilon}{3}, \quad \forall n, m \geq N_0$, y por la definición de la sucesión $(f_n)_{n\in\mathbb{N}}$ existe $N_1 \in \mathbb{N}$ tal que $\|f_n - \tilde{f}_n\|_0 < \frac{\epsilon}{3}, \quad \forall n \geq N_1$. Definimos $N = \text{máx}\{N_0, N_1\}$, entonces:

$$\|f_n - f_m\| = \|f_n - f_m\|_0 \leq \|f_n - \tilde{f}_n\|_0 + \|\tilde{f}_n - \tilde{f}_m\|_0 + \|\tilde{f}_m - f_m\|_0 < \epsilon, \quad \forall n, m \geq N,$$

tal que $(f_n)_{n\in\mathbb{N}} \subseteq V$ es de Cauchy, y por lo tanto $f_n \to f$ puntualmente, para alguna $f \in \tilde{V}$, pero esto a su vez implica, por lo demostrado un par de párrafos antes, que $\|f_n - f\|_0 \to 0$. Entonces:

$$\|\tilde{f}_n - f\|_0 \leq \|\tilde{f}_n - f_n\|_0 + \|f_n - f\|_0 \to 0,$$

a partir de lo cual podemos concluir que $\tilde{V}$ es completo.

Además, notemos que para toda $x \in X$:

$$|f(x)| = \lim_{n\to\infty} |f_n(x)| \leq \lim_{n\to\infty} M_x \|f_n\| = M_x \lim_{n\to\infty} \|f_n\| = M_x \|f\|_0, \quad \forall f \in \tilde{V}.$$

Con esto concluimos que $\tilde{V}$ es una completación de V en $\mathbb{R}^X$, y que además es un RKHS.

Finalmente, para probar la unicidad de $\tilde{V}$, sea $W \subseteq \mathbb{R}^X$ otra completación de V que es un RKHS con producto interior $\langle\cdot,\cdot\rangle_W$, que es extensión de $\langle\cdot,\cdot\rangle$.
Si $f \in W$, entonces existe $(f_n)_{n\in\mathbb{N}} \subseteq V$ de Cauchy tal que $\lim_{n\to\infty} \|f - f_n\|_W \to 0$, y por ser W un RKHS:

$$0 \leq \lim_{n\to\infty} |f(x) - f_n(x)| \leq M_x \lim_{n\to\infty} \|f - f_n\|_W \to 0,$$

de modo que f es el límite puntual de $(f_n)_{n\in\mathbb{N}} \subseteq V$ y por lo tanto $f \in \tilde{V}$.
$\therefore V \subseteq W \subseteq \tilde{V}$, y dado que V es denso en $\tilde{V}$, y $\tilde{V}$ es cerrado, entonces $\tilde{V} = W$. □

Teorema 1.2.2 (Teorema de Moore–Aronszajn)**.**
([2]) Sea X un conjunto. Para toda función $K : X \times X \to \mathbb{R}$ que satisfaga ser un kernel definido positivo y simétrico existe un único RKHS, $H \subseteq \mathbb{R}^X$, del cual K es kernel reproductor.

Demostración. Sea $K : X \times X \to \mathbb{R}$ un kernel definido positivo y simétrico.
Para toda $x \in X$ definimos la función $K_x : X \to \mathbb{R}$ dada por $K_x(y) = K(x, y)$, $\forall y \in X$, y denotamos con H_0 al espacio *span* $(\{K_x\}_{x\in X})$, y definimos la función $\langle\cdot,\cdot\rangle_0 : H_0 \times H_0 \to \mathbb{R}$, dada por:

$$\langle f, g\rangle_0 = \langle \sum_{j=1}^{n} a_j K_{x_j}, \sum_{i=1}^{m} b_i K_{x'_i} \rangle_0 = \sum_{j=1}^{n}\sum_{i=1}^{m} a_j b_i \langle K_{x_j}, K_{x'_i}\rangle_0 = \sum_{j=1}^{n}\sum_{i=1}^{m} a_j b_i K(x_j, x'_i), \quad \forall f, g \in H_0,$$

con $f = \sum_{j=1}^{n} a_j K_{x_j}$, $g = \sum_{i=1}^{m} b_i K_{x'_i}$. En lo subsecuente usaremos estas definiciones de f y g para referirnos a cualesquiera elementos f, g de H_0.

Notemos que la expansión de f y g no necesariamente es única, sin embargo:

$$\langle f, g\rangle_0 = \sum_{j=1}^{n} a_j \sum_{i=1}^{m} b_i K(x'_i, x_j) = \sum_{j=1}^{n} a_j g(x_j)$$

$$\langle f, g\rangle_0 = \sum_{i=1}^{m} b_i \sum_{j=1}^{n} a_j K(x_j, x_i') = \sum_{i=1}^{m} b_i f(x_i'),$$

tal que $\langle f, g\rangle_0$ no depende de las expansiones de g y f, de modo que esta bien definido.

P.d. $\langle \cdot, \cdot \rangle_0$ es un producto interior.

1. ***IP1*** y ***IP2***.
   La linealidad en la primer entrada y que sea simétrica se siguen de la definición de $\langle \cdot, \cdot \rangle_0$ y K, respectivamente.

2. ***IP3***.
   Notemos que para toda $x \in X$ y para toda $f \in H_0$:

   $$f(x) = \sum_{i=1}^{n} a_i K_{x_i}(x) = \sum_{i=1}^{n} a_i K(x_i, x) = \langle \sum_{i=1}^{n} a_i K_{x_i}, K_x \rangle = \langle f, K_x \rangle,$$

   de manera que se satisface la propiedad reproductiva.

   Por otro lado, para toda $f = \sum_{i=1}^{n} a_i K_{x_i} \in H_0$:

   $$\langle f, f\rangle_0 = \langle \sum_{i=1}^{n} a_i K_{x_i}, \sum_{i=1}^{n} a_i K_{x_i} \rangle_0 = \sum_{j=1}^{n} \sum_{i=1}^{n} a_j a_i K(x_j, x_i) = v^T M v \geq 0,$$

   donde $(M_{ij}) = K(x_i, k_j)$ es la matriz de Gram correspondiente a $A = \{x_1, ..., x_n\} \subseteq X$, que es semidefinida positiva, y $v^T = (a_1, ..., a_n)$.

   $\therefore \langle \cdot, \cdot \rangle_0$ es un semi-producto interior para H_0.
   Y si $\langle f, f\rangle_0 = 0$, entonces por la Proposición 1.1.2:

   $$|f(x)| = |\langle f, K_x\rangle_0| \leq \sqrt{\langle K_x, K_x\rangle_0}\sqrt{\langle f, f\rangle_0}, \quad \forall x \in X,$$

   tal que $f \equiv 0$.
   $\therefore \langle f, f\rangle_0 \neq 0$, si $f \neq 0$.

$\therefore (H_0, \langle \cdot, \cdot \rangle_0)$ es un espacio vectorial con producto interior.

Ahora probaremos que se satisfacen las condiciones del Teorema 1.2.1, para poder tener por completación un RKHS.

1. Por la propiedad reproductiva y por Cauchy:

   $$|f(x)| = |\langle f, K_x\rangle_0| \leq \|f\|_0 \|K_x\|_0, \quad \forall x \in X, \quad \forall f \in H_0,$$

   tal que la evaluación funcional es acotada, para toda $x \in X$.

2. Sea $(f_n)_{n\in\mathbb{N}} \subseteq H_0$ de Cauchy tal que converge puntualmente a cero, entonces por ser $(f_n)_{n\in\mathbb{N}}$ de Cauchy, existe $M > 0$ tal que $\|f_n\| < M$, para toda $n \in \mathbb{N}$. Por otro lado, sea $\epsilon > 0$, entonces por ser de Cauchy $\exists N \in \mathbb{N}$ tal que:

   $$\|f_n - f_N\|_0 < \frac{\epsilon}{M}, \quad \forall n \geq N,$$

donde f_N es de la forma $\sum_{i=1}^{k} a_i K_{x_i}$. Esto implica que:

$$\begin{aligned}
\|f_n\|_0^2 &= \langle f_n, f_n - f_N + f_N\rangle_0 \\
&= \langle f_n, f_n - f_N\rangle_0 + \langle f_n, f_N\rangle_0 \\
&\leq |\langle f_n, f_n - f_N\rangle_0| + \langle f_n, \sum_{i=1}^{k} a_i K_{x_i}\rangle_0 \\
&\leq \|f_n\|_0 \|f_n - f_N\|_0 + \sum_{i=1}^{k} a_i \langle f_n, K_{x_i}\rangle \\
&\leq M\frac{\epsilon}{M} + \sum_{i=1}^{k} a_i f_n(x_i) \\
&= \epsilon + \sum_{i=1}^{k} a_i f_n(x_i)
\end{aligned}$$

para toda $n \geq N$, por lo que:

$$\lim_{n\to\infty} \sup \|f_n\|_0 \leq \epsilon + \lim_{n\to\infty} \sup \sum_{i=1}^{k} a_i f_n(x_i) = \epsilon + \sum_{i=1}^{k} a_i \lim_{n\to\infty} f_n(x_i) = \epsilon,$$

$\therefore \forall \epsilon > 0, \quad 0 \leq \lim\sup_{n\to\infty} \|f_n\|_0 \leq \epsilon.$
$\therefore \|f_n\|_0 \to 0.$

Entonces se satisfacen las dos condiciones suficientes para que exista una única completación H de H_0 en $\mathbb{R}^X$, y, por continuidad del producto interior $\langle\cdot,\cdot\rangle_0$, K es kernel reproductor de H, pues H_0 es denso en H.

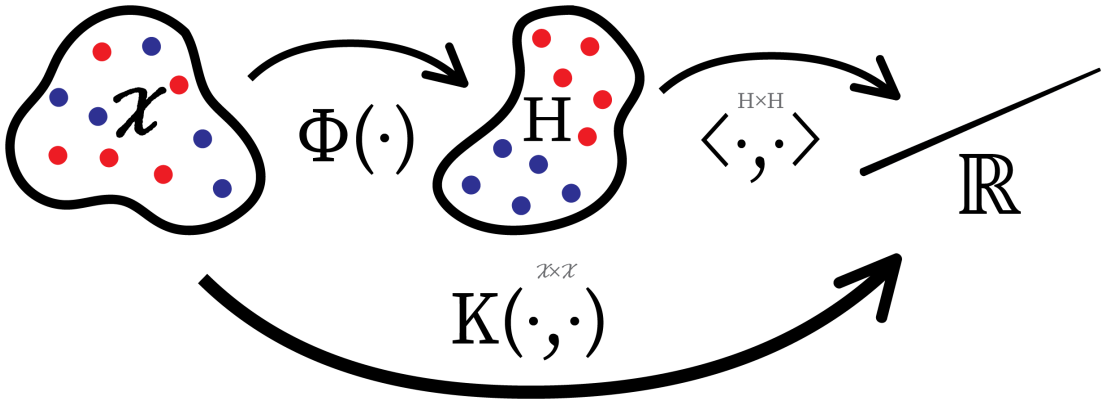
Finalmente, para probar la unicidad supongamos que existe $(H', \langle\cdot,\cdot\rangle_{'})$ un RKHS del cual K es kernel reproductor y $H' \neq H$, entonces $span\,(\{K_x\}_{x\in X}) \subseteq H'$ tal que $H \subseteq H'$, así por el Teorema 1.1.3, $H' = H \bigoplus H^{\perp}$, con $H^{\perp} \neq \{\mathbf{0}\}$.
Por otro lado, sea $f \in H^{\perp} \setminus \{\mathbf{0}\}$, tal que existe $x \in X$ para el cual $f(x) \neq 0$, pero por la propiedad reproductiva:

$$f(x) = \langle f, K_x\rangle_{'} =^{1} 0,$$

lo cual es una contradicción.
$\therefore$ Existe un único RKHS cuyo kernel reproductor es K. □

Por este teorema tenemos que dado un conjunto X, existe una biyección entre los RKHSs (en X) y los kernels definidos positivos (con dominio $X \times X$) y simétricos.

[1] $K_x \in H$



Más aún, dado X un conjunto y $K : X \times X \to \mathbb{R}$ un kernel definido positivo, entonces existen H un espacio de Hilbert (RKHS) y $\Phi : X \to H$ una función, no necesariamente única, tales que:

$$K(x, x') = \langle \Phi(x), \Phi(x') \rangle_H, \qquad \forall x, x' \in X.$$

Así, de manera implícita un kernel definido positivo nos permite operar conjuntos en espacios de Hilbert.

1.2.2. RKHS de funciones vectoriales

Toda la teoría que hemos desarrollado hasta ahora se puede extender para espacios vectoriales de funciones vectoriales. Brevemente introduciremos algunas definiciones y resultados útiles para nuestro propósito. Consideraremos funciones vectoriales con valores en $\mathbb{R}^d$, sin embargo, los resultados pueden enunciarse para un espacio de Hilbert arbitrario en lugar de $\mathbb{R}^d$.

Definición 1.2.4 (RKHS con valores en $\mathbb{R}^d$)**.**
Sean X un conjunto y $\mathcal{F}(X, \mathbb{R}^d)$ el espacio vectorial de funciones sobre X con codominio $\mathbb{R}^d$, equipado con la suma y el producto por escalar puntuales. Un subespacio vectorial $H \leqslant \mathcal{F}(X, \mathbb{R}^d)$ es un RKHS si satisface lo siguiente:

1. *Esta equipado con un producto interior, $\langle \cdot, \cdot \rangle_H$, que lo hace un espacio de Hilbert.*
2. *Para cada $x \in X$, la función $Ev_x : H \to \mathbb{R}^d$, dada por $Ev_x(f) = f(x), \quad \forall f \in H$, es acotada.*

Definición 1.2.5 (Kernel Reproductor de un RKHS con valores en $\mathbb{R}^d$)**.**
*Sea H un RKHS con valores en $\mathbb{R}^d$. Su **kernel reproductor**, $K : X \times X \to \mathcal{B}(\mathbb{R}^d) \cong M_d(\mathbb{R})$, esta dado por $K(x, x') = Ev_x Ev_{x'}^*$.*

$Ev_{(\cdot)}^*$ denota al operador adjunto, que también es acotado.

Definición 1.2.6 (Kernel Definido Positivo, en el sentido de Moore)**.**
*Dada una función $K : X \times X \to \mathcal{B}(\mathbb{R}^d)$, decimos que K es un **kernel definido positivo** si para todo subconjunto finito de X, digamos $\{x_1, ...x_n\} \subseteq X$, la matriz (de Gram), $(K(x_i, x_j)) \in M_n(\mathcal{B}(\mathbb{R}^d))$. es semidefinida positiva.*

Notemos que la condición de ser definida positiva podemos reescribirla de manera más sencilla de la siguiente forma:

- Sea $\{x_1, ...x_n\} \subseteq X$ un subconjunto finito de X, $(K(x_i, x_j)) \in M_n\left(\mathcal{B}\left(\mathbb{R}^k\right)\right)$ una matriz de operadores (matrices), $c = (c_1, ..., c_n) \in (\mathbb{R}^d)^n$ y $\langle \cdot, \cdot \rangle_\times$ el producto interior del espacio producto de acuerdo a la Proposición 1.1.7, entonces:

$$\langle (K(x_i, x_j))\, c, c \rangle_\times = \langle \begin{bmatrix} \sum_{j=1}^n K(x_1, x_j)\, c_j \\ \vdots \\ \sum_{j=1}^n K(x_n, x_j)\, c_j \end{bmatrix}, c \rangle_\times$$

$$= \sum_{i=1}^n \langle \sum_{j=1}^n K(x_i, x_j)\, c_j, c_i \rangle = \sum_{i=1}^n \sum_{j=1}^n \langle K(x_i, x_j)\, c_j, c_i \rangle = \sum_{i=1}^n \sum_{j=1}^n c_i^T K(x_i, x_j)\, c_j.$$

Así que la matriz de operadores es semidefinida positiva si:

$$\sum_{i=1}^n \sum_{j=1}^n c_i^T K(x_i, x_j)\, c_j \geq 0.$$

Esta última caracterización nos será de utilidad más adelante.

Proposición 1.2.2.
Sea H un RKHS con valores en $\mathbb{R}^d$ y kernel reproductor K. Entonces K es definido positivo en el sentido de Moore y $K(x, x')^T = K(x', x)$.

Demostración. Sean $\{x_1, ...x_n\} \subseteq X$ y $v = (v_1, ..., v_n) \in (\mathbb{R}^d)^n$, entonces:

$$\begin{aligned} \sum_{j=1}^n \sum_{i=1}^n \langle K(x_i, x_j)\, v_j, v_i \rangle &= \sum_{j=1}^n \sum_{i=1}^n \langle E v_{x_i} E v_{x_j}^* v_j, v_i \rangle \\ &= \sum_{j=1}^n \sum_{i=1}^n \langle E v_{x_j}^* v_j, E v_{x_i}^* v_i \rangle \\ &= \langle \sum_{j=1}^n E v_{x_j}^* v_j, \sum_{j=1}^n E v_{x_j}^* v_j \rangle = \| \sum_{j=1}^n E v_{x_j}^* v_j \|^2 \geq 0. \end{aligned}$$

Para probar la segunda propiedad notemos que:

$$K(x, x')^* = (E v_x E v_{x'}^*)^* = E v_{x'} E v_x^* = K(x', x).$$

$\therefore K(x, x')^T = K(x', x).$

□

Teorema 1.2.3 (Teorema de Moore–Aronszajn extendido)**.**
Sea $K : X \times X \to \mathcal{B}(\mathbb{R}^d)$ un kernel definido positivo (en el sentido de Moore) que satisface $K(x, x')^T = K(x', x)$. Entonces existe un único RKHS de funciones vectoriales, con codominio $\mathbb{R}^d$, tal que K es su kernel reproductor.

