Teoría en Optimización de Redes Neuronales Anchas

Víctor Suesta Arribas

Índice

1. Introducción

El aprendizaje profundo ha generado un notable impacto en la inteligencia artificial contemporánea, gracias a la eficacia práctica de las redes neuronales profundas para resolver tareas complejas. Sin embargo, su éxito plantea desafíos teóricos fundamentales: ¿por qué algoritmos locales como el descenso de gradiente logran minimizar funciones altamente no convexas? ¿Cómo pueden modelos con una capacidad de interpolación extrema —muchas veces con más parámetros que muestras— alcanzar generalización sin sobreajuste? Estas cuestiones han motivado una reevaluación del marco matemático tradicional para el análisis de modelos predictivos.

Una de las líneas más prometedoras en esta dirección ha sido el estudio del régimen de *redes anchas*, en el que el número de neuronas por capa crece significativamente. En este límite, se ha observado que la dinámica del entrenamiento puede aproximarse por una evolución lineal en un espacio funcional, y que la estructura del paisaje de pérdida se vuelve particularmente benigna. Herramientas como el *Kernel Tangente Neuronal* (NTK) y el análisis espectral del Hessiano han permitido caracterizar de forma rigurosa fenómenos como la convergencia funcional, la regularización implícita o la conectividad entre mínimos globales.

El presente trabajo desarrolla una caracterización matemática detallada del entrenamiento de redes neuronales anchas mediante descenso de gradiente, con el objetivo de entender por qué estos modelos optimizan con éxito y generalizan en condiciones de sobreparametrización. El enfoque adoptado es puramente teórico: todos los resultados provienen de una selección de diez artículos recientes fundamentales en este campo, y cada afirmación relevante se desarrolla con su demostración correspondiente, evitando cualquier dependencia externa no justificada.

A lo largo del texto se formaliza primero el modelo funcional de red ancha, incluyendo su comportamiento en el límite de ancho infinito y la linealización inducida por el NTK. A continuación, se analiza la dinámica de entrenamiento, tanto desde una perspectiva funcional como paramétrica, y se establecen resultados de convergencia y estabilidad. Posteriormente, se estudia la geometría del paisaje de pérdida, mostrando la ausencia de mínimos locales pobres, la estructura espectral degenerada del Hessiano, y la existencia de trayectorias suaves que conectan soluciones globales. Finalmente, se interpreta el fenómeno del doble descenso a partir de las herramientas desarrolladas, cerrando así un marco coherente que conecta optimización, estructura funcional y generalización en redes neuronales anchas.

2. Modelització funcional de redes anchas

2.1. Definición formal de una red neuronal profunda

Una red neuronal profunda es una función paramétrica compuesta que transforma una entrada de dimensión fija en una salida escalar (o vectorial) mediante una secuencia de capas intermedias no lineales. Sean $d \in \mathbb{N}$ la dimensión de entrada y $L \in \mathbb{N}$ el número de capas ocultas, definimos una red completamente conectada de L capas como una familia de funciones:

$$f_\theta : \mathbb{R}^d \to \mathbb{R}, \quad x \mapsto f_\theta(x),$$

parametrizadas por un conjunto de pesos $\theta = \{W^{(1)}, \ldots, W^{(L)}, a\}$, donde:

- $W^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$ representa la matriz de pesos de la capa ℓ,
- $a \in \mathbb{R}^{n_L}$ es el vector de pesos de la capa de salida (lineal),
- $n_0 := d$, es decir, la dimensión de entrada,
- n_ℓ es el número de neuronas en la capa ℓ.

Se asume $n_L = m$, donde $m = 1$ si la salida es escalar, o $m > 1$ si es vectorial.
La transformación a través de la red se define inductivamente como:

$$\begin{cases} x^{(0)} := x, \\ x^{(\ell)} := \phi\left(W^{(\ell)} x^{(\ell-1)}\right), \quad \text{para } \ell = 1, \ldots, L, \end{cases}$$

donde $\phi : \mathbb{R} \to \mathbb{R}$ es una función de activación aplicada componente a componente (por ejemplo, ReLU, tanh, erf). La salida final viene dada por:

$$f_\theta(x) := a^\top x^{(L)}.$$

Así, f_θ es una función paramétrica dependiente de todos los pesos θ. Esta arquitectura corresponde a una red totalmente conectada sin términos de sesgo, aunque el análisis puede extenderse fácilmente al caso con sesgos sin pérdida de generalidad [1, 2].
La dimensión total del espacio de parámetros es:

$$p := \sum_{\ell=1}^{L} n_\ell n_{\ell-1} + n_L.$$

A efectos analíticos, nos interesa no solo la función paramétrica f_θ, sino también el conjunto funcional que estas redes pueden generar. Para ello, introducimos el espacio:

$$F_\Theta := \{f_\theta : \theta \in \mathbb{R}^p\},$$

que es un subconjunto altamente no lineal del espacio de funciones $F = \{f : \mathbb{R}^d \to \mathbb{R}\}$.
Esta formulación tiene dos consecuencias fundamentales:

- La función f_θ depende de forma no lineal de los parámetros θ, debido a la composición repetida de capas con activación no lineal.
- La estructura de F_Θ varía radicalmente en función de la topología del modelo (profundidad L, activación ϕ, número de neuronas n_ℓ).

Una propiedad importante para el análisis posterior es la diferenciabilidad de f_θ respecto a θ. Bajo condiciones regulares sobre ϕ (por ejemplo, si ϕ es Lipschitz y diferenciable a trozos, como ReLU), la función $\theta \mapsto f_\theta(x)$ es diferenciable casi en todas partes. Este hecho es esencial para el estudio del descenso de gradiente y para el análisis espectral del Hessiano, como se desarrollará en los capítulos 3 y 4.