La prueba de este teorema, y la generalización de resultados previos, se pueden encontrar en [21]. Por este teorema tenemos nuevamente una biyección entre kernels definidos positivos, sobre $X \times X$, con una propiedad que generaliza la simetría del caso escalar, y los RKHSs de funciones vectoriales sobre X.

Capítulo 2

Machine learning

Productos interiores como medidas de similitud

Sean H, un espacio de Hilbert, y w, $z \in H$ dos vectores de norma fija. Notemos que:

$$\|w - z\|^2 = \|w\|^2 - 2\langle w, z\rangle + \|z\|^2,$$

por lo que, w y z son más similares (la distancias entre ellos es menor), cuanto más grande sea el producto interior entre ellos. Es por esta razón que se suele decir, en el contexto de machine learning, que el producto interior es una medida de similitud.

2.1. Ejemplo de un Kernel Definido Positivo: Kernel polinomial

Hasta este punto hemos visto a los *espacios de Hilbert de kernel reproductor* de manera abstracta, es por eso que aquí ejemplificaremos una aplicación con el fin de motivar su mejor comprensión como herramienta útil en machine learning.

En machine learning nuestros datos de entrada siempre los podemos ver como vectores en $\mathbb{R}^n$, sin importar su naturaleza. Ejemplo de esto son las imágenes digitales y los pixeles que las definen, cuyos valores se puedes acomodar en un vector, o con cadenas de caracteres para las cuales tenemos una amplia gama de algoritmos, que se agrupan bajo el nombre de *word embedding*, para realizar este trabajo.

Para este ejemplo, consideraremos $X = \mathbb{R}^n$, para algún $n \in \mathbb{N}$.

El siguiente ejemplo es tomado de [23], capítulo 1, y lo exponemos aquí haciendo hincapié en detalles que, aunque son elementales, resultan importantes.
Supongamos que tenemos un problema de reconocimiento de patrones con el siguiente conjunto de datos $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^m \subseteq \mathbb{R}^n \times \mathbb{R}$.
En esta tarea es de gran utilidad poder comparar elementos de nuestro espacio muestral entre sí. Para comparar dos elementos podríamos hacer uso del producto interior entre las n características con las que cuenta cada elemento, $\mathbf{x}_i$. Sin embargo, muchas veces es útil tener en cuenta la relación entre la j-ésima y la k-ésima características, la cual podemos considerar

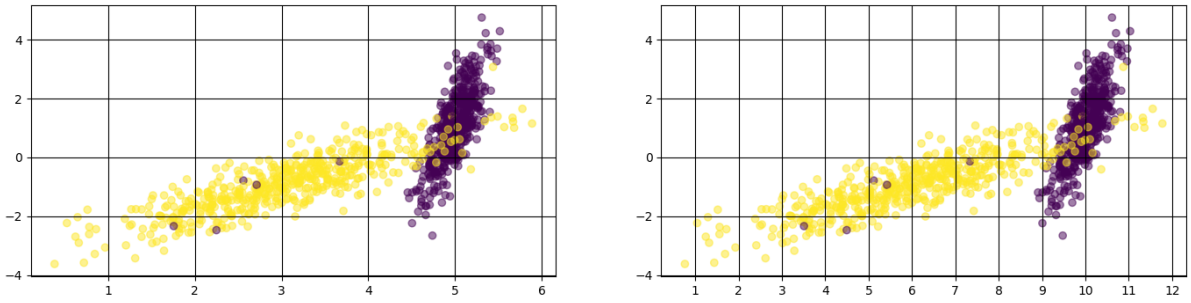


Figura 2.1: Es importante notar que el producto por escalar de una o más características no altera los patrones/regularidades en nuestros datos.

mediante el producto $x_j x_k$, de modo que ahora nuestros vectores de características pasan a $\mathbb{R}^{n+1}$ mediante la transformación:

$$\begin{aligned}\Phi :&\mathbb{R}^n \mapsto \mathbb{R}^{n+1}\\ &(x_1, ..., x_n) \mapsto (x_1, ..., x_n, x_j x_k)\,.\end{aligned}$$

De este modo, una vez que nuestros datos estén en el codominio de Φ, un espacio que suponemos expone de mejor manera el comportamiento de nuestros datos, podremos hacer uso del producto interior usual como medida de similitud para hacer el reconocimiento de patrones.

Ahora consideremos la hipotética situación en la que nos interesa considerar todos los monomios que son producto de r características (con reemplazo), estos serán de la forma:

$$x_{i_1} \cdots x_{i_r}, \quad \text{con } i_s \in \{1, ..., m\}, \quad \forall s \in \{1, ..., r\}.$$

En este caso, nuestra transfomación será de un espacio de dimensión n a uno de dimensión k, donde $k = \binom{r+n-1}{r}$, el número de las distintas combinaciones de tamaño r con elementos de un conjunto de tamaño n que se toman con reemplazo.
Aquí nos encontramos frente a un problema de eficiencia, pues suponiendo $n = 15$, $r = 5$ tenemos que $k = 11,628$, es decir, por cada elemento en T tendríamos que calcular $11,628$ productos, cada uno con 5 factores, y este número crecerá conforme la complejidad de nuestra tarea lo haga.

Es así que una *buena* transformación de nuestros datos podría ser costosa de implementar, sin embargo, podemos pensar el problema en términos de kernels. B. Schoelkopf (1997) en su tesis de doctorado expone un ejemplo y un par de resultados útiles para motivar el Kernel Polinomial y que aquí reproducimos en seguida.

Consideremos $n = r = 2$, y el mapeo:

$$C_2 : (x_1, x_2) \mapsto (x_1^2, x_2^2, x_1 x_2, x_2 x_1),$$

entonces tenemos la siguiente identidad para el producto interior de dos vectores $(x_1, x_2), (y_1, y_2) \in \mathbb{R}^2$, bajo el mapeo C_2:

$$\langle C_2(x_1, x_2), C_2(y_1, y_2)\rangle = x_1^2 y_1^2 + x_2^2 y_2^2 + 2x_1 x_2 y_1 y_2 = (x_1 y_1 + x_2 y_2)^2 = \langle (x_1, x_2), (y_1, y_2)\rangle^2.$$

Es importante notar que en C_2 estamos considerando todos los monomios de grado 2 que son distintos entre sí, aún en el orden de sus factores. Motivados por este ejemplo tenemos la siguiente proposición:

Proposición 2.1.1.
Sean n y r números naturales fijos, definimos a C_r como el mapeo que envía toda $x \in \mathbb{R}^n$ al vector real $C_r(x)$ cuyas entradas son todos los distintos monomios ordenados de grado r cuyos factores son las entradas del vector x. Entonces existe un kernel K que nos devuelve el valor del producto interior de cualesquiera vectores $x, y \in \mathbb{R}^n$ bajo la transformación C_r, y esta dado por:

$$K(x, y) = \langle C_r(x), C_r(y) \rangle = \langle x, y \rangle^r$$

Demostración. Haciendo los cálculos obtenemos:[1]

$$\begin{aligned}
\langle C_r(x), C_r(y) \rangle &= \sum_{j_1=1}^{n} \cdots \sum_{j_r=1}^{n} (x_{j_1} y_{j_1}) \cdots (x_{j_r} y_{j_r}) \\
&= \left(\sum_{j_1=1}^{n} x_{j_1} y_{j_1} \right) \cdots \left(\sum_{j_r=1}^{n} x_{j_r} y_{j_r} \right) \\
&= \left(\sum_{j=1}^{n} x_j y_j \right)^r = \langle x, y \rangle^r
\end{aligned}$$

$\therefore$ La igualdad se satisface y K es un kernel definido positivo y simétrico por estar definido como el producto interior de vectores bajo la transformación C_r. □

De este modo hemos transformado, de manera indirecta, nuestros datos y calculado el producto interior, como medida de similitud. Es importante notar que la transformación C_r asociada a K, bajo la cual se expresa como producto interior, no es única. Un ejemplo de esto lo tenemos al considerar $n = r = 2$, podemos notar que la transformación $\Phi : \mathbb{R}^2 \to \mathbb{R}^3$, dada por $\Phi(x_1, x_2) = \left(x_1^2, x_2^2, \sqrt{2} x_1 x_2\right)$, induce el mismo kernel, es decir:

$$K(x, y) = \langle C_2(x), C_2(y) \rangle = \langle \Phi(x), \Phi(y) \rangle$$

donde Φ es una transformación que obtiene todos los distintos monomios de grado 2, con uno de ellos escalado por una constante, lo cual no afecta de manera substancial los patrones/regularidades en los datos, como se puede ver en la figura 2.1. Esto mismo se puede hacer para C_r en general, obtener una transformación Φ equivalente, $\langle C_r(x), C_r(y) \rangle = \langle \Phi(x), \Phi(y) \rangle$, que solo considere los distintos monomios de grado r sin tener en cuenta el orden; sin embargo, obtener esta transformación es innecesario una vez que tenemos la forma compacta del kernel. Es así que en el contexto de machine learning, dado un kernel K, a cualquier función Φ que satisfaga $K(x, y) = \langle \Phi(x), \Phi(y) \rangle$ se le llama **mapeo de características**, sin importar la forma que tenga.

[1] No hacemos distinción explícita de los productos interiores para los distintos espacios de la forma $\mathbb{R}^n$, para alguna $n \in \mathbb{N}$, solo supondremos que se trata del producto interior usual en cada caso.

Finalmente, es importante mencionar que el kernel polinomial se puede generalizar de la siguiente forma:

$$K(x,y) = (\langle x,y\rangle + c)^r, \quad c \geq 0,$$

con el cual somos capaces de considerar todos los monomios de grado menor o igual a r.[2]

2.2. Planteamiento de la tarea de aprendizaje supervisado

La tarea de aprendizaje que nos limitaremos a abordar cae dentro del paradigma del aprendizaje supervisado, así que antes de continuar definiremos de manera formal tal paradigma.

2.2.1. Principios de teoría de la medida y probabilidad

La teoría expuesta en esta sección se toma de [14], en donde podemos hallar una introducción completa a la teoría de la medida. Del mismo modo que en el capítulo anterior, cada que hablemos de un conjunto X, este será distinto del vacío.

Definición 2.2.1.
Sean X un conjunto y $\mathcal{S} \subset \mathcal{P}(X)$. Decimos que $\mathcal{S}$ es una ***σ-álgebra de subconjuntos de*** *X si:*

- $\emptyset, X \in \mathcal{S}$;
- $A \setminus B \in \mathcal{S}, \quad \forall A, B \in \mathcal{S}$;
- $\bigcup_{n=1}^{\infty} A_n B \in \mathcal{S}, \quad \forall (A_n)_{n\in\mathbb{N}} \subseteq \mathcal{S}$.

Definición 2.2.2.
Definimos como ***espacio medible*** *al par ordenado $(X,\mathcal{S})$, donde X es un conjunto y $\mathcal{S} \subseteq \mathcal{P}(X)$ una σ-álgebra de subconjuntos de X.*

Definición 2.2.3 (Función medible)**.**
Sean $(X,\mathcal{S}_X)$, $(Y,\mathcal{S}_Y)$ espacios medibles y $f : X \to Y$ una función. Decimos que f es ***$\mathcal{S}_X$-medible*** *si:*

$$f^{-1}[A] \in \mathcal{S}_X, \quad \forall A \in \mathcal{S}_Y.$$

Definición 2.2.4.
Sea $(X,\mathcal{S})$ un espacio medible. Una función $\mu : \mathcal{S} \to \overline{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ es una ***medida*** *de $(X,\mathcal{S})$ si:*

- $\mu(\emptyset) = 0$;
- $\mu(A) \geq 0, \quad \forall A \in \mathcal{S}$;
- *[σ-aditividad] dada una sucesión $(A_n)_{n\in\mathbb{N}} \subseteq \mathcal{S}$ de elementos disjuntos por pares, se*

[2]Se debe hacer notar que cuando el valor de r es muy grande, entonces $K(x,y)$ es muy grande o muy cercano a cero.

cumple que:

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu\left(A_n\right).$$

Definición 2.2.5.
Sean $(X, \mathcal{S})$ un espacio medible y μ una medida de $(X, \mathcal{S})$. Decimos que μ es una ***medida finita*** *si $\mu(A) < +\infty, \quad \forall A \in \mathcal{S}$.*

Definición 2.2.6.
Definimos como ***espacio de medida*** *a la terna $(X, \mathcal{S}, \mu)$, donde X es un conjunto, $\mathcal{S} \subseteq \mathcal{P}(X)$ una σ-álgebra de subconjuntos de X y μ una medida de $(X, \mathcal{S})$.*

Definición 2.2.7 (Espacio de medida completo)**.**
Sea $(X, \mathcal{S}, \mu)$ un espacio de medida. Decimos que es ***completo*** *si:*

$$\forall A \in \mathcal{S},\ \forall B \subseteq A \quad \left(\mu(A) = 0 \implies B \in \mathcal{S}\right).$$

Definición 2.2.8.
Sea $(X, \mathcal{S}, \mu)$ un espacio de medida. Decimos que es un ***espacio de probabilidad*** *si:*

- *$(X, \mathcal{S}, \mu)$ es completo;*[3]
- *μ es una medida finita y $\mu(X) = 1$.*

A la medida de un espacio de probabilidad se suele denotar con $\mathbb{P}$, en lugar de μ.

Definición 2.2.9 (Variable Aleatoria)**.**
Sean $(X, \mathcal{S}_X, \mathbb{P})$ un espacio de probabilidad y $(\mathbb{R}^n, \mathcal{B})$ el espacio de medida para $\mathbb{R}^n$ con la $\sigma-$álgebra de Borel. La función $Z : X \to \mathbb{R}^n$ es una ***variable aleatoria*** *si Z es $\mathcal{S}_X$-medible.*

En adelante adoptaremos la siguiente notación:

$$\mathbb{P}[Z \in A] = \mathbb{P}\left(\{\omega \in X : Z(\omega) \in A\}\right),$$

con $A \in \mathcal{B}$.

Las siguientes definiciones las enunciamos para variables aleatorias reales.

Definición 2.2.10.
Sea $Z : X \to \mathbb{R}$ una variable aleatoria. Definimos el ***valor esperado*** *de Z como:*

$$\mathbb{E}[Z] = \int Z d\mathbb{P}.$$

En donde la integral corresponde a la de Lebesgue.

Definición 2.2.11.
Sea $Z : X \to \mathbb{R}$ una variable aleatoria. Definimos su ***función de distribución acumulada****, $F : \mathbb{R} \to \mathbb{R}$, como:*

$$F(x) = \mathbb{P}[Z \in (-\infty, x]].$$

[3]La completez no es necesaria en la definicón, sin embargo, en teoría de la probabilidad lo común es trabajar con espacios completos.

Modos de convergencia

Denotaremos con $(Z_i)_{i\in\mathbb{N}}$ a una sucesión de variables aleatorias sobre un mismo espacio medible, y con F_i a sus respectivas funciones de distribución acumulada.

Definición 2.2.12.
*Decimos que $(Z_i)_{i\in\mathbb{N}}$ converge **en media cuadrática** a una variable aleatoria Z si:*

$$\mathbb{E}\left[(Z_i - Z)^2\right] \to 0,$$

y lo denotamos como:

$$Z_i \xrightarrow{L^2} Z.$$

Definición 2.2.13.
*Decimos que $(Z_i)_{i\in\mathbb{N}}$ converge **en probabilidad** a una variable aleatoria Z si:*

$$\mathbb{P}\left[|Z_i - Z| > \epsilon\right] = \mathbb{P}\left[|Z_i - Z| \in (\epsilon, \infty)\right] \to 0, \qquad \forall \epsilon > 0,$$

y lo denotamos como:

$$Z_i \xrightarrow{P} Z.$$

Definición 2.2.14.
*Decimos que $(Z_i)_{i\in\mathbb{N}}$ converge **en distribución**, o **en ley**, a una variable aleatoria Z, con función de distribución F, si:*

$$F_i(x) \to F(x), \quad \forall x \in \mathbb{R},\ \textit{tal que } F \textit{ es continua en } x,$$

y lo denotamos como:

$$Z_i \xrightarrow{d} Z.$$

Definición 2.2.15.
*Decimos que $(Z_i)_{i\in\mathbb{N}}$ converge **casi seguro** a una variable aleatoria Z si:*

$$\mathbb{P}\left[\{\omega \in X; \lim_{i\to\infty} Z_i(\omega) = Z(\omega)\}\right] = 1,$$

y lo denotamos como:

$$Z_i \xrightarrow{c.s.} Z.$$

2.2.2. Modelo de aprendizaje supervisado

Para definir el paradigma de tipo supervisado seguiremos el mismo camino que en [25], definiendo cada una de las partes constitutivas.

- **Espacio Muestral**
  El espacio muestral es un conjunto de la forma $\mathcal{Z} \subseteq \mathcal{X} \times \mathcal{Y}$, en donde $\mathcal{X}$ es el conjunto de características con las cuales describimos los objetos que son de nuestro interés y $\mathcal{Y}$ es el conjunto de etiquetas posibles para estos objetos, es decir, $\mathcal{Z}$ es un conjunto que contiene todos los pares (x, y) cuya asignación de la etiqueta y para las características

x es válida. Tanto el conjunto $\mathcal{X}$ como el conjunto $\mathcal{Y}$ pueden ser de naturaleza muy variada, por lo que es necesario traducir las características y etiquetas a un lenguaje fácil de tratar, en nuestro caso lo que haremos es considerar una variable aleatoria $\widehat{\mathcal{Z}} : \mathcal{Z} \to \mathbb{R}^n \times \mathbb{R}^m$, en donde $\mathbb{R}^n$ y $\mathbb{R}^m$ son espacios donde se contendrá la información de las características y etiquetas, respectivamente. Denotaremos con $\mathcal{Z}$ tanto al espacio muestral como a la variable aleatoria, al ser esta última una traducción del espacio muestral, con medida de probabilidad según la Definición 2.2.9, y con A al conjunto de valores que toma la función $\widehat{\mathcal{Z}}$ en $\mathbb{R}^n$.