Finalmente, cabe destacar que el enfoque que adoptaremos a partir de la siguiente sección consiste en estudiar el límite en el que el número de neuronas por capa tiende a infinito, manteniendo fija la arquitectura de profundidad. En este régimen, las funciones f_θ exhiben un comportamiento sorprendentemente lineal en θ, lo cual permite una aproximación funcional que convierte el modelo en un objeto mucho más accesible desde el punto de vista teórico.

2.2. Límite de ancho infinito y linealización funcional

El análisis formal de redes neuronales profundas en el régimen de ancho infinito parte de considerar el comportamiento de dichas redes cuando el número de neuronas por capa tiende a infinito. Este límite, aunque idealizado, permite simplificar de forma notable la dinámica de entrenamiento y obtener una caracterización funcional de la red mediante objetos lineales y deterministas.

Sea $f_\theta : \mathbb{R}^d \to \mathbb{R}$ una red neuronal completamente conectada de L capas ocultas, donde cada capa tiene n neuronas, y sea $\theta \in \mathbb{R}^p$ el vector de parámetros. Supondremos que los pesos θ están inicializados aleatoriamente, con cada componente extraído de una ley normal independiente, adecuadamente normalizada (escalado por $1/\sqrt{n}$).

El proceso de entrenamiento se modela como una evolución continua en el tiempo bajo descenso de gradiente, es decir, se considera la dinámica:

$$\frac{d\theta(t)}{dt} = -\nabla_\theta \mathcal{L}(\theta(t)),$$

donde $\mathcal{L}$ es la pérdida empírica sobre un conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^n$. La pregunta que guía este apartado es: ¿cómo evoluciona la salida de la red $f_{\theta(t)}(x)$ a medida que se actualizan los parámetros?

Teorema (Linealización funcional en el régimen de ancho infinito). Sea f_θ una red neuronal profunda de ancho uniforme n en cada capa oculta, inicializada con pesos aleatorios de varianza escalada $1/n$. Suponiendo activaciones Lipschitz, datos acotados y pérdida diferenciable, entonces en el límite $n \to \infty$ se tiene que la dinámica funcional de $f_{\theta(t)}(x)$ verifica:

$$f_{\theta(t)}(x) \approx f_{\theta(0)}(x) - \int_0^t \sum_{i=1}^n \Theta(x, x_i) \frac{\partial \mathcal{L}}{\partial f}(f_{\theta(s)}(x_i))\, ds,$$

donde $\Theta(x, x') := \langle \nabla_\theta f_\theta(x), \nabla_\theta f_\theta(x') \rangle$ es el **Kernel Tangente Neuronal (NTK)** evaluado en la inicialización. En este régimen, Θ se vuelve determinista y constante en el tiempo.

Demostración. Comenzamos con el desarrollo de f_θ alrededor de la inicialización $\theta_0 := \theta(0)$ mediante una expansión de primer orden:

$$f_{\theta(t)}(x) = f_{\theta_0}(x) + \langle \nabla_\theta f_{\theta_0}(x), \theta(t) - \theta_0 \rangle + R(t),$$

donde $R(t)$ representa el resto de segundo orden de la expansión de Taylor. En el límite $n \to \infty$, el término de segundo orden $R(t)$ tiende a cero con alta probabilidad, debido a la concentración de medida en alta dimensión y a la estabilidad de las derivadas.

Puesto que la evolución de los parámetros está dada por:

$$\theta(t) - \theta_0 = -\int_0^t \nabla_\theta \mathcal{L}(\theta(s))\, ds,$$

se puede sustituir esta expresión en la expansión anterior:

$$f_{\theta(t)}(x) \approx f_{\theta_0}(x) - \int_0^t \langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta \mathcal{L}(\theta(s)) \rangle\, ds.$$

Por la regla de la cadena, se tiene:

$$\nabla_\theta \mathcal{L}(\theta) = \sum_{i=1}^{n} \frac{\partial \mathcal{L}}{\partial f}(f_\theta(x_i)) \nabla_\theta f_\theta(x_i).$$

Sustituyendo:

$$f_{\theta(t)}(x) \approx f_{\theta_0}(x) - \int_0^t \sum_{i=1}^{n} \langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta f_\theta(x_i) \rangle \cdot \frac{\partial \mathcal{L}}{\partial f}(f_\theta(x_i))\, ds.$$

Ahora, en el régimen $n \to \infty$, se cumple que:

$$\langle \nabla_\theta f_{\theta_0}(x), \nabla_\theta f_\theta(x_i) \rangle \longrightarrow \Theta(x, x_i),$$

donde Θ es el NTK evaluado en la inicialización. Además, se ha demostrado que este kernel se mantiene constante a lo largo del entrenamiento, es decir, $\Theta_{\theta(t)} \approx \Theta_{\theta_0}$ durante toda la dinámica.

Por tanto, se concluye que:

$$f_{\theta(t)}(x) \approx f_{\theta_0}(x) - \int_0^t \sum_{i=1}^{n} \Theta(x, x_i) \cdot \frac{\partial \mathcal{L}}{\partial f}(f_{\theta(s)}(x_i))\, ds,$$

lo que demuestra que la red neuronal se comporta funcionalmente como un sistema lineal evolutivo gobernado por el NTK.

Este resultado marca un cambio de paradigma en el estudio matemático de redes neuronales profundas. Mientras que en el régimen clásico la optimización se enfrenta a una función de pérdida altamente no convexa, en el régimen de ancho infinito el comportamiento de la red se aproxima por una evolución funcional lineal, lo cual permite desarrollar una teoría rigurosa de su dinámica de entrenamiento.

En el siguiente apartado, formalizaremos el concepto de **Kernel Tangente Neuronal (NTK)** y analizaremos sus propiedades espectrales, estructura funcional y papel central en la optimización de redes anchas.

2.3. Convergencia funcional mediante el NTK

En el régimen de ancho infinito, el comportamiento de una red neuronal profunda durante el entrenamiento queda gobernado por una dinámica funcional lineal, cuya evolución puede estudiarse rigurosamente mediante el **Kernel Tangente Neuronal (NTK)**. En esta sección formalizamos dicha convergencia y demostramos que, bajo condiciones razonables, el descenso

de gradiente aplicado a una red ancha conduce a una solución única y estable en el espacio funcional generado por el NTK.

Supongamos que $f_\theta : \mathbb{R}^d \to \mathbb{R}$ es la función inducida por una red neuronal profunda totalmente conectada, con parámetros inicializados aleatoriamente, y que se entrena sobre un conjunto de datos finito $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ minimizando la pérdida empírica:

$$\mathcal{L}(\theta) = \frac{1}{2} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i\right)^2.$$

El vector de salidas sobre los datos se denota por $f_\theta(X) := (f_\theta(x_1), \ldots, f_\theta(x_n))^\top \in \mathbb{R}^n$, y el vector de etiquetas como $y := (y_1, \ldots, y_n)^\top$.

Teorema (Convergencia funcional en el régimen de ancho infinito). Sea f_θ una red neuronal profunda de ancho n entrenada con descenso de gradiente continuo sobre la pérdida cuadrática, con pesos inicializados según una distribución gaussiana independiente y normalizada por $1/\sqrt{n}$. Sea $\Theta(x, x')$ el kernel tangente neuronal asociado a la inicialización $\theta(0)$. Entonces, en el límite $n \to \infty$, se cumple que:

$$f_{\theta(t)}(x) \to f^\infty(x) := \Theta(x, X)\Theta(X, X)^{-1} y,$$

para toda entrada $x \in \mathbb{R}^d$, donde $\Theta(x, X) := (\Theta(x, x_1), \ldots, \Theta(x, x_n))$. Además, se tiene convergencia funcional en norma L^2 sobre compactos, y la dinámica sigue:

$$f_{\theta(t)}(x) = f_{\theta(0)}(x) - \Theta(x, X)\left(I - e^{-t\Theta(X,X)}\right)\left(f_{\theta(0)}(X) - y\right).$$

Demostración. Partimos de la linealización obtenida en el apartado anterior:

$$f_{\theta(t)}(x) \approx f_{\theta(0)}(x) - \int_0^t \sum_{i=1}^{n} \Theta(x, x_i) \frac{\partial \mathcal{L}}{\partial f}(f_{\theta(s)}(x_i)) ds.$$

Como estamos en el caso de pérdida cuadrática:

$$\frac{\partial \mathcal{L}}{\partial f}(f_\theta(x_i)) = f_\theta(x_i) - y_i.$$

Por tanto, la dinámica funcional se reescribe como:

$$\frac{d}{dt} f_{\theta(t)}(x) = -\sum_{i=1}^{n} \Theta(x, x_i)(f_{\theta(t)}(x_i) - y_i).$$

En forma matricial, para todo $x \in \mathbb{R}^d$:

$$\frac{d}{dt} f_{\theta(t)}(x) = -\Theta(x, X)(f_{\theta(t)}(X) - y).$$

Esta es una ecuación diferencial lineal inhomogénea en el espacio funcional generado por el NTK. Definimos el vector $\delta(t) := f_{\theta(t)}(X) - y$, de manera que:

$$\frac{d}{dt}\delta(t) = -\Theta(X, X)\delta(t).$$

La solución de esta ecuación lineal matricial es:

$$\delta(t) = e^{-t\Theta(X,X)}\delta(0).$$

De ahí se deduce que:

$$f_{\theta(t)}(X) = y + e^{-t\Theta(X,X)}(f_{\theta(0)}(X) - y),$$

y por tanto:

$$f_{\theta(t)}(x) = f_{\theta(0)}(x) - \Theta(x, X)\left(I - e^{-t\Theta(X,X)}\right)(f_{\theta(0)}(X) - y).$$

Tomando el límite $t \to \infty$, se observa que la matriz exponencial $e^{-t\Theta(X,X)}$ converge a cero. Esto se debe a que $\Theta(X, X)$ es una matriz simétrica semidefinida positiva, lo cual garantiza que sus valores propios son reales y no negativos, y por tanto $e^{-t\lambda_i} \to 0$ para cada valor propio λ_i. Como consecuencia:

$$f_{\theta(t)}(x) \longrightarrow f^{\infty}(x) := \Theta(x, X)\Theta(X, X)^{-1}y.$$

Esta expresión corresponde exactamente a la solución del sistema lineal que minimiza la pérdida cuadrática empírica en el espacio funcional generado por el kernel Θ. Es decir, f^{∞} es la solución de mínima norma funcional dentro de la *Reproducing Kernel Hilbert Space* (RKHS) inducida por Θ, tal y como se demuestra formalmente en los trabajos [1, 3, 2].

Este resultado es clave: establece que en el régimen de ancho infinito, el entrenamiento de la red profunda por descenso de gradiente equivale funcionalmente a resolver un sistema lineal en el espacio RKHS generado por el NTK. La red actúa como un modelo de kernel no entrenado, y la solución obtenida es la de **mínima complejidad funcional**, entre todas las que interpolan los datos.

En los próximos apartados analizaremos cómo esta convergencia funcional explica, desde el punto de vista matemático, los fenómenos de estabilidad, regularización implícita y capacidad de generalización observados en redes neuronales anchas.