Para $\mathcal{Z}$ suponemos que existe un espacio de probabilidad $(\mathcal{Z}, \mathcal{S}, \mu)$ desconocido que se ajusta a la realidad observable, cuya distribución denotaremos con $\mathcal{D}$, así como una función medible, $f : A \subseteq \mathbb{R}^n \to \mathbb{R}^m$, que nos da la asignación real de etiquetas para cada elemento de A.[4]

El fin del algoritmo de aprendizaje será aproximar la función f, esto lo conseguiremos alimentando el algoritmo con N sampleos independientes de $\mathcal{Z}$, $\{(x_i, y_i)\}_{i=1}^N$, con N suficientemente grande de manera que $\{(x_i, y_i)\}_{i=1}^N$ sea una muestra representativa, es decir, que aproxime de manera empírica a la distribución real de $\mathcal{Z}$.

Una vez que tenemos nuestro muestreo aleatorio, lo particionamos en los siguientes subconjuntos[5]:

- **Conjunto de entrenamiento:** ($\sim 80\,\%$ del total de datos)
  Con estos datos ajustamos nuestro modelo para aproximar el comportamiento de los datos.
- **Conjunto de validación:** ($\sim 10\,\%$ del total de datos)
  Este conjunto es útil cuando se desea encontrar una combinación óptima de hiperparámetros para el algoritmo. Para esto se suele entrenar el modelo con cada combinación de hiperparámetros, después se compara el rendimiento de estos modelos en el conjunto de validación para finalmente quedarnos con el de mejor rendimiento.
- **Conjunto de test:** ($\sim 10\,\%$ del total de datos)
  Una vez finalizado el entrenamiento con los hiperparámetros que mejor nos funcionaron, se evalúa el modelo en el conjunto de test, y es esta evaluación la que consideraremos para tomar la medida generalizada del error.

Es importante recalcar que estos subconjuntos los creamos de manera *aleatoria*, con la misma probabilidad ($\frac{1}{N}$) para cada elemento del conjunto $\{(x_i, y_i)\}_{i=1}^N$ y sin reemplazo.

- **Medida de error**
  A lo largo del algoritmo es necesaria una *medida* que nos indique cuan cerca está nuestro modelo de imitar el comportamiento descrito por los datos, que es la información disponible para conocer f. De manera teórica podemos definir el **error real** como aquel

[4] Aquí estamos suponiendo que existe una asignación determinista dada por f, lo cual no siempre es cierto pues la asignación puede darse por una variable aleatoria condicionada con A.

[5] Los porcentajes indicados en cada caso son los más comúnmente usados, sin embargo, pueden variar dependiendo del problema.

que se basa en la distribución real, $\mathcal{D}$, de $\mathcal{Z}$ y esta dado por:

$$\mathcal{L}_{(\mathcal{D})}\left(f_\theta\right) = \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[f_\theta\left(x\right) \neq y\right] = \mathbb{P}_{(x,y)\sim\mathcal{D}}\left[f_\theta\left(x\right) \neq f\left(x\right)\right].$$

Donde f_θ es el modelo que deseamos evaluar, que depende de un conjunto de variables θ. Pero dado que desconocemos $\mathcal{D}$, lo que haremos será aproximar este error con una medida empírica del error, en base a los datos disponibles, la cual denotaremos por $\mathcal{L}$.

El principio según el cual minimizar el **error empírico** equivale a minimizar el error real es conocido como **minimización empírica del riesgo** (Empirical Risk Minimization, ERM).

La forma del error empírico puede variar según la naturaleza de nuestro problema, como se ejemplifica en seguida:

- **Norma p**
  La función del error empírico esta dada por:

$$\mathcal{L}\left(f_\theta, \{(x_i, y_i)\}_{i=1}^N\right) = \left(\sum_{i=1}^{N}\left(y_i - f_\theta\left(x_i\right)\right)^p\right)^{1/p}.$$

  Esta medida del error la ocupamos cuando nuestro problema es de regresión, es decir, con y a valores continuos. La relevancia del índice p es que entre mayor sea su valor, la función de error le dará mayor importancia a los datos en los que erramos más.

- **Entropía cruzada**
  La función del error empírico esta dada por:

$$\mathcal{L}\left(f_\theta, \{(x_i, y_i)\}_{i=1}^N\right) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log\left(f_\theta(x_i)\right) + \left(1 - y_i\right) \cdot log\left(1 - f_\theta\left(x_i\right)\right).$$

  Esta función es usada cuando nuestro problema es de clasificación binaria, $\mathcal{Y} = \{0, 1\}$. En este caso nuestro modelo f_θ calcula la probabilidad de etiquetar con 1 a las características con que se alimenta el modelo.

A la función $\mathcal{L}$ también se le conoce como función de pérdida, función de costo o simplemente función de error.

- **Entrenamiento**
  Una vez que conocemos nuestro problema, fijamos una arquitectura adecuada[6] del modelo mediante el cual queremos aproximar a f. Esta arquitectura posee valores variables, θ, con los cuales seremos capaces de modificar el comportamiento del modelo a la hora de hacer las asignaciones de etiquetas, $x \mapsto y$.
  La idea del algoritmo es hallar una θ adecuada para que $f_\theta \approx f$. Esto se consigue durante la fase de entrenamiento, en la cual, con ayuda de los datos de entrenamiento, se ajusta θ con el objetivo de que la función de error se minimice.

[6] Adecuada según la experiencia que se tiene.

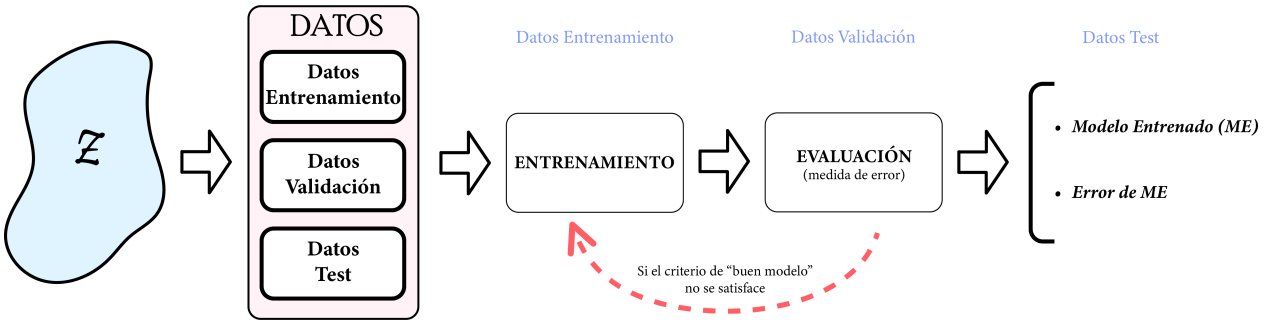


Figura 2.2: Bosquejo del aprendizaje máquina.

2.3. Redes Neuronales

Es durante el siglo XX que da inicio la era de la inteligencia artificial con grandes personajes como Alan Turing, con sus artículos *On Computable Numbers*(1936), que representa el inicio del estudio de la computación teórica, y *Computing Machinery and Intelligence*(1950), en el cual se embarca de manera formal a responder la pregunta: ¿Pueden pensar las máquinas?. En 1943, motivados por el artículo de Turing de 1936 e inspirados en las neuronas biológicas, **McCulloch y Pitts** (1943) publican su modelo matemático de red neuronal, que se reconoce hoy en día como el primer modelo matemático neuronal moderno.

Pese a la novedad del modelo de neurona de McCulloch y Pitts, este contaba con severas deficiencias. El modelo está dado por una función $\phi : \{0,1\}^n \to \{0,1\}$, que se define como $\phi(x_1, ..., x_n) = \mathbb{1}_{[\theta,\infty)}\left(\sum_{i=1}^n x_i\right)$, donde θ es un valor fijado previamente según el problema a resolver. De manera escueta, la lógica detrás de este modelo es la siguiente:

- Fijamos un umbral θ, que podemos traducir como el mínimo de energía necesaria para excitar la neurona.
- La neurona recibe n señales, cada una de las cuales puede ser una *señal excitada* (1) o una *señal no excitada* (0).
- Se suman las contribuciones de las señales recibidas y si el total de la suma es mayor o igual al umbral, θ, entonces la neurona se excita y retorna el valor de 1, en otro caso devuelve 0.

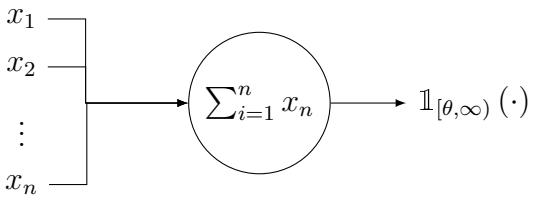


Figura 2.3: Modelo de neurona de McCulloch-Pitts.

En este modelo la información es binaria y la función de activación es una función umbral, lo que limita mucho la aplicación que podamos darle, debido a que la complejidad de las redes con este tipo de neuronas sería muy grande.

En 1958 **Rosenblatt**, inspirado en la teoría hebbiana de plasticidad sináptica[7], propuso un modelo de neurona artificial al cual nombró como perceptrón, así como un algoritmo de aprendizaje. El modelo es el siguiente:

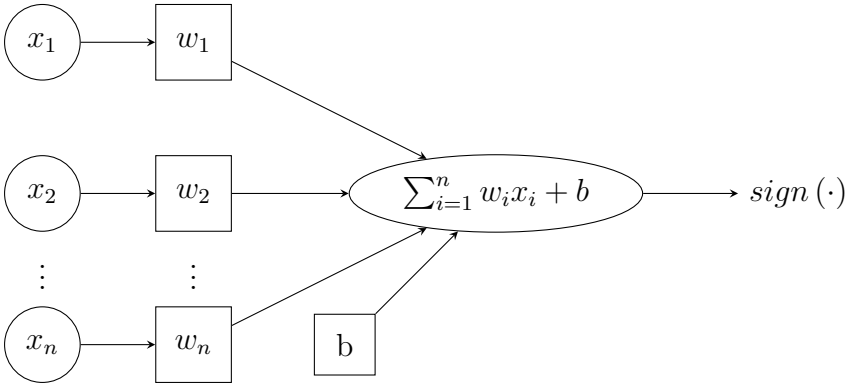


Figura 2.4: Perceptrón de Rosenblatt.

En este modelo las señales de entrada son números reales y cada señal x_i es ponderada por un real w_i, de manera que habrá señales con mayor influencia que otras, así como señales inhibitorias (con valores negativos). Por otro lado, la función de activación es fija:[8]

$$sign(x) = \begin{cases} 1 & \text{si } x > 0 \\ 0 & \text{si } x = 0 \\ -1 & \text{si } x < 0, \end{cases}$$

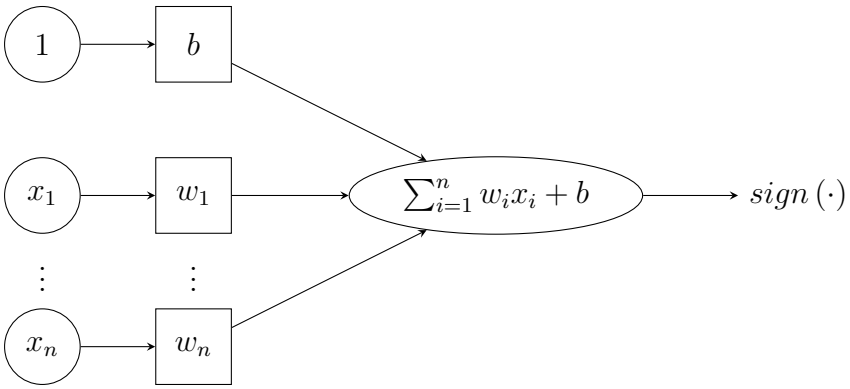
en donde el umbral de activación[9] esta dado por el peso b (bias) cuyo valor se aprende junto con los pesos que ponderan a las señales.

Este modelo es de clasificación binaria y en esencia lo que hace es dividir $\mathbb{R}^n$ mediante un subespacio vectorial afín $(n-1) - dimensional$ (un hiperplano). Notemos que mediante el mapeo $(x_1, ..., x_n) \mapsto (1, x_1, ..., x_n)$ podemos reescribir nuestro modelo de manera equivalente ignorando el bias, el cual pasa a ser el peso que corresponde a la primer dimensión, y nuestra tarea de aprendizaje ya no es encontrar un subespacio afín sino solamente un subespacio vectorial $n - dimensional$ en $\mathbb{R}^{n+1}$.

[7]La capacidad de adaptación de las neuronas del cerebro durante el proceso de aprendizaje.

[8]Una practica común es usar $\mathbb{1}_{[0,\infty)}(\cdot)$ como función de activación, lo que es una forma completamente equivalente de reescribir este modelo.

[9]La activación aquí equivale a tener signo positivo.



La relevancia de esta propuesta se debió en parte al algoritmo iterativo de aprendizaje, que en seguida presentamos, en donde suponemos que los datos están según la transformación descrita en el párrafo anterior:

Algorithm 1 Entrenamiento del perceptrón de Rosenblatt

Input: $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^{n+1} \times \{-1, 1\}$ un conjunto de datos linealmente separables, $\mathbf{w} = (0, ..., 0) \in \mathbb{R}^{n+1}$.
  while $\exists m \in \{1, ..., N\}$ tal que $\langle \mathbf{w}, \mathbf{x}_m \rangle y_m \leq 0$ **do**
    for $i = 1$ **to** k **do**
      if $\langle \mathbf{w}, \mathbf{x}_i \rangle y_i \leq 0$ **then**
        $\mathbf{w} \leftarrow \mathbf{w} + (y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle) \cdot \mathbf{x}_i$
      end if
    end for
    Test de convergencia ▷ Se prueba la condición del bucle **while**
  end while
Output: **w**

La prueba de la convergencia de este algoritmo se puede consultar en [25].

Posteriormente, con el paso de los años se realizaron diversas contribuciones, tanto en el modelo como en los algoritmos de aprendizaje, para llegar a lo que actualmente conocemos como una red neuronal artificial (ANN), y a la retropropagación (backpropagation) como fundamento del proceso de aprendizaje.

Modelo de Red Neuronal Artificial

El caso prototípico de una ANN es el perceptrón multicapa, para el cual necesitamos definir lo siguiente:

- Número de neuronas en la capa entrada: n_0, coincide con la dimensión del espacio en el que se encuentran las características del espacio muestral.
- Número de neuronas en la capa de salida: n_L.
- Profundidad de la red: L.
- Número de capas ocultas: $L - 1$.

- Número de neuronas en la $i-$ésima capa: n_i.
- Función de activación (no linear en la mayoría de los casos), $\sigma : \mathbb{R} \longrightarrow \mathbb{R}$:
  - **Identidad**: Equivale a no considerar ninguna función de activación.
  - **Función logística**: Mayormente conocida como *función sigmoide*.[10]

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

  - **Tangente hiperbólica**:

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

  - **ReLU** (Rectified Linear Unit):

$$\sigma(x) = \text{máx}\{0, x\}$$

  - **etc...**

Definimos el modelo de red neuronal, $f_\theta(\cdot) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$, a partir de la siguiente recursión:

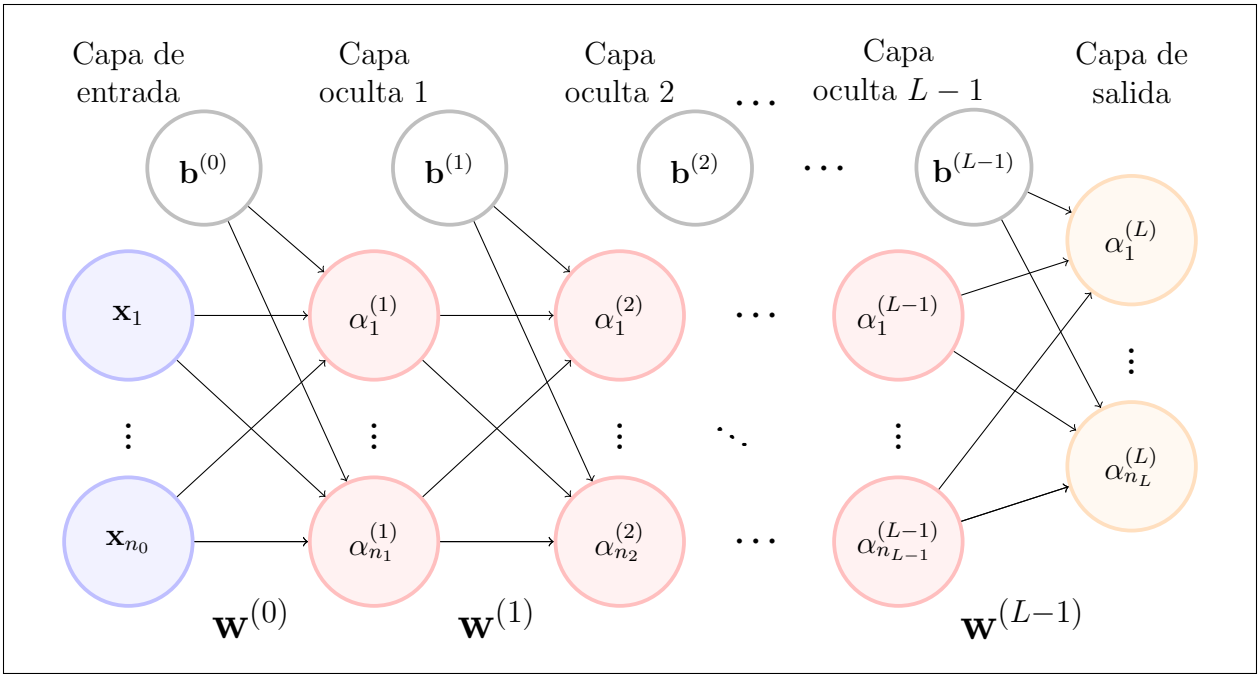
$$\alpha^{(0)}(\mathbf{x};\theta) = \mathbf{x},$$

$$\tilde{\alpha}^{(l+1)}(\mathbf{x};\theta) = \mathbf{W}^{(l)}\alpha^{(l)}(\mathbf{x};\theta) + \mathbf{b}^{(l)},$$
$$\alpha^{(l)}(\mathbf{x};\theta) = \sigma\left(\tilde{\alpha}^{(l)}(\mathbf{x};\theta)\right),$$

en donde $\mathbf{W}^{(l)} \in \mathbb{R}^{n_{l+1}\times n_l}$, $\mathbf{b}^{(l)} \in \mathbb{R}^{n_{l+1}}$, $f_\theta = \tilde{\alpha}^{(L)}(\cdot;\theta)$ y la evaluación en σ es puntual. A las funciones $\tilde{\alpha}^{(l)}(\cdot;\theta) : \mathbb{R}_{n_0} \to \mathbb{R}^{n_l}$ y $\alpha^{(l)}(\cdot;\theta) : \mathbb{R}_{n_0} \to \mathbb{R}^{n_l}$ las llamaremos **pre-activación** y **post-actiación** de la capa (l), respectivamente, con θ denotaremos a los pesos la red, $\{(\mathbf{W}^{(l)}, \mathbf{b}^{(l)})\}_{l=1}^{L-1}$, que son los valores que se han de aprender y con P a la cantidad total de pesos.