3. Optimización via NTK: dinámica i estabilidad

3.1. Convergencia paramétrica del descenso de gradiente

El análisis funcional anterior ha mostrado que, en el régimen de ancho infinito, la salida de la red neuronal sigue una evolución linealmente gobernada por el kernel tangente neuronal (NTK). No obstante, para una caracterización completa de la optimización en redes anchas, es necesario estudiar también la evolución del vector de parámetros $\theta(t)$ durante el entrenamiento. Este apartado se centra en la dinámica paramétrica inducida por el descenso de gradiente y en las condiciones que garantizan su convergencia hacia una solución global.

Hipótesis y marco de trabajo Consideramos una red neuronal de dos capas con función de activación $\sigma : \mathbb{R} \to \mathbb{R}$, definida como

$$f_\theta(x) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(w_r^\top x),$$

donde $\theta = \{(a_r, w_r)\}_{r=1}^m$, con $a_r \in \mathbb{R}$, $w_r \in \mathbb{R}^d$, y m es el número de neuronas de la capa oculta. El factor de normalización $1/\sqrt{m}$ es crucial para el comportamiento asintótico correcto cuando $m \to \infty$. El conjunto de entrenamiento es $\{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$.

Suponemos que:

- Las entradas x_i están acotadas: $\|x_i\| \leq 1$.
- La función σ es Lipschitz y diferenciable.
- Los pesos iniciales $a_r(0), w_r(0)$ son independientes y siguen distribuciones normales centradas con varianza adecuada.
- La pérdida es la cuadrática empírica:

$$\mathcal{L}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} (f_\theta(x_i) - y_i)^2.$$

Teorema (Convergencia paramétrica bajo sobreparametrización) Sea $\theta(t)$ la evolución de los parámetros bajo descenso de gradiente continuo. Entonces, si el ancho m de la red es suficientemente grande (polinomialmente en n), con alta probabilidad sobre la inicialización, se cumple que:

$$\mathcal{L}(\theta(t)) \leq e^{-\lambda t} \mathcal{L}(\theta(0)),$$

para algún $\lambda > 0$ independiente de m. En particular, la red converge a una solución global de la pérdida empírica:

$$\lim_{t \to \infty} \mathcal{L}(\theta(t)) = 0, \quad \lim_{t \to \infty} f_{\theta(t)}(x_i) = y_i, \quad \forall i.$$

Demostración La prueba se basa en la estimación de la derivada temporal de la pérdida empírica. Sea

$$\delta(t) := f_{\theta(t)}(X) - y \in \mathbb{R}^n,$$

donde $f_{\theta(t)}(X) = (f_{\theta(t)}(x_1), \ldots, f_{\theta(t)}(x_n))$.

Consideramos la dinámica del descenso de gradiente continuo:

$$\frac{d\theta(t)}{dt} = -\nabla_\theta \mathcal{L}(\theta(t)),$$

y aplicamos la regla de la cadena:

$$\frac{d}{dt} \mathcal{L}(\theta(t)) = \langle \nabla_\theta \mathcal{L}(\theta(t)), \frac{d\theta(t)}{dt} \rangle = -\|\nabla_\theta \mathcal{L}(\theta(t))\|^2.$$

El gradiente de la pérdida se expresa como:

$$\nabla_\theta \mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \delta_i(t) \nabla_\theta f_\theta(x_i),$$

con lo que la norma cuadrática del gradiente es:

$$\|\nabla_\theta \mathcal{L}\|^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} \delta_i(t) \nabla_\theta f_\theta(x_i) \right\|^2 .$$

Ahora definimos el **Kernel de Tangente Neuronal paramétrico**:

$$\Theta_\theta(x_i, x_j) := \langle \nabla_\theta f_\theta(x_i), \nabla_\theta f_\theta(x_j) \rangle .$$

El vector $\delta(t) \in \mathbb{R}^n$ satisface:

$$\frac{d}{dt}\mathcal{L} = -\frac{1}{n^2} \delta(t)^\top \Theta_\theta(X, X) \delta(t),$$

y si $\Theta_\theta(X, X) \succeq \lambda I$ (es decir, tiene valor propio mínimo acotado por $\lambda > 0$), entonces:

$$\frac{d}{dt}\mathcal{L} \leq -\frac{\lambda}{n^2} \|\delta(t)\|^2 = -\frac{2\lambda}{n} \mathcal{L}(\theta(t)).$$

Esto implica:

$$\frac{d}{dt}\mathcal{L} \leq -\rho \mathcal{L}, \quad \text{con } \rho := \frac{2\lambda}{n},$$

y por tanto:

$$\mathcal{L}(\theta(t)) \leq e^{-\rho t} \mathcal{L}(\theta(0)).$$

Esto muestra que la pérdida decae exponencialmente, y como es positiva, tiende a cero:

$$\lim_{t \to \infty} \mathcal{L}(\theta(t)) = 0.$$

En particular, esto implica:

$$f_{\theta(t)}(x_i) \to y_i \quad \text{para todo } i = 1, \ldots, n,$$

como queríamos demostrar.

Este resultado demuestra que, bajo una condición de ancho suficientemente grande (concretamente, que garantice que el NTK inicial está bien condicionado), el descenso de gradiente es capaz de encontrar una **solución global** de la pérdida empírica, incluso en presencia de no convexidad. Este hecho —sorprendente desde el punto de vista de la optimización clásica— se fundamenta en la geometría especial del paisaje de pérdida en redes anchas [4] , que exploraremos en los apartados siguientes.