Cabe recalcar que existen varias formas de inicializar θ, la mayoría de ellas de manera aleatoria. En cuestión de notación, en adelante omitiremos θ en las pre-activaciones y post-activaciones, pero es importan tener presente su influencia.

[10] Matemáticamente hablando, una función sigmoide no es una función en particular sino una familia de funciones, a la cual pertenece esta función.



Aquí los datos se procesan por capas, tal que la salida de la capa $(l-1)$ es la entrada de la capa l, más un sesgo $\mathbf{b}^{(l-1)}$.

Existen dos formas posibles de describir esta red neuronal, primero esta aquella en la que, para una θ fija, la red neuronal es una función que evalúa un vector de características y le asigna una etiqueta, $f_\theta(\cdot) : \mathbb{R}^{n_0} \longrightarrow \mathbb{R}^{n_L}$, y por otro lado la red neuronal puede verse en función de θ, mediante el mapeo $\theta \mapsto f_\theta$, en donde el codominio es un espacio de funciones. La elección de la interpretación de la red dependerá de lo que pretendamos estudiar.

Retropropagación (Backpropagation)

Las pruebas que se omiten en este apartado acerca de la derivada direccional pueden encontrarse en [10].

- **Descenso de gradiente**
  Sea $g : \mathbb{R}^n \longrightarrow [0, \infty)$ una función, $g \in C^1(\mathbb{R}^n)$, definimos el **gradiente** de g como la función $\nabla g : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ dada por:

$$\nabla g(\mathbf{p}) = \begin{bmatrix} \frac{\partial g}{\partial x_1}(\mathbf{p}) \\ \vdots \\ \frac{\partial g}{\partial x_n}(\mathbf{p}) \end{bmatrix}.$$

  Una concepto importante en este punto es el de **derivada direccional**. Dados un punto $\mathbf{p} \in \mathbb{R}^n$ y un vector $\mathbf{v} \in \mathbb{R}^n$, definimos la derivada direccional de g en $\mathbf{p}$ con dirección $\mathbf{v}$ como:

$$\mathbf{D}_\mathbf{v} g(\mathbf{p}) = \lim_{h \to 0} \frac{g(\mathbf{p} + h\mathbf{v}) - g(\mathbf{p})}{h}.$$

A grandes rasgos, la derivada direccional en $\mathbf{p}$ con dirección $\mathbf{v}$ nos indica la tasa de cambio en el punto $\mathbf{p}$ en la dirección de $\mathbf{v}$.

Dos resultados útiles de cálculo son los siguientes:

$$\mathbf{D}_{\mathbf{v}}g(\mathbf{p}) = \nabla g(\mathbf{p}) \cdot \mathbf{v},$$

$$\mathbf{D}_{c\mathbf{v}}g(\mathbf{p}) = c\mathbf{D}_{\mathbf{v}}g(\mathbf{p}), \quad \text{con} \quad c \in \mathbb{R}.$$

Son relevantes para la comprensión del método de descenso de gradiente, pues fijando un punto $\mathbf{p} \in \mathbb{R}^n$, en que cual $\nabla g(\mathbf{p}) \neq \mathbf{0}$, y en el caso hipotético que nos interese conocer la dirección en la que la función g tiene una mayor tasa de crecimiento, es decir, buscamos maximizar $\mathbf{D}_{\mathbf{v}}g(\mathbf{p})$ en función de $\mathbf{v}$, con $\|\mathbf{v}\|$ constante (supongamos que $\|\mathbf{v}\| = 1$), entonces es importante notar que:

$$\mathbf{D}_{\mathbf{v}}g(\mathbf{p}) = \nabla g(\mathbf{p}) \cdot \mathbf{v} = \|\nabla g(\mathbf{p})\| cos(\theta_{\mathbf{v}}),$$

donde $\theta_{\mathbf{v}}$ es en ángulo entre los vectores $\nabla g(\mathbf{p})$ y $\mathbf{v}$. De este modo, maximizar $\mathbf{D}_{\mathbf{v}}g(\mathbf{p})$ equivale a maximizar $cos(\theta_{\mathbf{v}})$, lo cual sucede con $\theta_{\mathbf{v}} = 0$, lo que a su vez implica $\mathbf{v} = \frac{\nabla g(\mathbf{p})}{\|\nabla g(\mathbf{p})\|}$, es decir, la dirección del gradiente coincide con aquella en la que la tasa de cambio es mayor, y por la homogeneidad de la derivada direccional tenemos que la dirección del vector $-\nabla g(\mathbf{p})$ coincide con aquella en la que la función disminuye más rápidamente, para una vecindad de $\mathbf{p}$.

La idea detrás del **descenso del gradiente** es la siguiente: dado un punto $\mathbf{p} \in \mathbb{R}^n$, con $\nabla g(\mathbf{p}) \neq \mathbf{0}$, sabemos que si nos movemos en $\mathbb{R}^n$ siguiendo la dirección del vector $-\nabla g(\mathbf{p})$, entonces hallaremos valores menores a $g(\mathbf{p})$, esto al menos en una vecindad muy pequeña de $\mathbf{p}$. El hecho de que alcanzar valores menores solo se pueda garantizar en vecindades pequeñas nos lleva a la necesidad de introducir lo que nombraremos **learning rate** (o **tasa de aprendizaje**), ϵ, que es un real positivo con el que escalaremos el vector $-\nabla g(\mathbf{p})$, buscando que se satisfaga $g\left(\mathbf{p} - \epsilon \nabla g(\mathbf{p})\right) < g\left(\mathbf{p}\right)$.

Existen diversos métodos basados en el *descenso de gradiente*, con muy buenos resultados, sin embargo, nos limitaremos a enunciar el algoritmo básico, que será el que ocuparemos más adelante.

Algorithm 2 Descenso de Gradiente

Input: $g : \mathbb{R}^n \longrightarrow [0, \infty)$, $g \in C^1\left(\mathbb{R}^n\right)$, $\mathbf{w} = (0, ..., 0) \in \mathbb{R}^n$, $0 < \epsilon < 1$.
  while $\nabla g\left(\mathbf{w}\right) \neq \mathbf{0}$ **do**
    $\mathbf{w} \leftarrow \mathbf{w} - \epsilon \nabla g\left(\mathbf{w}\right)$
  end while
Output: $\mathbf{w}$

Este algoritmo, tal como se enuncia, no necesariamente converge, sin embargo, es el primer paso para métodos más complejos, con mejores resultados en la práctica.

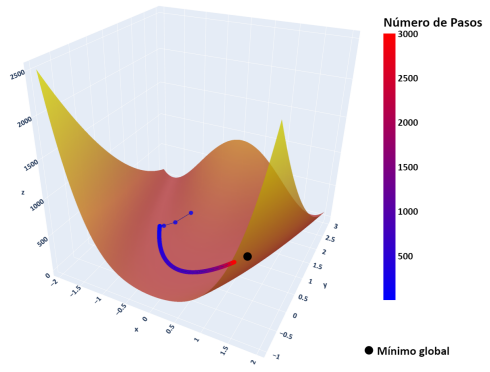


Figura 2.5: Ejemplificamos el descenso de gradiente poniendo como criterio de paro $3,000$ número de iteraciones, en vez del criterio de convergencia, con la función de Rosenbrock, $f(x,y) = (1-x)^2 + 100(y-x^2)^2$, comenzando en $(-0.5, 1.5)$, con learning rate igual a 0.001 y un mínimo global en $(1,1)$. Esta función tiene la peculiaridad de que su gráfica posee un valle muy largo y sin pendientes pronunciadas a lo largo del valle, razón por la cual mediante descenso de gradiente es fácil llegar al valle pero un vez ahí los pasos son muy pequeños porque la norma del gradiente a lo largo del valle es muy pequeña, esto lo podemos notar en la gráfica, pues en pocos pasos se llega al valle y la mayoría de los pasos se concentran en torno al mínimo global a lo largo del valle.

Supongamos que tenemos un conjunto $T = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N} \subseteq \mathbb{R}^{n_0} \times \mathbb{R}^{n_L}$, cuyos elementos vienen del espacio muestral del cual deseamos imitar su distribución. Sea $f_\theta : \mathbb{R}^{n_0} \longrightarrow \mathbb{R}^{n_L}$ una ANN que sigue el tipo de arquitectura antes descrita, con función de activación continuamente diferenciable[11], y consideremos una función de pérdida, $\mathcal{L}(\cdot; f_\cdot, T) : \mathbb{R}^P \longrightarrow [0, \infty)$, sobre el espacio de los pesos que definen a la red, θ, tal que $\mathcal{L}$ da una medida del error de la red en T mediante la aplicación de funciones diferenciables a las salidas de la red, de modo que $\mathcal{L}$ es diferenciable.

Una forma de optimizar $\mathcal{L}$ fue propuesta en el año 1986 cuando se publicó *Learning Representations by Back-Propagating Errors*, artículo que introduce de manera formal el método de backpropagation mediante el cual se siguen entrenando las redes neuronales hoy en día. Este método es útil para calcular las derivadas parciales necesarias para hacer descenso de gradiente sobre $\mathcal{L}(\cdot; f_\cdot, T)$, y lo podemos dividir en dos grandes pasos: paso hacia adelante (forward) y paso hacia atrás (backward).

El primer paso es evaluar $\mathcal{L}(\cdot; f_\cdot, T)$ en una θ específica, y posteriormente, a través del paso hacia atrás, se calculan las derivadas parciales (el gradiente).
Para explicar el segundo paso, consideraremos la función de error dada por:

$$\mathcal{L}(\theta; f_\theta, T) = \frac{1}{N} \sum_{i=1}^{N} \frac{\| y_i - f_\theta(\mathbf{x}_i) \|^2}{2} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{\mathbf{x}_i}(\theta; f_\theta, T),$$

en donde el error total es igual al promedio de los errores por elemento. Pese a que estamos considerando una función de pérdida particular, pasar a cualquier otra función de pérdida (válida) no debería significar ninguna dificultad salvo, que el cálculo de los errores (total y por elemento) sea complejo.

Comenzamos con el cálculo del gradiente para el error de un elemento cualquiera de T, denotaremos a este elemento con $(\mathbf{x}, y)$ y a f_θ con f.

El error individual esta dado por:

[11] La función de activación la elegimos así con el fin de simplificar los cálculos, pero en la práctica es posible elegir otro tipo de función de activación.

$$\mathcal{L}_{\mathbf{x}}\left(\theta; f, T\right) = \frac{\|y - f\left(\mathbf{x}\right)\|^2}{2} = \frac{\sum_{i=1}^{n_L}(y_i - f_i\left(\mathbf{x}\right))^2}{2}.$$

Consideraremos la siguiente notación:

$$z^l = \tilde{\alpha}^{(l+1)}\left(\mathbf{x}\right) =, \quad \text{tal que} \quad z_i^l = \sum_{j=1}^{n_l} \mathbf{W}_{i,j}^{(l)} \alpha_j^{(l)}\left(\mathbf{x}\right) + \mathbf{b}_i^{(l)},$$

$$\delta_i^l = \frac{\partial \mathcal{L}_x}{\partial z_i^l}, \quad \text{y} \quad \delta^l = \left[\delta_1^l, ..., \delta_{n_{l+1}}^l\right]^T.$$

El método de backpropagation nos da una forma de calcular δ^l para cada capa, a partir de los cuales podremos deducir el gradiente.

1. El algoritmo comienza con el cálculo de δ^{L-1}, que corresponde a la capa de salida. Por regla de la cadena:

$$\delta_i^{L-1} = \frac{\partial \mathcal{L}_x}{\partial f_i} \frac{\partial f_i}{z_i^{L-1}} = (f_i(\mathbf{x}) - y_i)\sigma_0'\left(z_i^{L-1}\right),$$

en donde σ_0 es una función de activación, que suele ser la identidad, para la salida de la red. Entonces de manera vectorial tenemos lo siguiente:

$$\delta^{L-1} = (f(\mathbf{x}) - y) \odot \sigma_0'(z^{L-1}),$$

donde $\odot$ es el producto de Hadamard, o producto entrada a entrada.

2. Ahora supongamos que conocemos δ^l, con $l \in \{1, ..., L-1\}$, y deseamos calcular δ^{l-1}. Por la regla de la cadena:

$$\begin{aligned}
\delta_i^{l-1} &= \frac{\partial \mathcal{L}_x}{\partial z^l} \frac{\partial z^l}{\partial z_i^{l-1}} \\
&= \left(\delta^l\right)^T \left[\frac{\partial z_1^l}{\partial z_i^{l-1}}, ..., \frac{\partial z_{n_{l+1}}^l}{\partial z_i^{l-1}}\right]^T \\
&= \left(\delta^l\right)^T \left[\frac{\partial z_1^l}{\partial \alpha_i^{(l)}} \frac{\partial \alpha_i^{(l)}}{\partial z_i^{l-1}}, ..., \frac{\partial z_{n_{l+1}}^l}{\partial \alpha_i^{(l)}} \frac{\partial \alpha_i^{(l)}}{\partial z_i^{l-1}}\right]^T \\
&= \left(\left[\mathbf{W}_{1,i}^{(l)} \sigma'\left(z_i^{l-1}\right), ..., \mathbf{W}_{n_{l+1},i}^{(l)} \sigma'\left(z_i^{l-1}\right)\right] \delta^l\right)^T \\
&= \sum_{j=1}^{n_{l+1}} \delta_j^l \mathbf{W}_{j,i}^{(l)} \sigma'\left(z_i^{l-1}\right).
\end{aligned}$$

Entonces:

$$\delta^{l-1} = \begin{bmatrix} \sum_{j=1}^{n_{l+1}} \delta_j^l \mathbf{W}_{j,1}^l \\ \vdots \\ \sum_{j=1}^{n_{l+1}} \delta_j^l \mathbf{W}_{j,n_{l+1}}^l \end{bmatrix} \odot \sigma'\left(z^{l-1}\right) = \left(\mathbf{W}^{(l)^T} \delta^l\right) \odot \sigma'\left(z^{l-1}\right).$$

3. Una vez que hemos calculado δ^l, para $l \in \{0, ..., L-1\}$, es útil notar que:

$$\frac{\partial \mathcal{L}_x}{\partial \mathbf{b}_i^l} = \frac{\partial \mathcal{L}}{\partial z_i^l} \frac{\partial z_i^l}{\partial \mathbf{b}_i^{(l)}} = \delta_i^l,$$

$$\frac{\partial \mathcal{L}_x}{\partial \mathbf{W}_{i,j}^{(l)}} = \frac{\partial \mathcal{L}}{\partial z_i^l} \frac{\partial z_i^l}{\partial \mathbf{W}_{i,j}^{(l)}} = \delta_i^l \left(\alpha_j^{(l)}(\mathbf{x})\right).$$

   A partir de esto, el cálculo de $\nabla \mathcal{L}_x$ es inmediato.

Y por linealidad del gradiente:

$$\nabla \mathcal{L}\left(\theta; f_\cdot, T\right) = \frac{1}{n} \sum_{\mathbf{x} \in T} \nabla \mathcal{L}_{\mathbf{x}}.$$

En el paso hacia adelante (forward) se calculan $\alpha_j^{(l)}(\mathbf{x})$ y en el paso hacia atrás δ^l, que son las cantidades son necesarias para calcular el gradiente.

2.3.1. Flujo del gradiente

Para simplificar la explicación de este punto supondremos que la función de activación es de clase $C^1(\mathbb{R})$, sin embargo, esto se puede generalizar tomando subgradientes y gradientes débiles.

Definición 2.3.1.
*Un **campo vectorial** en $U \subseteq \mathbb{R}^n$, con U un subconjunto abierto, esta definido por una función $F : U \to \mathbb{R}^n$.*

Sean $f_\cdot : \mathbb{R}^P \to \mathcal{F}$, la función dada por el mapeo $\theta \mapsto f_\theta$, y $\mathcal{L} : \mathcal{F} \to \mathbb{R}$ una función de costo, tal que la composición $h = \mathcal{L} \circ f : \mathbb{R}^P \to \mathbb{R}$ es de clase $C^1\left(\mathbb{R}^P\right)$. Definimos el siguiente campo vectorial:

$$F(\mathbf{x}) = -\nabla h(\mathbf{x}). \tag{2.1}$$

Dado que h es continuamente diferenciable, entonces F define un tipo de campo vectorial al que llamamos **campo vectorial conservativo**.

Una solución del campo vectorial dado por (2.1), con condición inicial $\theta(0) = \mathbf{x}_0$, es una función $\theta : I \subseteq \mathbb{R} \to \mathbb{R}^P$, para un intervalo abierto I, que aparte de la condición inicial también satisface:

$$\frac{d}{dt}\theta\left(t\right) = -\nabla h\left(\theta\left(t\right)\right).$$

Tomando $\mathbf{x}_0$ igual a la inicialización de los pesos de nuestro modelo, podemos garantizar la existencia de una única solución para una vecindad de $\mathbf{x}_0$, gracias al **teorema de existencia y unicidad**. En el contexto de redes neuronales llamamos a esta solución como **flujo del gradiente**.

Pasar del descenso de gradiente (GD) al flujo del gradiente (FG) (de una optimización discreta a una continua) tiene la ventaja de que la teoría matemática esta mejor fundamentada y desarrollada, de modo que existe gran interés en saber cuán próxima es la dinámica del GD a la del FG, conforme el tamaño de la tasa de aprendizaje disminuye. La dificultad está en la complejidad que puede tener la función de pérdida compuesta con la red, que casi siempre es no convexa en un grado muy alto.

Capítulo 3

Kernel Tangente Neuronal (NTK)

Los resultados de probabilidad usados en esta sección son parte de la teoría básica y las pruebas pueden hallarse en [22], [17].

3.1. Redes neuronales poco profundas de ancho infinito

Para comenzar este capítulo, introduciremos el trabajo de Radford M. Neal(1996) sobre la distribución a priori de redes neuronales bayesianas poco profundas (una sola capa oculta en el trabajo de Neal) de ancho infinito. En las redes neuronales bayesianas, a diferencia de las ANNs, los pesos dejan de ser fijos y pasan a ser variables aleatorias, con una distribución a priori, la cual se actualiza durante el entrenamiento.

Comenzamos con un modelo de red neuronal $f : \mathbb{R}^{n_0} \longrightarrow \mathbb{R}^{n_L}$, con una capa oculta y con una función de activación acotada. Entonces, matricialmente, el modelo es el siguiente:

$$f(\mathbf{x}) = \mathbf{W}^{(1)}\alpha^{(1)}(\mathbf{x}) + \mathbf{b}^{(1)},$$

Supondremos que todos los pesos corresponden a variables aleatorias independientes con distribuciones normales centradas:

$$\mathbf{W}_{ij}^{(0)} \sim N\left(0, \sigma_{w0}^2\right), \quad \mathbf{W}_{ij}^{(1)} \sim N\left(0, \sigma_{w1}^2\right), \quad \mathbf{b}_i^{(0)} \sim N\left(0, \sigma_{b0}^2\right), \quad \mathbf{b}_i^{(1)} \sim N\left(0, \sigma_{b1}^2\right).$$

Enfoquémonos en una salida $f_i(\mathbf{x})$, cualquiera pero fija, con valores de entrada $\mathbf{x}$, fijos también, tal que:

$$f_i(\mathbf{x}) = \sum_{j=1}^{n_1} \mathbf{W}_{i,j}^{(1)}\alpha_j^{(1)}(\mathbf{x}) + \mathbf{b}_i^{(1)}.$$

Notemos que por independencia:

$$\mathbb{E}\left[\mathbf{W}_{i,j}^{(1)}\alpha_j^{(1)}(\mathbf{x})\right] = \mathbb{E}\left[\mathbf{W}_{i,j}^{(1)}\right]\mathbb{E}\left[\alpha_j^{(1)}(\mathbf{x})\right] = (0)\,\mathbb{E}\left[\alpha_j^{(1)}(\mathbf{x})\right] = 0,$$

$$\begin{aligned} Var\left(\mathbf{W}_{i,j}^{(1)}\alpha_j^{(1)}(\mathbf{x})(x)\right) &= \mathbb{E}\left[\left(\mathbf{W}_{i,j}^{(1)}\alpha_j^{(1)}(\mathbf{x})\right)^2\right] - \mathbb{E}\left[\mathbf{W}_{i,j}^{(1)}\alpha_j^{(1)}(\mathbf{x})\right]^2 \\ &= \mathbb{E}\left[\left(\mathbf{W}_{i,j}^{(1)}\right)^2\right]\mathbb{E}\left[\left(\alpha_j^{(1)}(\mathbf{x})\right)^2\right] \\ &= \sigma_{w1}^2\mathbb{E}\left[\left(\alpha_j^{(1)}(\mathbf{x})\right)^2\right], \end{aligned}$$

donde $\mathbb{E}\left[\left(\alpha_j^{(1)}(\mathbf{x})\right)^2\right] < \infty$, por ser acotada la función de activación, y más aún, $\alpha_j^{(1)}(\mathbf{x})$ y $\alpha_i^{(1)}(\mathbf{x})$ se distribuyen igual para cualesquiera $i, j \in \{1, ..., n_1\}$, por esta razón, y con el fin de simplificar la notación, definimos $\mathcal{V} := \mathbb{E}\left[\left(\alpha_j^{(1)}(\mathbf{x})\right)^2\right]$.

El ancho de la capa oculta de nuestro modelo esta dado por n_1, y es la cantidad que deseamos incrementar, sin embargo, es necesario que esta cantidad se relacione con la varianza para controlar el tamaño de los sumandos en $f_i(\mathbf{x})$, de modo que el límite de las sumas no diverja. Con este fin definimos:

$$\sigma_{w1} = \frac{C_{w1}}{\sqrt{n_1}},$$

en donde C_{w1} es una constante positiva, tal que $Var\left(\mathbf{W}_{i,j}^{(1)}\alpha_j^{(1)}(\mathbf{x})\right) = \frac{C_{w1}^2}{n_1}\mathcal{V}$.

El propósito ahora es probar que $f_i(\mathbf{x})$ converge en ley a una distribución normal cuando $n_1 \to \infty$. Para esto necesitamos una serie de definiciones y resultados de la teoría de la probabilidad que en seguida enunciamos.

Proposición 3.1.1 (Desigualdad Chebyshev-Bienaymé)**.**
Sea X una variable aleatoria con media μ y varianza σ^2 finita, entonces:

$$\mathbb{P}\left[|X - \mu| \geq \epsilon\right] \leq \frac{\sigma^2}{\epsilon^2}, \quad \forall \epsilon > 0.$$

Definición 3.1.1.
*A un conjunto de variables aleatorias de la forma $\{\psi_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$ la nombramos como **arreglo triangular de variables aleatorias**.*

$$\begin{array}{cccccc} X_{1,1} & & & & & \\ X_{2,1} & X_{2,2} & & & & \\ X_{3,1} & X_{3,2} & X_{3,3} & & & \\ \cdots & \cdots & \cdots & \cdots & & \\ X_{n,1} & X_{n,2} & X_{n,3} & \cdots & X_{n,n} & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{array}$$

Definición 3.1.2.
*Sea $\{\psi_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$, un arreglo triangular de variables aleatorias, decimos que es un **arreglo nulo** si:*

- *$\{\psi_{n,k}\}_{k=1}^n$ son variables aleatorias independientes, para toda $n \in \mathbb{N}$.*

- $\sup_k \mathbb{E}\left[|\psi_{n,k}| \, ||\psi_{n,k}| \leq 1\right] \to 0$.

Definición 3.1.3.
Sea $\{X_i\}_{i\in\Lambda}$ un conjunto de variables aleatorias, con Λ un conjunto de índices. Decimos que tal conjunto es un ***proceso gausiano*** *si para todo $\{X_i\}_{i\in\Omega} \subseteq \{X_i\}_{i\in\Lambda}$ subconjunto finito, el vector $(X_i)_{i\in\Omega}$ tiene distribución gausiana multivariada.*

La media y matriz de covarianzas de un proceso gausiano están dadas por las funciones $\mu(\cdot) : \Lambda \to \mathbb{R}$ y $K(\cdot,\cdot) : \Lambda \times \Lambda \to \mathbb{R}$, respectivamente, en donde K es un kernel definido positivo (en el sentido de Moore). Todo proceso se define por estas dos funciones de modo que el proceso se escribe como:

$$\mathcal{GP}(\mu, K).$$

Teorema 3.1.1 (Convergencia gaussiana, Feller, Lévi)**.**
Sea $\{\psi_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$ un arreglo nulo y ψ una variable aleatoria que se distribuye $\mathcal{N}(b,c)$, con c y b constantes. Entonces $\sum_k \psi_{n,k} \xrightarrow{d} \psi$ si y solo si estas tres condiciones se cumplen:

1. $\sum_k \mathbb{P}\left[|\psi_{n,k}| > \epsilon\right] \to 0 \quad$ *para todo $\epsilon > 0$;*
2. $\sum_k \mathbb{E}\left[\psi_{n,k} \, ||\psi_{n,k}| \leq 1\right] \to b$;
3. $\sum_k Var\left[\psi_{n,k} \, ||\psi_{n,k}| \leq 1\right] \to c$.

La prueba de este teorema se puede encontrar en [17], capítulo 4.

Definimos $\psi_{n,k} := \mathbf{W}^{(1)}_{i,k}\alpha^{(1)}_k(\mathbf{x})$, con n el ancho de la capa oculta de la red a la que corresponde la variable aleatoria $\mathbf{W}_{\cdot,\cdot}$; notemos que podemos obviar el subíndice i por tratarse de variables aleatorias independientes e idénticamente distribuidas. Es claro que $\{\psi_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$ es un arreglo triangular de variables aleatorias independientes e idénticamente distribuidas con media igual a cero, para cada $n \in \mathbb{N}$, tal que la condición:

$$\sup_k \mathbb{E}\left[|\psi_{n,k}| \, ||\psi_{n,k}| \leq 1\right] \to 0,$$

es equivalente a

$$\mathbb{E}\left[|\psi_{n,k}| \, ||\psi_{n,k}| \leq 1\right] \to 0, \tag{3.1}$$

al ser variables aleatorias independientes e idénticamente distribuidas las de las filas en el arreglo triangular.

Por otro lado sabemos que:

$$\mathbb{E}\left[(|\psi_{n,k}| - 0)^2\right] = \mathbb{E}\left[(\psi_{n,k} - 0)^2\right] = \frac{C^2_{w1}}{n}\mathcal{V} \to 0,$$

es decir, $|\psi_{n,k}| \to 0$ en media cuadrática, tal que podemos concluir que la Ecuación 3.1 se satisface y por lo tanto $\{\psi_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$ es un arreglo nulo. Resta probar que satisface las condiciones del Teorema 3.1.1 para concluir la convergencia en distribución.

1. Siguiendo a [17], capítulo 4, la primer condición equivale a:

$$\sup_k |\psi_{n,k}| \xrightarrow{P} 0,$$

y para nuestro caso particular lo podemos reescribir como:

$$|\psi_{n,k}| \xrightarrow{P} 0.$$

Por la desigualdad Chebyshev-Bienaymé tenemos que:

$$\mathbb{P}\left[|\psi_{n,k}| \geq \epsilon\right] \leq \frac{C_{w1}^2 \mathcal{V}}{n\epsilon^2}, \qquad \forall \epsilon > 0,$$

tal que $\forall k \in \mathbb{N}$ y $\forall \epsilon > 0$:

$$\mathbb{P}\left[|\psi_{n,k}| \geq \epsilon\right] \to 0.$$

2. Gracias a que tenemos la convergencia a 0 en media cuadrática podemos reescribir la condición 2 como:

$$\sum_k \mathbb{E}\left[\psi_{n,k}\right] \to b.$$

Notemos que:

$$\sum_k \mathbb{E}\left[\psi_{n,k}\right] = \sum_k 0 \to 0,$$

tal que la segunda condición se satisface, para $b = 0$.

3. Por el mismo argumento del inciso anterior podemos reescribir la condición 3 como:

$$\sum_k Var\left[\psi_{n,k}\right] \to c.$$

Por otro lado:

$$\sum_k Var\left[\psi_{n,k}\right] = \sum_k \frac{C_{w1}^2 \mathcal{V}}{n} = C_{w1}^2 \mathcal{V} \to C_{w1}^2 \mathcal{V},$$

tal que la tercer condición se satisface, para $c = C_{w1}^2 \mathcal{V}$

De este modo por el Teorema 3.1.1 podemos concluir que:

$$\sum_{j=1}^{n_1} \mathbf{W}_{i,j}^{(1)} \alpha_j^{(1)}(\mathbf{x}) \xrightarrow{d} \mathcal{N}\left(0, C_{w1}^2 \mathcal{V}\right), \quad \text{conforme } n_1 \text{ tiende a infinito,}$$

y más aún:

$$f_i\left(\mathbf{x}\right) = \sum_{j=1}^{n_1} \mathbf{W}_{i,j}^{(1)} \alpha_j^{(1)}(\mathbf{x}) + \mathbf{b}_i^{(1)} \xrightarrow{d} \mathcal{N}\left(0, \sigma_{b1}^2 + C_{w1}^2 \mathcal{V}\right).$$

A partir de este resultado Neal, en [20], concluye, sin detallar mucho, que distintas salidas de la red, f_i, f_j, con $i \neq j$, tienen una distribución conjunta normal en el límite cuando

$n_1 \to \infty$, de modo que cuando $n_1 \to \infty$, $f(\cdot)$, en su inicialización, converge en distribución a un proceso gausiano centrado, $\mathcal{GP}(\mathbf{0}, \Sigma)$, donde matriz de covarianzas dada por:

$$\Sigma\left(f_i(\mathbf{x}) f_j(\mathbf{x}')\right) = \begin{cases} \sigma_{b1}^2 + C_{w1}^2 \mathbb{E}\left[\alpha_1^{(1)}(\mathbf{x})\alpha_1^{(1)}(\mathbf{x}')\right] & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}. \tag{3.2}$$

Un tratamiento detallado de esto último podemos encontrarlo en [12]. Nosotros probaremos esto más adelante, bajo condiciones particulares, pero equivalentes.

3.2. Kernel Tangente Neuronal y su convergencia

En adelante nos basaremos en el artículo [16], en las 4 primeras secciones, enfocándonos en desarrollar una explicación clara de los primeros resultados del mismo.

Comencemos dando una parametrización oportuna de una red neuronal artificial $f(\cdot) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$, con función de activación σ acotada de clase $C^1(\mathbb{R})$:

$$\alpha^{(0)}(\mathbf{x}) = \mathbf{x}$$

$$\tilde{\alpha}^{(l+1)}(\mathbf{x}) = \frac{1}{\sqrt{n_l}} \mathbf{W}^{(l)} \alpha^{(l)}(\mathbf{x}) + \beta \mathbf{b}^{(l)}$$

$$\alpha^{(l)}(\mathbf{x}) = \sigma\left(\tilde{\alpha}^{(l)}(\mathbf{x})\right),$$

en donde cada peso es inicializado a partir de variables aleatorias independientes con distribución normal estándar , $\mathbf{W}_{ij}^{(l)}, \mathbf{b}_i^{(l)} \sim \mathcal{N}(0,1)$, y $\beta > 0$.

Notemos que la distribución de la inicialización de la red coincide con la expuesta en la sección anterior, ya que:

$$Si \quad \mathbf{X} \sim \mathcal{N}(0, a) \Longrightarrow c\mathbf{X} \sim \mathcal{N}\left(0, c^2 a\right), \quad \forall c > 0,$$

sin embargo, el cálculo de las derivadas con esta parametrización conserva el escalamiento, además, nos permitirá escribir de mejor manera ciertas ecuaciones.

Consideremos ahora el espacio de funciones $\mathcal{F} := \{h : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}\}$, y un conjunto de entrenamiento $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N \subseteq \mathbb{R}^{n_0} \times \mathbb{R}^{n_L}$, cuyo espacio muestral de características, $\mathcal{X} \subseteq \mathbb{R}^{n_0}$, tiene una distribución p^{in}. Sobre el espacio $\mathcal{F}$ consideremos la siguiente forma bilineal (que a su vez induce una seminorma):

$$\langle f, g \rangle_{p^{in}} := \mathbb{E}_{\mathbf{x} \sim p^{in}}\left[f(\mathbf{x})^T g(\mathbf{x})\right].$$

En nuestro caso desconocemos p^{in}, pero es posible aproximarla (sustituirla) mediante la distribución empírica[1], que esta dada por:

$$p^0 = \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_i},$$

[1]Siempre que el conjunto de entrenamiento sea una muestra representativa.

tal que la forma bilineal anterior podemos reescribirla como:

$$\langle f, g\rangle_{p^0} := \frac{1}{N}\sum_{i=1}^{N} f\left(\mathbf{x}_i\right)^T g\left(\mathbf{x}_i\right).$$

Otro espacio útil será $\mathcal{F}^* := \{\langle d, \cdot\rangle_{p_0} : \mathcal{F} \to \mathbb{R} | d \in \mathcal{F}\}$.
La forma bilineal antes mencionada es degenerada, ya que depende de los valores de las funciones en un número finito de puntos en el dominio, pero es posible considerar otro espacio, para el que esto no ocurra. Consideremos la siguiente relación de equivalencia:

$$f \sim g \iff \langle f - g, \cdot\rangle_{p^0} \equiv 0 \iff f(\mathbf{x}_i) = g(\mathbf{x}_i), \quad \forall i \in \{1, ..., N\},$$

entonces $\mathcal{F}/\sim$ es un espacio vectorial finito ya que todo $\overline{f} \in \mathcal{F}/\sim$ esta definido de manera única por sus valores en $\{\mathbf{x}_i\}_{i=1}^N$:

- Denotemos con $\overline{f}_{ik} \in \mathcal{F}/\sim$ a la función (clase de equivalencia) cero en todas partes, excepto cuando se evalúa en $\mathbf{x}_i$, en este caso devuelve el vector cuya única entrada distinta de cero es la $k-$ésima, con valor de 1. Entonces $\forall \overline{f} \in \mathcal{F}/\sim$ existe $\{\alpha_{ik}\}_{i\in\{1,...,N\}, k\in\{1,...,n_L\}} \subseteq \mathbb{R}$, tal que:

$$\overline{f} = \sum_{i=1}^{N}\sum_{k=1}^{n_L} \alpha_{ik}\overline{f}_{ik}.$$

Es así que $\tilde{\mathcal{F}} := \mathcal{F}/\sim$ es un espacio de Hilbert por ser finito dimensional con producto interior, $\langle\cdot,\cdot\rangle_{p_0}$, y más aún $\mathcal{F}^* \cong \tilde{\mathcal{F}}$ por el mapeo:

$$\langle f, \cdot\rangle_{p^0} \mapsto \overline{f}.$$

En lo sucesivo escribiremos sin $\bar{\cdot}$ a los elementos de $\tilde{\mathcal{F}}$, sin riesgo a cometer errores cruciales al confundirlos con elementos de $\mathcal{F}$, pues las funciones de $\mathcal{F}$ solo nos interesan por su evaluación en el conjunto de entrenamiento.