3.2. Regularización implícita inducida por el descenso de gradiente

Una de las propiedades más notables de las redes neuronales profundas entrenadas mediante descenso de gradiente es su capacidad para generalizar, incluso en contextos de sobreparametrización extrema. Este comportamiento, aparentemente contradictorio con la intuición clásica basada en el sobreajuste, se explica por un fenómeno denominado **regularización implícita**: sin necesidad de añadir términos explícitos de penalización en la función de pérdida, el algoritmo de optimización selecciona preferentemente soluciones con ciertas propiedades de simplicidad o estructura funcional favorable.

En el caso de las redes anchas entrenadas con descenso de gradiente, se ha demostrado que la solución convergente obtenida no es arbitraria, sino que pertenece a un subespacio funcional de norma mínima en la RKHS (Reproducing Kernel Hilbert Space) inducida por el NTK.

Teorema (Convergencia hacia la solución de mínima norma en la RKHS) Sea f_θ una red neuronal de dos capas, suficientemente ancha y entrenada mediante descenso de gradiente sobre la pérdida logística. Supongamos que los datos son linealmente separables en la RKHS inducida por el NTK Θ. Entonces, con alta probabilidad sobre la inicialización, se cumple que:

$$\lim_{t\to\infty} f_{\theta(t)}(x) = f^\star(x),$$

donde $f^\star \in \mathcal{H}_\Theta$ es la solución de margen máximo, es decir, aquella función f que separa los datos correctamente y minimiza la norma $\|f\|_{\mathcal{H}_\Theta}$.

Demostración Consideramos un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$ con $y_i \in \{-1, 1\}$, linealmente separables en la RKHS asociada al kernel Θ. Entrenamos una red de dos capas f_θ con activación σ, inicializada aleatoriamente y con ancho m suficientemente grande. La función de pérdida logística empírica se define como:

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i f_\theta(x_i)}).$$

La evolución de los parámetros bajo descenso de gradiente continuo está dada por:

$$\frac{d\theta(t)}{dt} = -\nabla_\theta \mathcal{L}(\theta(t)).$$

El objetivo es demostrar que, en el límite $t \to \infty$, la función $f_{\theta(t)}$ converge hacia la **solución de margen máximo funcional** en la RKHS. La clave es analizar la evolución funcional $f_{\theta(t)}$ y mostrar que:

1. La secuencia de funciones $f_{\theta(t)}$ permanece en $\mathcal{H}_\Theta$ para todo t.
2. La pérdida logística decae hasta cero: $\lim_{t\to\infty} \mathcal{L}(\theta(t)) = 0$, por lo que $f_{\theta(t)}(x_i) \to +\infty$ si $y_i = 1$ y $\to -\infty$ si $y_i = -1$.

3. La dirección de $f_{\theta(t)}$, renormalizada, converge hacia la función $f^\star \in \mathcal{H}_\Theta$ que maximiza el margen funcional:

$$\max_{\|f\|_{\mathcal{H}_\Theta} \leq 1} \min_i y_i f(x_i).$$

Este último paso requiere demostrar que, aunque la norma de $f_{\theta(t)}$ diverge, la dirección funcional en $\mathcal{H}_\Theta$ converge hacia aquella de norma mínima que realiza la separación.

En [5] se demuestra formalmente que esta convergencia es consecuencia de:

- La linealización funcional del modelo en el régimen ancho.
- La estructura convexa de la pérdida logística.
- La dinámica de descenso de gradiente, que sigue trayectorias funcionales que minimizan la pérdida sin abandonar el cono generado por las derivadas funcionales en la RKHS.

Por tanto,

$$\frac{f_{\theta(t)}}{\|f_{\theta(t)}\|_{\mathcal{H}_\Theta}} \longrightarrow \frac{f^\star}{\|f^\star\|_{\mathcal{H}_\Theta}},$$

donde $f^\star$ es la única función de margen máximo.

Este resultado es especialmente relevante: incluso en ausencia de regularización explícita, el algoritmo de entrenamiento impone de forma implícita un sesgo inductivo hacia soluciones estructuralmente simples. En redes anchas, esta simplicidad se traduce en **baja complejidad funcional en la RKHS**, lo que ofrece una explicación matemática de su sorprendente capacidad de generalización, aún en contextos de interpolación perfecta.

3.3. Estabilidad funcional y control del error

Tal como hemos visto en apartados anteriores, la evolución funcional de una red neuronal ancha puede aproximarse por un sistema dinámico lineal inducido por el NTK. Esta linealización es válida si el término de segundo orden de la expansión de Taylor de la función f_θ, denotado por $R(t)$, permanece suficientemente pequeño durante el entrenamiento. En esta sección formalizamos y probamos este hecho, demostrando que la **trayectoria funcional real** y su **aproximación linealizada** se mantienen cercanas en todo el intervalo de entrenamiento.

Planteamiento del problema Recordemos la expansión de primer orden de $f_{\theta(t)}(x)$ alrededor de la inicialización:

$$f_{\theta(t)}(x) = f_{\theta(0)}(x) + \langle \nabla_\theta f_{\theta(0)}(x), \theta(t) - \theta(0) \rangle + R(t, x),$$

donde $R(t, x)$ representa los términos de segundo orden (y superiores) de la expansión. Nuestro objetivo es demostrar que:

$$\sup_{x \in \mathcal{D}} |R(t, x)| \longrightarrow 0 \quad \text{cuando } m \to \infty,$$

uniformemente en $t \in [0, T]$, para dominios $\mathcal{D} \subset \mathbb{R}^d$ compactos y acotados.