Por otra parte, sea $K : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L} \times \mathbb{R}^{n_L}$ un kernel definido positivo que satisface $K\left(\mathbf{x}, \mathbf{x}'\right)^T = K\left(\mathbf{x}', \mathbf{x}\right)$, al que nombraremos como *kernel multidimensional* dentro de este capítulo. Entonces, bajo las condiciones de la distribución empírica[2], definimos la siguiente forma bilineal:

$$\langle f, g\rangle_K := \mathbb{E}_{\mathbf{x},\mathbf{x}'\sim p^0}\left[f\left(\mathbf{x}\right)^T K\left(\mathbf{x}, \mathbf{x}'\right) g\left(\mathbf{x}'\right)\right] = \frac{1}{N^2}\sum_{i,j=1}^{N} f\left(\mathbf{x}_i\right)^T K\left(\mathbf{x}_i, \mathbf{x}_j\right) g\left(\mathbf{x}_j\right).$$

Para las dos formas bilineales que hemos descrito hasta ahora, $\langle f, g\rangle_{p^0}$, $\langle f, g\rangle_K$, definimos

$$\|f\|_{p^0} := \langle f, f\rangle_{p^0}, \quad \forall f \in \mathcal{F},$$

$$\|f\|_K := \langle f, f\rangle_K, \quad \forall f \in \mathcal{F}.$$

[2]La distribución empírica nos permite dar una manera de calcular la forma bilineal en función de los datos.

Definición 3.2.1.
Decimos que K está ***definido positivo respecto a la seminorma*** $\|\cdot\|_{p^0}$ *si:*

$$\|f\|_{p^0} > 0 \Longrightarrow \|f\|_K > 0, \quad \forall f \in \mathcal{F}.$$

3.2.1. Kernel gradiente

Sea $\mathcal{L} : \mathcal{F} \to [0, \infty)$ una función de costo. Consideremos el siguiente siguiente mapeo, con f fijo:

$$\phi \mapsto \lim_{\epsilon \to 0} \frac{\mathcal{L}(f + \epsilon\phi) + \mathcal{L}(f)}{\epsilon}, \quad \forall \phi \in \mathcal{F} \cong \tilde{\mathcal{F}}.$$

En cálculo de variaciones este límite se toma con una aproximación lineal (variación lineal)[3], de donde concluyen que el mapeo es lineal, y por ser $\tilde{\mathcal{F}}$ un espacio de Hilbert finito dimensional, entonces el mapeo también es acotado. Así, por el teorema de representación de Riesz, existe un único $\frac{\partial \mathcal{L}}{\partial f} \in \tilde{\mathcal{F}}$ tal que:

$$\langle \frac{\partial \mathcal{L}}{\partial f}, \phi \rangle_{p^0} = \lim_{\epsilon \to 0} \frac{\mathcal{L}(f + \epsilon\phi) + \mathcal{L}(f)}{\epsilon}, \qquad \forall \phi \in \tilde{\mathcal{F}},$$

a $\frac{\partial \mathcal{L}}{\partial f}$ se le conoce como la **derivada funcional** de $\mathcal{L}$ en f, cuyo significado puede compararse con el gradiente al hablar de derivadas direccionales, $\mathbf{D}_{\mathbf{v}} g(\mathbf{x}) = \nabla g(\mathbf{x}) \cdot \mathbf{v} = \langle \nabla g(\mathbf{x}), \mathbf{v} \rangle$. Notemos que $\frac{\partial \mathcal{L}}{\partial f}$ formalmente es una clase de equivalencia, sin embargo nos interesa solo por su evaluación en los datos de entrenamiento.

Retomando K, el kernel multidimesional, consideremos la función $K(\mathbf{x}, \cdot) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L} \times \mathbb{R}^{n_L}$, con $\mathbf{x}$ fija, entonces con $K_{i,\cdot}(\mathbf{x}, \cdot) : \mathbb{R}^{n_0} \to \mathbb{R}^{n_L}$ denotamos a la función que va de un vector $\mathbf{x}' \in \mathbb{R}^{n_0}$ al vector correspondiente a la $i-$ésima fila de la matriz $K(\mathbf{x}, \mathbf{x}')$. Ahora definamos la función $\Phi_K : \mathcal{F}^* \to \mathcal{F}$, tal que para toda $\mu = \langle d, \cdot \rangle \in \mathcal{F}^*$, $\Phi_K(\mu)$ esta dada por:

$$\Phi_K(\mu)(\mathbf{x}) = (\mu(K_{i,\cdot}(\mathbf{x}, \cdot)))_{i=1,\ldots,n_L} = (\langle d, K_{i,\cdot}(\mathbf{x}, \cdot) \rangle_{p^0})_{i=1,\ldots,n_L}, \quad \forall \mathbf{x} \in \mathbb{R}^{n_0}.$$

A $\Phi_K\left(\langle \frac{\partial \mathcal{L}}{\partial f}, \cdot \rangle_{p^0}\right)$, con f una red neuronal, lo llamamos **kernel gradiente** y la denotamos con $\nabla_K \mathcal{L}|_f$. Cuando una red neuronal $f_{\theta(\cdot)} : [0, a) \to \mathcal{F}$, con sus parámetros en función de tiempo y $a \in (0, \infty)$, satisface la ecuación:

$$\frac{d}{dt} f_{\theta(t)}(\mathbf{x}) = -\nabla_K \mathcal{L}|_{f_{\theta(t)}}(\mathbf{x}), \qquad \forall \mathbf{x} \in \mathbb{R}^{n_0},$$

decimos que la red evoluciona en el tiempo siguiendo el kernel gradiente definido por a K; $\theta(\cdot)$ es lo que en el capítulo anterior llamamos *flujo del gradiente*. Notemos que:

[3]Para mayor información puedo consultarse [13].

$$
\begin{aligned}
-\nabla_K \mathcal{L}|_f(\mathbf{x}) = -\Phi_K\left(\langle\frac{\partial\mathcal{L}}{\partial f},\cdot\rangle_{p^0}\right)(\mathbf{x}) &= \left(-\langle\frac{\partial\mathcal{L}}{\partial f}, K_{i,\cdot}(\mathbf{x},\cdot)\rangle_{p^0}\right)_{i=1,\dots,n_L} \\
&= \left(\frac{-1}{N}\sum_{j=1}^{N} K_{i,\cdot}(\mathbf{x},\mathbf{x}_j)^T\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_j)\right)_{i=1,\dots,n_L} \\
&= \frac{-1}{N}\sum_{j=1}^{N}\left(K_{i,\cdot}(\mathbf{x},\mathbf{x}_j)^T\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_j)\right)_{i=1,\dots,n_L} \\
&= \frac{-1}{N}\sum_{j=1}^{N} K(\mathbf{x},\mathbf{x}_j)\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_j),
\end{aligned}
$$

tal que la evolución de f en $\mathcal{F}$ estaría determinada por una máquina kernel.

Suponiendo lo anterior, podemos calcular la *derivada* de $\mathcal{L}$ en $\mathcal{F}$ respecto al tiempo en dirección del kernel gradiente:

$$
\begin{aligned}
\partial_t\mathcal{L}|_{f(t)} = \langle\frac{\partial\mathcal{L}}{\partial f}, -\nabla_K\mathcal{L}|_f\rangle_{p^0} &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_i)^T\left(-\nabla_K\mathcal{L}|_f(\mathbf{x}_i)\right) \\
&= \frac{1}{N}\sum_{i,j=1}^{N}\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_i)^T\left(\frac{-1}{N}\sum_{j=1}^{N}K(\mathbf{x},\mathbf{x}_j)\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_j)\right) \\
&= \frac{-1}{N^2}\sum_{i,j=1}^{N}\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_i)^T K(\mathbf{x}_i,\mathbf{x}_j)\frac{\partial\mathcal{L}}{\partial f}(\mathbf{x}_j) \\
&= -\langle\frac{\partial\mathcal{L}}{\partial f},\frac{\partial\mathcal{L}}{\partial f}\rangle_K \\
&= -\|\frac{\partial\mathcal{L}}{\partial f}\|_K.
\end{aligned}
$$

Si K es definido positivo con respecto a $\|\cdot\|_{p^0}$, entonces $\mathcal{L}|_{f(t)}$ es una función decreciente con puntos críticos en donde la derivada funcional tiene norma igual a cero respecto $\|\cdot\|_{p^0}$. Teóricamente, eligiendo una función de costo adecuada (convexa y acotada) es posible tener un entrenamiento convexo en $\mathcal{F}$.

3.2.2. NTK de una red neuronal artificial

Consideremos una función de pérdida $\mathcal{L}:\mathbb{R}^P\to[0,\infty)$ que sea el promedio de los errores por dato de entrenamiento, $c:\mathbb{R}^P\times\mathbb{R}^{n_0}\times\mathbb{R}^{n_L}\to\mathbb{R}$, tal que:

$$
\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} c\left(f(\theta,\mathbf{x}_i), y_i\right)
$$

con c continuamente diferenciable, tal que la función $\mathcal{L}$ también lo es. Para abreviar notación definimos $c_i(\theta) := c\left(f(\theta,\mathbf{x}_i), y_i\right)$.

Por cálculo multivariable tenemos:

$$\nabla_\theta \mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \frac{\partial}{\partial\theta_1} f_1(\theta, \mathbf{x}_i) & \cdots & \frac{\partial}{\partial\theta_1} f_{n_L}(\theta, \mathbf{x}_i) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta_P} f_1(\theta, \mathbf{x}_i) & \cdots & \frac{\partial}{\partial\theta_P} f_{n_L}(\theta, \mathbf{x}_i) \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial f_1} c_i(\theta) \\ \vdots \\ \frac{\partial}{\partial f_{n_L}} c_i(\theta) \end{bmatrix}$$

A partir de esto definimos el siguiente campo vectorial en $\mathbb{R}^P$:

$$F(\theta) = -\nabla_\theta \mathcal{L}(\theta).$$

Por el teorema de existencia y unicidad, para toda condición inicial $\mathbf{x}_0$ existe una única solución $\theta(\cdot) : [0, a) \to \mathbb{R}^P$, para alguna $a \in (0, \infty) \cup \{\infty\}$:

$$\frac{d}{dt}\theta(t) = -\nabla_\theta \mathcal{L}(\theta(t)), \quad \theta(0) = \mathbf{x}_0.$$

Entonces por regla de la cadena, la evolución de la red neuronal a través del flujo del gradiente esta dado por:

$$\begin{aligned} \frac{d}{dt} f(\theta(t), \mathbf{x}) &= \frac{d}{d\theta} f(\theta(t), \mathbf{x}) \frac{d}{dt}\theta(t) \\ &= \begin{bmatrix} \frac{\partial}{\partial\theta_1} f_1(\theta, \mathbf{x}) & \cdots & \frac{\partial}{\partial\theta_P} f_1(\theta, \mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta_1} f_{n_L}(\theta, \mathbf{x}) & \cdots & \frac{\partial}{\partial\theta_P} f_{n_L}(\theta, \mathbf{x}) \end{bmatrix} \\ &\qquad \left(\frac{-1}{N}\sum_{i=1}^{n} \begin{bmatrix} \frac{\partial}{\partial\theta 1} f_1(\theta, \mathbf{x}_i) & \cdots & \frac{\partial}{\partial\theta 1} f_{n_L}(\theta, \mathbf{x}_i) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial\theta P} f_1(\theta, \mathbf{x}_i) & \cdots & \frac{\partial}{\partial\theta P} f_{n_L}(\theta, \mathbf{x}_i) \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial f_1} c_i(\theta) \\ \vdots \\ \frac{\partial}{\partial f_{n_L}} c_i(\theta) \end{bmatrix} \right) \\ &= \frac{-1}{N}\sum_{i=1}^{N} \mathbf{J} f(\theta(t), \mathbf{x}) \mathbf{J} f(\theta(t), \mathbf{x}_i)^T \nabla c_i(\theta) \\ &= \frac{-1}{N}\sum_{i=1}^{N} \Theta(\mathbf{x}, \mathbf{x}_i) \nabla c_i(\theta), \end{aligned}$$

donde $\mathbf{J} f(\theta(t), \mathbf{x})$ denota al jacobiano de $f(\theta(t), \mathbf{x})$ en función de θ.

La función $\Theta : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}^{n_L} \times \mathbb{R}^{n_L}$, dada por:

$$\Theta(\mathbf{x}, \mathbf{x}') = \mathbf{J} f(\theta, \mathbf{x}) \mathbf{J} f(\theta, \mathbf{x}')^T,$$

es definida positiva en el sentido de Moore ya que :

$$\sum_{i=1}^{r}\sum_{j=1}^{r} v_i^T \Theta(s_i, s_j) v_j = \sum_{i=1}^{r}\sum_{j=1}^{r} v_i^T \mathbf{J} f(\theta, s_i) \mathbf{J} f(\theta, s_j)^T v_j$$

$$
\begin{aligned}
&= \sum_{i=1}^{r}\sum_{j=1}^{r}\left(\mathbf{J}f\left(\theta, s_i\right)^T v_i\right)^T\left(\mathbf{J}f\left(\theta, s_j\right)^T v_j\right)\\
&= \left(\sum_{i=1}^{r}\mathbf{J}f\left(\theta, s_i\right)^T v_i\right)^T\left(\sum_{j=1}^{r}\mathbf{J}f\left(\theta, s_j\right)^T v_j\right)\\
&= \|\sum_{i=1}^{r}\mathbf{J}f\left(\theta, s_i\right)^T v_i\|^2 \geq 0,
\end{aligned}
$$

para todo $\{s_1, ..., s_r\} \subseteq \mathbb{R}^{n_0}$, $\{v_1, ..., v_r\} \subseteq \mathbb{R}^{n_0}$, $r \in \mathbb{N}$. Por otro lado:

$$\Theta\left(\mathbf{x}, \mathbf{x}'\right)^T = \mathbf{J}f\left(\theta, \mathbf{x}'\right)\mathbf{J}f\left(\theta, \mathbf{x}\right)^T = \Theta\left(\mathbf{x}', \mathbf{x}\right),$$

tal que Θ es un kernel reproductor al cual llamamos **Kernel Neuronal Tangente**. Retomando lo descrito páginas anteriores, la evolución de la red a lo largo del flujo del gradiente esta descrita por el kernel gradiente asociado a Θ, con $\frac{\partial \mathcal{L}}{\partial f} = \nabla c\left(f\left(\theta, \cdot\right)\right)$, y además, si Θ es definido positivo respecto a $\|\cdot\|_{p^0}$, entonces:

$$\partial_t \mathcal{L}|_{f(t)} = \langle \frac{\partial \mathcal{L}}{\partial f}, -\nabla_K \mathcal{L}|_f \rangle_{p^0} = -\|\nabla c\left(f\left(\theta, \cdot\right)\right)\|_\Theta$$

En este punto es importante notar que el kernel Θ tiene un componente aleatorio debido a la inicialización aleatoria de los pesos, que también influye en la condición inicial del flujo del gradiente.

3.3. Convergencia del NTK

Esta última sección se compone en dos resultados expuestos de manera escueta en [16], y que aquí desarrollaremos detalladamente, comenzando por la convergencia de la red en su estado inicial a un proceso gausiano, conforme el ancho de las capas ocultas tiende a infinito, y después tenemos la convergencia de Θ a un kernel determinista en la inicialización de la red, bajo el mismo límite.

Para la prueba utilizaremos una definición y dos resultados clásicos de probabilidad que en seguida enunciamos.

Definición 3.3.1 (Distribución normal multivariada)**.**
Sea $X = (X_1, ..., X_n)$ un vector aleatorio. Decimos que X se sigue una ***distribución normal multivariada*** *si alguna de las siguientes condiciones equivalentes se satisface:*

1. *Existen $Z = (z_1, ..., z_m)$, un vector de variables aleatorias indenpendientes con distribución normal estándar, $A \in \mathbb{R}^{n\times m}$ y $\mu \in \mathbb{R}^n$, tales que:*

$$X = AZ + \mu$$

2. *Toda combinación lineal de $\{X_i\}_{i=1}^n$ tiene distribución normal.*

Proposición 3.3.1 (Ley débil de los grandes números)**.**
Sea $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ *una sucesión de variables aleatorias, con* $\mathbb{E}[\boldsymbol{X}_i] = \mu$, $Var(\boldsymbol{X}_i) = \sigma^2 < \infty$, $\forall i \in \mathbb{N}$, *entonces:*

$$\sum_{i=1}^{n} \frac{\boldsymbol{X}_i}{n} \xrightarrow{P} \mu, \quad n \to \infty.$$

Teorema 3.3.1 (Teorema multivariado del límite central)**.**
Sea $(\boldsymbol{X}_i)_{i\in\mathbb{N}}$ *una sucesión de vectores aleatorios independientes e idénticamente distribuidos, con* $\mathbb{E}[\boldsymbol{X}_i] = \mu < \infty$ *y matriz de covarianza* Σ, *entonces:*

$$\sum_{i=1}^{n} \frac{(\boldsymbol{X}_i - \mu)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \Sigma), \quad n \to \infty.$$

Proposición 3.3.2 (Convergencia en distribución a un proceso gausiano)**.**
En la inicialización de la red, si el ancho de las capas ocultas tiende a infinito secuencialmente desde la primera hasta la última, entonces los vectores $\{\left(\tilde{\alpha}_k^{(L)}(\boldsymbol{x}_1), ..., \tilde{\alpha}_k^{(L)}(\boldsymbol{x}_N)\right)\}_{k=1}^{n_L}$ *convergen en ley a una distribución normal centrada multivariada con matriz de covarianzas* $\Sigma^{(L)}$ *definida recursivamente como:*

$$\Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{n_0}\boldsymbol{x}^T\boldsymbol{x}' + \beta^2,$$

$$\lambda^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \begin{bmatrix} \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}) & \Sigma^{(l)}(\boldsymbol{x}, \boldsymbol{x}') \\ \Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}) & \Sigma^{(l)}(\boldsymbol{x}', \boldsymbol{x}') \end{bmatrix},$$

$$\Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\left(\tilde{\alpha}^l(\boldsymbol{x}), \tilde{\alpha}^l(\boldsymbol{x}')\right) \sim \mathcal{N}\left(0, \lambda^{(l)}\right)} \left[\alpha(\boldsymbol{x})\,\alpha(\boldsymbol{x}')\right] + \beta^2,$$

y los vectores son independientes entre sí, en el límite.

Demostración. Comenzaremos definiendo notación que nos simplificará la escritura:

- $\tilde{\alpha}_i^0(\mathbf{x}_j) = (\mathbf{x}_j)_i$.
- $T = \{\mathbf{x}_i\}_{i=1}^N$, las características de entrada de nuestro conjunto de entrenamiento.
- $\cdot_k$, el subíndice k indica una salida cualquiera de la red en alguna capa.
- $\lambda_l := \left(\tilde{\alpha}^l(\mathbf{x}), \tilde{\alpha}^l(\mathbf{x}')\right) \sim \mathcal{N}\left(0, \lambda^{(l)}\right)$.
- $\mathbf{V}_{k,l} = \left(\tilde{\alpha}_k^l(\mathbf{x}_1), ..., \tilde{\alpha}_k^l(\mathbf{x}_N)\right)$.
- $\mathbf{V}_l = \left(\tilde{\alpha}_1^l(\mathbf{x}_1), ..., \tilde{\alpha}_{n_l}^l(\mathbf{x}_1), ..., \tilde{\alpha}_1^l(\mathbf{x}_N), ..., \tilde{\alpha}_{n_l}^l(\mathbf{x}_N)\right)$.

Por inducción sobre la profundidad, L, de la red tenemos lo siguiente.

1. **Caso base**: $L = 1$
   En este caso nuestra red se reduce a algo de la forma:

$$\frac{1}{\sqrt{n_0}}\mathbf{W}^{(0)}\mathbf{x} + \beta\mathbf{b}^{(0)}$$

Es claro que las salidas de la red son independientes e idénticamente distribuidas (con media igual a cero), de modo que los vectores aleatorios $\{\mathbf{V}_{i,1}\}_{i=1}^{n_0}$ son independientes e idénticamente distribuidos, con distribución normal centrada multivariada, pues:

$$\mathbf{V}_{k,1} = \begin{bmatrix} (\mathbf{x}+_1)^T \\ \vdots \\ (\mathbf{x}+_N)^T \end{bmatrix} \underbrace{\begin{bmatrix} \mathbf{W}_{k,1}^{(0)} \\ \vdots \\ \mathbf{W}_{k,n_0}^{(0)} \\ \mathbf{b}_k^{(0)} \end{bmatrix}}_{\mathbf{Z}}$$

en donde el vector $(\mathbf{x}+_i)^T = \left(\frac{1}{\sqrt{n_0}}(\mathbf{x}_i)_1, ..., \frac{1}{\sqrt{n_0}}(\mathbf{x}_i)_{n_0}, \beta\right)$ es constante y $\mathbf{Z}$ es un vector de variables aleatorias con distribución normal estándar.

Dado que cada elemento del vector $\mathbf{V}_{k,1}$ esta especificado por el punto en que se evalúa, la matriz de covarianzas podemos verla en función de estos puntos:

$$\begin{aligned}
\Sigma^{(1)}(\mathbf{x}, \mathbf{x}') &= Cov\left(\tilde{\alpha}_k^1(\mathbf{x}), \tilde{\alpha}_k^1(\mathbf{x}')\right) \\
&= \mathbb{E}\left[\left(\frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\mathbf{W}_{k,i}^{(0)}(\mathbf{x})_i + \beta\mathbf{b}_k^{(0)}\right)\left(\frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\mathbf{W}_{k,j}^{(0)}(\mathbf{x}')_j + \beta\mathbf{b}_k^{(0)}\right)\right] \\
&= \mathbb{E}\left[\frac{1}{n_0}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\mathbf{W}_{k,i}^{(0)}(\mathbf{x})_i\mathbf{W}_{k,j}^{(0)}(\mathbf{x}')_j + \frac{1}{\sqrt{n_0}}\sum_{i=1}^{n_0}\mathbf{W}_{k,i}^{(0)}(\mathbf{x})_i\beta\mathbf{b}_k^{(0)}\right. \\
&\qquad \left. + \frac{1}{\sqrt{n_0}}\sum_{j=1}^{n_0}\mathbf{W}_{k,j}^{(0)}(\mathbf{x}')_j\beta\mathbf{b}_k^{(0)} + \beta^2\left(\mathbf{b}_k^{(0)}\right)^2\right] \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}(\mathbf{x})_i(\mathbf{x}')_j\mathbb{E}\left[\mathbf{W}_{k,i}^{(0)}\mathbf{W}_{k,j}^{(0)}\right] + \frac{\beta}{\sqrt{n_0}}\sum_{j=1}^{n_0}(\mathbf{x}')_j\mathbb{E}\left[\mathbf{W}_{k,j}^{(0)}\mathbf{b}_k^{(0)}\right] \\
&\quad + \frac{\beta}{\sqrt{n_0}}\sum_{i=1}^{n_0}(\mathbf{x})_i\mathbb{E}\left[\mathbf{W}_{k,i}^{(0)}\mathbf{b}_k^{(0)}\right] + \beta^2\mathbb{E}\left[\left(\mathbf{b}_k^{(0)}\right)^2\right] \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}(\mathbf{x})_i(\mathbf{x}')_j Cov\left(\mathbf{W}_{k,i}^{(0)}\mathbf{W}_{k,j}^{(0)}\right) \\
&\quad + \frac{\beta}{\sqrt{n_0}}Cov\left(\mathbf{W}_{k,j}^{(0)}, \mathbf{b}_k^{(0)}\right)\left(\sum_{j=1}^{n_0}(\mathbf{x})_i + (\mathbf{x}')_j\right) + \beta^2 Var\left(\mathbf{b}_k^{(0)}\right) \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0}(\mathbf{x})_i(\mathbf{x}')_i + 0 + 0 + \beta^2 \\
&= \frac{1}{n_0}\mathbf{x}^T\mathbf{x}' + \beta^2
\end{aligned}$$

Aquí no hay un proceso límite, de modo que tenemos probado el caso base.

2. **Hipótesis inductiva**:
   Supongamos que para $L = R - 1$ se satisface que tomando límites secuenciales:

   a) Salidas distintas de la red son independientes entre sí:

$$\tilde{\alpha}_i^{(R-1)}(\mathbf{x}) \perp \tilde{\alpha}_j^{(R-1)}(\mathbf{x}'), \qquad \forall \mathbf{x}, \mathbf{x}' \in T, \text{ con } i \neq j.$$

   b) Los vectores $\{\mathbf{V}_{i,R-1}\}_{i=1}^{n_{R-1}}$ son independientes e idénticamente distribuidos, con distribución normal multivariada centrada y matriz de covarianza dada por:

$$\Sigma^{(R-1)}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\lambda_{R-2}}\left[\alpha_k^{(R-2)}(\mathbf{x})\,\alpha^{(R-2)}(\mathbf{x}')\right] + \beta^2.$$

3. **Paso inductivo:**
   Sean $L = R$. Notemos que:

$$\begin{aligned}\mathbf{V}_{k,R} &= \left(\tilde{\alpha}_k^{(R)}(\mathbf{x}_1), ..., \tilde{\alpha}_k^{(R)}(\mathbf{x}_N)\right)^T \\ &= \underbrace{\frac{1}{\sqrt{n_{R-1}}}\left(\sum_{i=1}^{n_{R-1}} \begin{bmatrix} \mathbf{W}_{k,i}^{(R-1)}\alpha_i^{(R-1)}(\mathbf{x}_1) \\ \vdots \\ \mathbf{W}_{k,i}^{(R-1)}\alpha_i^{(R-1)}(\mathbf{x}_N) \end{bmatrix}\right)}_{\Gamma} + \beta \underbrace{\begin{bmatrix} \mathbf{b}_k^{(R-1)} \\ \vdots \\ \mathbf{b}_k^{(R-1)} \end{bmatrix}}_{\Lambda},\end{aligned}$$

   en donde los vectores de la suma en Γ son independientes e idénticamente distribuidos, pues por hipótesis los vectores $\{\mathbf{V}_{k,R-1}\}_{k=1}^{n_{R-1}}$ lo son. Más aún, cada sumando tiene media cero, ya que:

$$\mathbb{E}\left[\mathbf{W}_{k,i}^{(R-1)}\alpha_i^{(R-1)}(\mathbf{x}_1)\right] = \mathbb{E}\left[\mathbf{W}_{k,i}^{(R-1)}\right]\mathbb{E}\left[\alpha_i^{(R-1)}(\mathbf{x}_1)\right] = (0)\,\mathbb{E}\left[\alpha_i^{(R-1)}(\mathbf{x}_1)\right] = 0,$$

   entonces por el teorema multivariado del límite central Γ converge en ley a una distribución normal multivariada centrada,

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}),$$

   cuando $n_{R-1} \to \infty$. Así, por definición

$$\begin{matrix} \mathbf{X}_1 = & y_{11}\mathbf{Z}_1 + & \cdots & + y_{1t}\mathbf{Z}_t \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}_N = & y_{N1}\mathbf{Z}_1 + & \cdots & + y_{Nt}\mathbf{Z}_t \end{matrix}, \qquad \mathbf{Z}_i \sim \mathcal{N}(0,1).$$

   Por otra parte, el vector que se multiplica por beta no varía en el límite, tal que el vector $\mathbf{V}_{k,R}$ converge en distribución al vector $\mathbf{W}$, cuando $n_{R-1} \to \infty$, el cual esta dado por:

$$\begin{array}{ccccccc} \mathbf{W}_1 = & \beta\mathbf{Z}_0 & +y_{11}\mathbf{Z}_1+ & \cdots & y_{1t}\mathbf{Z}_t & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & , & \mathbf{Z}_i \sim \mathcal{N}(0,1), \\ \mathbf{W}_N = & \beta\mathbf{Z}_0 & +y_{N1}\mathbf{Z}_1+ & \cdots & +y_{Nt}\mathbf{Z}_t & & \end{array}$$

de modo que $\mathbf{W}$ sigue una distribución normal centrada multivariada, $\mathbf{W} \sim \mathcal{N}\left(\mathbf{0}, \Sigma^{(R)}\right)$.

Para el cálculo de la matriz de covarianzas basta analizar el comportamiento de la matriz de covarianzas de $\mathbf{V}_{k,R}$ conforme $n_{R-1} \to \infty$. Para comenzar, consideremos las salidas de la red condicionadas con $\mathbf{V}_{R-1}$, de manera que son independientes e idénticamente distribuidas las salidas, y en particular el vector aleatorio $\mathbf{V}_{k,R}|\mathbf{V}_{R-1}$ sigue una distribución normal multivariada centrada, con matriz de covarianza:

$$\tilde{\Sigma}^{(R)}(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[\mathbf{S}\mathbf{S}' \middle| \mathbf{V}_{R-1}\right] = \frac{1}{n_{R-1}} \left(\alpha^{(R-1)}(\mathbf{x})\right)^T \alpha^{(R-1)}(\mathbf{x}') + \beta^2,$$

con

$$\mathbf{S} = \left(\frac{1}{\sqrt{n_{R-1}}} \sum_{i=1}^{n_{R-1}} \mathbf{W}_{k,i}^{(R-1)} \alpha_i^{(R-1)}(\mathbf{x}) + \beta \mathbf{b}_k^{(R-1)}\right)$$

$$\mathbf{S}' = \left(\frac{1}{\sqrt{n_{R-1}}} \sum_{j=1}^{n_{R-1}} \mathbf{W}_{k,j}^{(R-1)} \alpha_j^{(R-1)}(\mathbf{x}') + \beta \mathbf{b}_k^{(R-1)}\right).$$

Dado que por hipótesis las post-activaciones de la capa $R-1$ son independientes e idénticamente distribuidas, entonces por la ley débil de los grandes números:

$$\frac{1}{n_{R-1}} \alpha^{R-1}(\mathbf{x})^T \alpha^{R-1}(\mathbf{x}') \xrightarrow{P} \mathbb{E}\left[\left(\alpha_k^{R-1}(\mathbf{x})\right)^T \alpha_k^{(R-1)}(\mathbf{x}'_k)\right], \qquad n_{R-1} \to \infty,$$

de modo que:

$$\tilde{\Sigma}^{(R)}(\mathbf{x}, \mathbf{x}') \xrightarrow{P} \mathbb{E}\left[\alpha_k^{R-1}(\mathbf{x})^T \alpha_k^{R-1}(\mathbf{x}'_k)\right] + \beta^2, \qquad n_{R-1} \to \infty.$$

La matriz de covarianzas del vector condicionado converge casi seguramente a una matriz determinista que no depende de la realizaciones de $\alpha^R(\mathbf{x})$ como variable aleatoria, y por lo tanto el límite corresponde a $\Sigma^{(R)}(\mathbf{x}, \mathbf{x}')$, es decir,

$$\Sigma^{(R)}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{\iota_{R-1}}\left[\alpha_k^{(R-1)}(\mathbf{x})\, \alpha^{(R-1)}(\mathbf{x}')\right] + \beta^2.$$

Para demostrar la independencia entre salidas distintas consideraremos $\tilde{\alpha}_k^{(R)}(\mathbf{x})$, $\tilde{\alpha}_{k'}^{(R)}(\mathbf{x}')$, con $k \neq k'$ y $\mathbf{x}, \mathbf{x}' \in T$. Entonces:

$$\mathbf{V} = \begin{bmatrix} \tilde{\alpha}_k^{(R)}(\mathbf{x}) \\ \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \end{bmatrix} = \frac{1}{\sqrt{n_{R-1}}} \underbrace{\sum_{i=1}^{n_{R-1}} \begin{bmatrix} \mathbf{W}_{k,i}^{(R-1)} \alpha_i^{(R-1)}(\mathbf{x}) \\ \mathbf{W}_{k',i}^{(R-1)} \alpha_i^{(R-1)}(\mathbf{x}') \end{bmatrix}}_{\Gamma_0} + \beta \begin{bmatrix} \mathbf{b}_k^{(R)} \\ \mathbf{b}_{k'}^{(R)} \end{bmatrix}.$$

Los vectores en Γ_0 son independientes e idénticamente distribuidos gracias a la hipótesis de inducción. Entonces, procediendo de la misma manera que en párrafos anteriores, podemos concluir que $\mathbf{V}$ converge en ley a una distribución normal centrada bivariada. Por otra parte,

$$\begin{aligned}
Cov\left(\tilde{\alpha}_k^{(R)}(\mathbf{x}), \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}')\right) &= \mathbb{E}\left[\left(\frac{1}{\sqrt{n_{R-1}}}\sum_{i=1}^{n_{R-1}} \mathbf{W}_{k,i}^{(R-1)} \alpha_i^{(R-1)}(\mathbf{x}) + \beta \mathbf{b}_k^{(R)}\right)\right. \\
&\qquad \left.\left(\frac{1}{\sqrt{n_{R-1}}}\sum_{i=1}^{n_{R-1}} \mathbf{W}_{k',i}^{(R-1)} \alpha_i^{(R-1)}(\mathbf{x}') + \beta \mathbf{b}_{k'}^{(R)}\right)\right] \\
&= \frac{1}{n_0}\sum_{i,j=1}^{n_0} \mathbb{E}\left[\alpha_i^{(R-1)}(\mathbf{x})\,\alpha_j^{(R-1)}(\mathbf{x}')\right] \mathbb{E}\left[\mathbf{W}_{k,i}^{(R-1)}\right] \mathbb{E}\left[\mathbf{W}_{k',j}^{(R-1)}\right] \\
&\quad + \frac{\beta}{\sqrt{n_{R-1}}}\sum_{i=1}^{n_0} \mathbb{E}\left[\alpha_i^{(R-1)}(\mathbf{x})\right] \mathbb{E}\left[\mathbf{W}_{k,i}^{(R-1)}\right] \mathbb{E}\left[\mathbf{b}_{k'}^{(R-1)}\right] \\
&\quad + \frac{\beta}{\sqrt{n_{R-1}}}\sum_{j=1}^{n_0} \mathbb{E}\left[\alpha_j^{(R-1)}(\mathbf{x}')\right] \beta \mathbb{E}\left[\mathbf{W}_{k',j}^{(R-1)}\right] \mathbb{E}\left[\mathbf{b}_k^{(R-1)}\right] \\
&\quad + \beta^2 \mathbb{E}\left[\mathbf{b}_k^{(R-1)}\right] \mathbb{E}\left[\mathbf{b}_{k'}^{(R-1)}\right] \\
&= \frac{1}{n_0}\sum_{i,j=1}^{n_0} \mathbb{E}\left[\alpha_i^{(R-1)}(\mathbf{x})\,\alpha_j^{(R-1)}(\mathbf{x}')\right](0)(0) \\
&\quad + \frac{\beta}{\sqrt{n_{R-1}}}\sum_{i=1}^{n_0} \mathbb{E}\left[\alpha_i^{(R-1)}(\mathbf{x})\right](0)(0) \\
&\quad + \frac{\beta}{\sqrt{n_{R-1}}}\sum_{j=1}^{n_0} \mathbb{E}\left[\alpha_j^{(R-1)}(\mathbf{x}')\right]\beta(0)(0) + \beta^2(0)(0) \\
&= 0
\end{aligned}$$

tal que en el límite la covarianza es cero.