Teorema (Control del error de linealización en redes anchas) Sea f_θ una red neuronal de dos capas con activación Lipschitz σ, entrenada mediante descenso de gradiente continuo con inicialización aleatoria y normalización $1/\sqrt{m}$. Entonces, para todo $T > 0$, con alta probabilidad sobre la inicialización, se cumple:

$$\sup_{t\in[0,T]} \sup_{x\in\mathcal{D}} |f_{\theta(t)}(x) - f_{\text{lin}}(t,x)| \to 0 \quad \text{cuando } m \to \infty,$$

donde $f_{\text{lin}}(t,x) := f_{\theta(0)}(x) - \int_0^t \Theta(x,X)\nabla\mathcal{L}(f_{\theta(s)}(X))ds$ es la evolución funcional linealizada inducida por el NTK.

Demostración La demostración se divide en tres pasos:

Paso 1: Acotación del gradiente funcional Por propiedades de la arquitectura y la inicialización aleatoria normalizada (véase [5], [4]), se tiene que:

$$\|\nabla_\theta f_\theta(x)\| \leq C,$$

uniformemente en θ y en $x \in \mathcal{D}$, para alguna constante C independiente de m, gracias a la normalización $1/\sqrt{m}$ y a la acotación de σ y x.

Paso 2: Expansión del segundo orden La expresión de Taylor con resto de segundo orden para $f_{\theta(t)}(x)$ alrededor de $\theta(0)$ es:

$$f_{\theta(t)}(x) = f_{\theta(0)}(x) + \langle\nabla_\theta f_{\theta(0)}(x), \theta(t) - \theta(0)\rangle + \frac{1}{2}(\theta(t) - \theta(0))^\top \nabla_\theta^2 f_\xi(x)(\theta(t) - \theta(0)),$$

para algún punto intermedio ξ entre $\theta(0)$ y $\theta(t)$. Denotamos este término cuadrático como:

$$R(t,x) = \frac{1}{2}(\theta(t) - \theta(0))^\top \nabla_\theta^2 f_\xi(x)(\theta(t) - \theta(0)).$$

Paso 3: Control de la segunda derivada y de la trayectoria Según los resultados combinados de [4] y [5], se puede acotar el hessiano $\nabla_\theta^2 f_\xi(x)$ por un valor del orden $\mathcal{O}(1/\sqrt{m})$ con alta probabilidad. Esta estimación se debe a que cada derivada segunda depende de la derivada de σ, que está acotada, y a que la arquitectura incluye un escalado por $1/\sqrt{m}$, lo que reduce la magnitud de los términos cuadráticos.

Además, se puede probar (véase [4], Thm 2.1) que la trayectoria $\theta(t) - \theta(0)$ permanece acotada por $\mathcal{O}(1)$ en norma euclídea durante $t \in [0,T]$.

Combinando ambos hechos:

$$|R(t,x)| \leq \frac{1}{2}\|\theta(t) - \theta(0)\|^2 \cdot \|\nabla_\theta^2 f_\xi(x)\| \leq C \cdot \frac{1}{\sqrt{m}}.$$

Por tanto,

$$\sup_{t\in[0,T]} \sup_{x\in\mathcal{D}} |R(t,x)| \leq \frac{C}{\sqrt{m}} \to 0 \quad \text{cuando } m \to \infty.$$

■

Este resultado valida de manera rigurosa el modelo linealizado que hemos utilizado en los apartados anteriores, y justifica el uso del NTK como una aproximación precisa de la dinámica funcional real, siempre que la anchura sea suficientemente grande. La estabilidad de esta aproximación es fundamental para el análisis teórico de la convergencia y generalización de redes neuronales anchas.

4. Análisis del paisaje de pérdida

4.1. Aproximación del Hessiano en el régimen funcional

En los apartados anteriores se ha demostrado que, en el régimen de ancho infinito, la función implementada por la red neuronal evoluciona según una dinámica funcional gobernada por un sistema lineal derivado del kernel tangente neuronal (NTK). Este comportamiento funcional permite reemplazar el análisis tradicional sobre el espacio de parámetros por un estudio directo sobre el espacio de funciones inducido por el NTK. En este contexto, la aproximación del Hessiano funcional adquiere un papel central, ya que permite describir la curvatura de la pérdida en torno a trayectorias de entrenamiento, identificando regiones planas o abruptas en el paisaje.

Nuestro objetivo en este apartado es demostrar que el Hessiano del funcional de pérdida respecto a los parámetros θ puede ser aproximado, en el régimen ancho, por una versión funcional definida en términos del NTK. Esta aproximación será válida mientras la dinámica funcional linealizada siga siendo una buena aproximación, hecho que fue formalizado en el Teorema de estabilidad del apartado 3.3.

Planteamiento y notación Recordamos que la red f_θ se entrena minimizando la pérdida empírica cuadrática:

$$\mathcal{L}(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i\right)^2,$$

y que su gradiente se puede expresar como:

$$\nabla_\theta \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(f_\theta(x_i) - y_i\right) \nabla_\theta f_\theta(x_i).$$

El Hessiano de la pérdida se define como:

$$H(\theta) := \nabla_\theta^2 \mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_\theta f_\theta(x_i) \nabla_\theta f_\theta(x_i)^\top + \left(f_\theta(x_i) - y_i\right) \nabla_\theta^2 f_\theta(x_i)\right].$$

Este operador tiene dos componentes: un término positivo semi-definido (producto exterior del gradiente) y un término de segundo orden que depende explícitamente del error y del Hessiano segundo de la red.

Teorema (Aproximación funcional del Hessiano) Sea f_θ una red neuronal suficientemente ancha entrenada con pérdida cuadrática, y sea $H(\theta)$ el Hessiano de la pérdida empírica. Supongamos que:

1. La red tiene activaciones Lipschitz y diferenciables a trozos.
2. La arquitectura es de dos capas o más, con normalización $1/\sqrt{m}$.
3. Los datos (x_i, y_i) están acotados y el error de predicción es pequeño en norma ℓ^2.