En teoría de la probabilidad, si dos variables aleatorias tienen una distribución conjunta normal bivariada y covarianza cero, entonces las variables aleatorias son independientes.

$$\therefore \tilde{\alpha}_k^{(R)}(\mathbf{x}) \perp \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}'), \qquad \forall \mathbf{x}, \mathbf{x}' \in T, \text{ con } i \neq j.$$

$\therefore$ Los vectores aleatorios $\{\mathbf{V}_{i,R}\}_{i=1}^{n_R}$ son independientes e idénticamente distribuidos, con distribución normal multivariada centrada y matriz de covarianza dada por:

$$\Sigma^{(R)}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_{l_{R-1}}\left[\alpha_k^{(R-1)}(\mathbf{x})\,\alpha^{(R-1)}(\mathbf{x}')\right] + \beta^2.$$

□

De la proposición anterior se sigue inmediatamente que las salidas de la red convergen en distribución a un proceso gausiano, pues la marginalización de un vector aleatorio con dis-

tribución normal multivariada se sigue distribuyendo normal, y dos variables aleatorias independientes con distribución normal tienen distribución conjunta normal multivariada.

Un resultado inmediato de la proposición anterior es que tomando límites secuenciales en las capas ocultas de la red, la inicialización del kernel Θ es determinista. Seguiremos usando la notación definida en la proposición anterior. Además, con $\Theta^{(l)}$ nos referiremos al NTK asociado a una red con profundidad l.

Proposición 3.3.3.
Bajo las condiciones de la proposición anterior, la inicialización de kernel $\Theta^{(L)}$ converge en probabilidad a un kernel determinista, dado por $\Theta_\infty^{(L)} \otimes Id_{n_L}$, donde $\Theta_\infty^{(L)}(\cdot,\cdot) : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$, es una función escalar que se define mediante la siguiente recursión (con límites secuenciales):

$$\Theta_\infty^{(1)}(\boldsymbol{x}, \boldsymbol{x}') = \Sigma^{(1)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\Theta_\infty^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \Theta_\infty^{(l)}(\boldsymbol{x}, \boldsymbol{x}')\, \dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') + \Sigma^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}')$$

$$\dot{\Sigma}^{(l+1)}(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\left(\tilde{\alpha}^l(\boldsymbol{x}), \tilde{\alpha}^l(\boldsymbol{x}')\right) \sim \mathcal{N}\left(0, \lambda^{(l)}\right)} \left[\dot{\sigma}\left(\tilde{\alpha}^l(\boldsymbol{x})\right) \dot{\sigma}\left(\tilde{\alpha}^l(\boldsymbol{x}')\right)\right] + \beta^2$$

Demostración. La prueba la haremos por inducción sobre la profundidad

1. **Caso base:** $L = 1$.

$$\begin{aligned}
\Theta_{k,k'}^{(1)}(\mathbf{x}, \mathbf{x}') &= \left(\mathbf{J}f(\theta, \mathbf{x})\, \mathbf{J}f(\theta, \mathbf{x}')^T\right)_{k,k'} \\
&= \nabla_\theta \tilde{\alpha}_k^{(1)}(\mathbf{x})^T \nabla_\theta \tilde{\alpha}_{k'}^{(1)}(\mathbf{x}') \\
&= \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \frac{\partial \tilde{\alpha}_k^{(1)}}{\partial \mathbf{W}_{i,j}^{(0)}}(\mathbf{x}) \frac{\partial \tilde{\alpha}_{k'}^{(1)}}{\partial \mathbf{W}_{i,j}^{(0)}}(\mathbf{x}') + \sum_{i=1}^{n_1} \frac{\partial \tilde{\alpha}_k^{(1)}}{\partial \mathbf{b}_i^{(0)}}(\mathbf{x}) \frac{\partial \tilde{\alpha}_{k'}^{(1)}}{\partial \mathbf{b}_i^{(0)}}(\mathbf{x}') \\
&= \sum_{i=1}^{n_1} \sum_{j=1}^{n_0} \frac{\mathbf{x}_j}{\sqrt{n_0}} \delta_{i,k} \frac{\mathbf{x}'_j}{\sqrt{n_0}} \delta_{i,k'} + \sum_{i=1}^{n_1} \beta \delta_{i,k} \beta \delta_{i,k'} \\
&= \frac{1}{n_0} \sum_{j=1}^{n_0} \mathbf{x}_j \mathbf{x}'_j \delta_{k,k'} + \beta^2 \delta_{k,k'} \\
&= \left(\frac{1}{n_0} \mathbf{x}^T \mathbf{x}' + \beta^2\right) \delta_{k,k'} \\
&= \Sigma^{(1)}(\mathbf{x}, \mathbf{x}')\, \delta_{k,k'}
\end{aligned}$$

   Aquí no hay un proceso límite, de modo que

$$\Theta^{(1)}(\mathbf{x}, \mathbf{x}') = \Sigma^{(1)}(\mathbf{x}, \mathbf{x}') \otimes Id_{n_1} = \Theta_\infty^{(1)}(\mathbf{x}, \mathbf{x}') \otimes Id_{n_1}.$$

2. **Hipótesis inductiva:** $L = R - 1$
   Supongamos que tomando límites secuencialmente, hasta la capa $R-1$, el kernel $\Theta^{(R-1)}$

converge en probabilidad a $\Theta_\infty^{(R-1)} \otimes Id_{n_{R-1}}$. Retomando notación de [16], con $\tilde{\theta}$ denotaremos a los pesos que corresponden hasta la capa $R-1$ de una red neuronal, entonces la hipótesis inductiva se reescribe como:

$$\left(\nabla_{\tilde{\theta}} \tilde{\alpha}_k^{R-1}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}} \tilde{\alpha}_{k'}^{R-1}(\mathbf{x}') \xrightarrow{P} \Theta_\infty^{R-1}(\mathbf{x}, \mathbf{x}') \delta_{k,k'}$$

Seguiremos con esta notación en lo que resta de la prueba.

3. **Paso inductivo:** $L = R$
Notemos que:

$$\begin{aligned}
\left(\nabla_\theta \tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_\theta \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') &= \left(\nabla_{\tilde{\theta}} \tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}} \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \\
&\quad + \left(\nabla_{\mathbf{W}^{(R-1)}} \tilde{\alpha}_k^{R}(\mathbf{x})\right)^T \nabla_{\mathbf{W}^{(R-1)}} \tilde{\alpha}_{k'}^{R}(\mathbf{x}') \\
&\quad + \left(\nabla_{\mathbf{b}^{(R-1)}} \tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\mathbf{b}^{(R-1)}} \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \\
&= \left(\nabla_{\tilde{\theta}} \tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}} \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \\
&\quad + \frac{1}{n_{R-1}} \sum_{j=1}^{n_{R-1}} \sum_{i=1}^{n_R} \alpha_j^{(R-1)}(\mathbf{x}) \alpha_j^{(R-1)}(\mathbf{x}') \delta_{k,i} \delta_{k',i} + \beta^2 \delta_{k,k'} \\
&= \left(\nabla_{\tilde{\theta}} \tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}} \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \\
&\quad + \left(\frac{1}{n_{R-1}} \alpha^{(R-1)}(\mathbf{x})^T \alpha^{(R-1)}(\mathbf{x}') + \beta^2\right) \delta_{k,k'}.
\end{aligned}$$

Enfoquémonos en el primero de los sumandos. Notemos que:

$$\begin{aligned}
\nabla_{\tilde{\theta}} \tilde{\alpha}_k^R(\mathbf{x}) &= \frac{1}{\sqrt{n_{R-1}}} \sum_{i=1}^{n_{R-1}} \mathbf{W}_{k,i}^{(R-1)} \nabla_{\tilde{\theta}} \alpha_i^{(R-1)}(\mathbf{x}) \\
&= \frac{1}{\sqrt{n_{R-1}}} \sum_{i=1}^{n_{R-1}} \mathbf{W}_{k,i}^{(R-1)} \dot{\sigma}\left(\tilde{\alpha}_i^{(R-1)}(\mathbf{x})\right) \nabla_{\tilde{\theta}} \tilde{\alpha}_i^{(R-1)}(\mathbf{x}).
\end{aligned}$$

Para simplificar la escritura definimos la siguiente notación:

$$\boxed{\Xi_{ij} := \dot{\sigma}\left(\tilde{\alpha}_i^{(R-1)}(\mathbf{x})\right) \dot{\sigma}\left(\tilde{\alpha}_j^{(R-1)}(\mathbf{x}')\right)},$$

tal que:

$$\begin{aligned}
\left(\nabla_{\tilde{\theta}} \tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}} \tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') &= \frac{1}{n_{R-1}} \sum_{i,j=1}^{n_{R-1}} \left[\left(\nabla_{\tilde{\theta}} \tilde{\alpha}_i^{(R-1)}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}} \tilde{\alpha}_j^{(R-1)}(\mathbf{x}')\right] \\
&\qquad \left(\Xi_{ij} \mathbf{W}_{k,i}^{(R-1)} \mathbf{W}_{k',j}^{(R-1)}\right) \\
&= \frac{1}{n_{R-1}} \sum_{i,j=1}^{n_{R-1}} \Theta_{i,j}^{(R-1)}(\mathbf{x}, \mathbf{x}') \Xi_{ij} \mathbf{W}_{k,i}^{(R-1)} \mathbf{W}_{k',j}^{(R-1)}.
\end{aligned}$$

Así, por la hipótesis de inducción:

$$\left(\nabla_{\tilde{\theta}}\tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\tilde{\theta}}\tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \xrightarrow{P} \frac{1}{n_{R-1}}\sum_{i,j=1}^{n_{R-1}} \Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\delta_{i,j}\Xi_{i,j}\mathbf{W}_{k,i}^{(R-1)}\mathbf{W}_{k',j}^{(R-1)}$$
$$= \frac{1}{n_{R-1}}\sum_{i=1}^{n_{R-1}} \underbrace{\Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\Xi_{ii}\mathbf{W}_{k,i}^{(R-1)}\mathbf{W}_{k',i}^{(R-1)}}_{\mathbf{X}_i}$$

cuando secuencialmente se toman límites hasta la capa $R-2$.

Como consecuencia de la proposición anterior, las variables aleatorias $\mathbf{X}_i$ son independientes e idénticamente distribuidas, con media finita, entonces por la ley débil de los grandes números:

$$\frac{\sum_{i=1}^{n_{R-1}} \Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\Xi_{ii}\mathbf{W}_{k,i}^{(R-1)}\mathbf{W}_{k',i}^{(R-1)}}{n_{R-1}} \xrightarrow{P} \mathbb{E}\left[\Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\Xi_{ii}\mathbf{W}_{k,i}^{(R-1)}\mathbf{W}_{k',i}^{(R-1)}\right]$$
$$= \Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\mathbb{E}_{\iota_{R-1}}[\Xi_{ii}]\,\mathbb{E}\left[\mathbf{W}_{k,i}^{(R-1)}\mathbf{W}_{k',i}^{(R-1)}\right]$$
$$= \Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\dot{\Sigma}^{(R)}(\mathbf{x},\mathbf{x}')\,\delta_{k,k'},$$

tomando el límite $n_{R-1}\to\infty$.

Por otro lado, para el segundo sumando sabemos que:

$$\left(\frac{1}{n_{R-1}}\alpha^{(R-1)}(\mathbf{x})^T\alpha^{(R-1)}(\mathbf{x}')+\beta^2\right)\delta_{k,k'} \xrightarrow{P} \Sigma^R(\mathbf{x},\mathbf{x}')\,\delta_{k,k'}, \quad n_{R-1}\to\infty.$$

Entonces:

$$\left(\nabla_{\theta}\tilde{\alpha}_k^{(R)}(\mathbf{x})\right)^T \nabla_{\theta}\tilde{\alpha}_{k'}^{(R)}(\mathbf{x}') \xrightarrow{P} \Theta_\infty^{(R-1)}(\mathbf{x},\mathbf{x}')\,\dot{\Sigma}^{(R)}(\mathbf{x},\mathbf{x}')\,\delta_{k,k'} + \Sigma^R(\mathbf{x},\mathbf{x}')\,\delta_{k,k'}$$
$$= \Theta_\infty^R(\mathbf{x},\mathbf{x}')\,\delta_{k,k'}$$

□

Discusión

Los resultados y la teoría antes presentada dieron pie a un nuevo enfoque en el estudio de las redes neuronales artificiales, más en específico han motivado el estudio de la interpretabilidad de estas, pues si bien las redes neuronales tienen gran influencia en el desarrollo actual del mundo, lo que se conoce de ellas muchas veces se limita a meras descripciones de su desempeño bajo condiciones ideales, con un fundamento matemático que muchas veces se supone trivial.

Entre los trabajos que han recibido influencia del kernel neuronal tangente podemos destacar los siguientes:

1. **Modelos sobreparametrizados y modelos lineales:** Sea ha desarrollado teoría en torno a lo que podemos denominar *régimen kernel*, que define las condiciones bajo las cuales el desempeño de una ANN se describe (o aproxima) con la teoría discutida en este capítulo. En un reciente artículo [3] se dan condiciones para que con un ancho suficientemente grande, pero finito, de las capas ocultas, la dinámica del entrenamiento se encuentre dentro del régimen kernel.
   Esto va de la mano con el estudio de la dinámica de modelos sobreajustados[1][27][6], cuya generalización ha resultado ser buena.

   Por otro lado, los modelos de ANN anchos en sus capas ocultas pueden aproximarse linealmente por expansiones de Taylor, y su dinámica se explica por el NTK [19][7].

2. **Proximidad entre ANN y métodos kernel:** Citando textualmente a P. Domingos[9]: *"Our result builds on the concept of neural tangent kernel"* .
   P. Domingos probó que una ANN entrenada por descenso de gradiente (en su versión más sencilla), aproximando al flujo del gradiente, equivale a una máquina kernel, cuyo kernel se construye sobre el NTK. Este resultado muestra una clara relación entre métodos que en el pasado fueron arduamente estudiados (métodos kernel) y las ANNs.

Bibliografía

[1] Allen-Zhu, Z., Li, Y., and Liang, Y. (2020). Learning and generalization in overparameterized neural networks, going beyond two layers.

[2] Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404.

[3] Arora, S., S. Du, S., Hu, W., Li, Z., Salakhutdinov, R., and Wang, R. (2019). On exact computation with an infinitely wide neural net.

[4] Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media.

[5] Brezis, H. (2010). *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer.

[6] Bu, Z., Xu, S., and Chen, K. (2021). A dynamical view on optimization algorithms of overparameterized neural networks. In Banerjee, A. and Fukumizu, K., editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3187–3195. PMLR.

[7] Chizat, L., Oyallon, E., and Bach, F. (2019). On lazy training in differentiable programming. *Advances in neural information processing systems*, 32.

[8] Clapp, M. (2015). *Análisis matemático*. papirhos, IM-UNAM, México.

[9] Domingos, P. (2020). Every model learned by gradient descent is approximately a kernel machine.

[10] Edwards, C. H. (1973). *Advanced Calculus of Several Variables*. Dover Publications.

[11] Elkabetz, O. and Cohen, N. (2021). Continuous vs. discrete optimization of deep neural networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*.

[12] G. de G. Matthews, A., Hron, J., Rowland, M., E. Turner, R., and Ghahramani, Z. (2018). Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*.

[13] Giaquinta, M. and Hildebrandt, S. (1996). *Calculus of Variations I*, volume 310 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag Berlin Heidelberg.

[14] Grabinsky, G. (2013). *Teoría de la medida*. Facultad de Ciencias, UNAM. 3er reimpresión.

[15] H. Manton, J. and Amblard, P.-O. (2015). A primer on reproducing kernel hilbert spaces.

[16] Jacot, A., Gabriel, F., and Hongler, C. (2020). Neural tangent kernel: Convergence and generalization in neural networks.

[17] Kallenberg, O. (2002). *Foundations of Modern Probability*. Springer, 2 edition.

[18] Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. (2017). Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*.

[19] Lee, J., Xiao, L., S Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2020). Wide neural networks of any depth evolve as linear models under gradient descent*. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124002.

[20] Neal, R. M. (1996). *Priors for Infinite Networks*, pages 29–53. Springer New York, New York, NY.

[21] Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge University Press.

[22] Ross, S. M. (1998). *A First Course in Probability*. Prentice Hall, Upper Saddle River, N.J., fifth edition.

[23] Schölkopf, B. (1997). *Support Vector Learning*. PhD thesis, Royal Holloway, University of London.

[24] Schölkopf, B. and Smola, A. J. (2002). *Learning with kernels : support vector machines, regularization, optimization, and beyond*. Adaptive computation and machine learning. MIT Press.

[25] Shalev-Shwartz, S. and Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

[26] Tarmoun, S., Franca, G., Haeffele, B. D., and Vidal, R. (2021). Understanding the dynamics of gradient flow in overparameterized linear models. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR.

[27] Zhang, R. and Zhang, S. (2021). Rethinking influence functions of neural networks in the over-parameterized regime.