Entonces, en el régimen $m \to \infty$, con alta probabilidad sobre la inicialización y mientras la dinámica permanezca en el régimen linealizado, el Hessiano $H(\theta)$ satisface:

$$H(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_{\theta(0)}(x_i) \nabla_\theta f_{\theta(0)}(x_i)^\top =: \Theta_{\theta(0)}(X, X),$$

donde $\Theta_{\theta(0)}$ es la matriz del NTK evaluado en la inicialización.

Demostración La demostración se basa en separar el Hessiano en dos componentes:

$$H(\theta) = H^{(1)}(\theta) + H^{(2)}(\theta),$$

donde:

$$H^{(1)}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_\theta(x_i) \nabla_\theta f_\theta(x_i)^\top,$$

$$H^{(2)}(\theta) := \frac{1}{n} \sum_{i=1}^{n} (f_\theta(x_i) - y_i) \nabla_\theta^2 f_\theta(x_i).$$

El primer término representa el componente positivo semidefinido que domina la curvatura en el régimen funcional. El segundo término es un corrector que se anula cuando la red interpola perfectamente los datos. Por tanto, si el error es pequeño (como se ha demostrado en la convergencia funcional del apartado 2.3), este término puede despreciarse.

Además, por los resultados de control del error de linealización (cf. Sección 3.3), se tiene que:

$$\nabla_\theta f_\theta(x_i) \approx \nabla_\theta f_{\theta(0)}(x_i),$$

con error de orden $\mathcal{O}(1/\sqrt{m})$ en norma, uniformemente en i, durante todo el entrenamiento. Por tanto:

$$H^{(1)}(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_{\theta(0)}(x_i) \nabla_\theta f_{\theta(0)}(x_i)^\top.$$

Finalmente, como $H^{(2)}(\theta)$ es pequeño si el error de predicción es bajo, y $H^{(1)}(\theta)$ se aproxima por su evaluación en la inicialización, concluimos:

$$H(\theta) \approx \Theta_{\theta(0)}(X, X),$$

lo cual demuestra la afirmación.

■

Comentario Este resultado es crucial: nos permite sustituir el análisis del Hessiano de la pérdida en redes anchas por el estudio de una matriz fija y determinista, el NTK inicial. Esto implica que la geometría local del paisaje de pérdida puede entenderse sin necesidad de seguir la trayectoria paramétrica completa, y que la curvatura efectiva del modelo está determinada desde el principio del entrenamiento. Esta propiedad será la base para el análisis espectral que llevaremos a cabo en el apartado 4.2, donde estudiaremos cómo la acumulación espectral en cero y el rango efectivo explican la presencia de trayectorias suaves de optimización.

4.2. Condicionamiento del Hessiano y curvatura de la pérdida

Como consecuencia directa de la aproximación funcional del Hessiano demostrada en el apartadoapartado 4.1, resulta natural analizar la estructura espectral que este operador induce en el paisaje de pérdida. En particular, en redes neuronales profundas y anchas, el Hessiano presenta una degeneración marcada que afecta directamente al comportamiento del descenso de gradiente.

Este apartado se centra en caracterizar esa degeneración espectral, evidenciada en el régimen de sobreparametrización, donde la mayor parte del espectro de $\nabla^2_\theta \mathcal{L}(\theta)$ se acumula en torno a cero, mientras que solo unos pocos autovalores presentan contribuciones significativas. Este fenómeno tiene implicaciones relevantes sobre la estabilidad, el rango efectivo y la convergencia del entrenamiento.

Modelo y contexto. Consideramos una red neuronal profunda f_θ, completamente conectada, entrenada sobre una pérdida empírica de la forma:

$$\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i),$$

donde ℓ es diferenciable dos veces y $\theta \in \mathbb{R}^p$ representa el vector de parámetros.
El Hessiano de la pérdida se define como:

$$H(\theta) := \nabla^2_\theta \mathcal{L}(\theta).$$

Queremos estudiar la estructura espectral de $H(\theta)$ en el régimen de ancho infinito, bajo inicialización aleatoria, y durante la evolución por descenso de gradiente.

Descomposición del Hessiano. Derivando dos veces la pérdida, obtenemos:

$$\nabla^2_\theta \mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \left[\nabla^2_\theta f_\theta(x_i) \cdot \frac{\partial \ell}{\partial f}(f_\theta(x_i)) + \nabla_\theta f_\theta(x_i) \nabla_\theta f_\theta(x_i)^\top \cdot \frac{\partial^2 \ell}{\partial f^2}(f_\theta(x_i)) \right].$$

En particular, cuando ℓ es la pérdida cuadrática, esta expresión se reduce a:

$$H(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\nabla_\theta f_\theta(x_i) \nabla_\theta f_\theta(x_i)^\top + \delta_i(t) \cdot \nabla^2_\theta f_\theta(x_i) \right],$$

donde $\delta_i(t) = f_\theta(x_i) - y_i$.

Teorema (Predominio de la componente de primer orden). Sea f_θ una red profunda con activaciones suaves y normalización $1/\sqrt{m}$, entrenada mediante descenso de gradiente continuo. En el régimen $m \to \infty$, con alta probabilidad sobre la inicialización, se cumple:

$$H(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta f_\theta(x_i) \nabla_\theta f_\theta(x_i)^\top + \mathcal{O}\left(\frac{1}{\sqrt{m}}\right),$$

es decir, la componente de segundo orden es asintóticamente despreciable frente al término de primer orden.

Demostración. Bajo inicialización aleatoria con pesos de varianza $1/m$, las derivadas segundas de f_θ respecto a θ tienen magnitud esperada de orden $\mathcal{O}(1/\sqrt{m})$, como se mostró en el apartado 3.3.

Asimismo, durante el entrenamiento, la trayectoria $\theta(t)$ se mantiene en una vecindad acotada de la inicialización, lo que implica que el tamaño de $\delta_i(t)$ está también acotado. Por tanto, el producto $\delta_i(t) \cdot \nabla_\theta^2 f_\theta(x_i)$ es de orden $\mathcal{O}(1/\sqrt{m})$.

La suma sobre i divide por n, por lo que la contribución total de los términos de segundo orden sigue siendo $\mathcal{O}(1/\sqrt{m})$. Esto justifica la aproximación dada.

Implicaciones espectrales. Esta estructura implica que el Hessiano es aproximadamente igual a una matriz de Gram del tipo:

$$H \approx \frac{1}{n} J J^\top, \quad \text{donde } J := [\nabla_\theta f_\theta(x_1), \ldots, \nabla_\theta f_\theta(x_n)] \in \mathbb{R}^{p \times n}.$$

El rango de H está entonces acotado por n, y si $p \gg n$, como ocurre en el régimen de sobreparametrización, la mayoría de los autovalores de H son exactamente cero. Tal como se demuestra en [6], el espectro de H se concentra en un pequeño número de valores positivos, mientras que la mayor parte de la masa espectral se acumula cerca de cero.

Esto sugiere que:

- El paisaje de pérdida tiene direcciones planas dominantes. - Las trayectorias del descenso de gradiente tienden a moverse en subespacios de baja curvatura. - La optimización se ve facilitada por esta estructura degenerada, especialmente en redes anchas.

Observación. Aunque esta estructura espectral puede hacer que la convergencia sea lenta en ciertas direcciones (debido a pequeños autovalores), también explica por qué el descenso de gradiente converge exitosamente a mínimos globales: no hay pozos estrechos que atrapen al optimizador, y el paisaje está compuesto por amplias regiones planas conectadas.

En la siguiente sección se demostrará formalmente la existencia de trayectorias de mínima pérdida continua que conectan distintos mínimos globales, utilizando esta estructura.

4.3. Trayectorias de baja pérdida y geometría suave

Como consecuencia directa de la aproximación funcional del Hessiano demostrada en el apartado 4.1, y del estudio espectral desarrollado en el apartado 4.2, se deduce que el paisaje de pérdida en redes neuronales anchas posee una estructura geométrica favorable para la

optimización. En particular, se ha demostrado que, bajo condiciones de sobreparametrización suficientes, existen trayectorias continuas en el espacio de parámetros que conectan soluciones globales con pérdida cercana a cero durante todo el recorrido.

Planteamiento funcional. Sea f_θ una red neuronal ancha, y sea $\mathcal{L}(\theta)$ la pérdida empírica sobre un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$. Consideramos el conjunto de parámetros que interpolan perfectamente los datos:

$$\mathcal{M} := \{\theta \in \mathbb{R}^p \mid f_\theta(x_i) = y_i \ \forall i\}.$$

Nuestro objetivo es probar que, si $\theta_0, \theta_1 \in \mathcal{M}$, entonces existen trayectorias $\gamma : [0,1] \to \mathbb{R}^p$ tales que $\gamma(0) = \theta_0$, $\gamma(1) = \theta_1$, y $\mathcal{L}(\gamma(t))$ se mantiene pequeña a lo largo de todo el recorrido.

Teorema (Existencia de trayectorias de pérdida baja entre mínimos globales). Sea f_θ una red neuronal suficientemente ancha con activaciones suaves y pesos inicializados aleatoriamente. Sean $\theta_0, \theta_1 \in \mathcal{M}$ dos soluciones que interpolan perfectamente los datos. Entonces, con alta probabilidad cuando el ancho $m \to \infty$, existe una curva continua $\gamma : [0,1] \to \mathbb{R}^p$ tal que:

$$\gamma(0) = \theta_0, \quad \gamma(1) = \theta_1, \quad \text{y} \quad \mathcal{L}(\gamma(t)) \leq \varepsilon \quad \forall t \in [0,1],$$

donde $\varepsilon \to 0$ cuando $m \to \infty$.

Demostración. La prueba se basa en una construcción funcional, aprovechando el hecho de que en el régimen ancho el modelo se comporta de forma aproximadamente lineal en el espacio funcional generado por el NTK.

(1) Como se mostró en el apartado 3.3, la red ancha se aproxima por un modelo funcional lineal en el espacio RKHS asociado al NTK, es decir:

$$f_{\theta(t)}(x) \approx f_{\theta(0)}(x) - \Theta(x, X)\left(I - e^{-t\Theta(X,X)}\right)(f_{\theta(0)}(X) - y).$$

(2) Este modelo funcional es lineal en $f_{\theta(0)}(X)$, por lo que si dos funciones f_{θ_0} y f_{θ_1} interpolan exactamente los datos, también lo hace su interpolación convexa:

$$f_t := (1-t)f_{\theta_0} + t f_{\theta_1}.$$

Como $f_{\theta_0}(x_i) = f_{\theta_1}(x_i) = y_i$ para todo i, entonces $f_t(x_i) = y_i$ para todo $t \in [0,1]$.

(3) Dado que el modelo funcional es aproximadamente lineal, existe una trayectoria en el espacio de parámetros $\gamma(t)$ tal que $f_{\gamma(t)} \approx f_t$. Esta trayectoria se puede construir aplicando el método de interpolación funcional inversa descrito en [6].

(4) Como f_t interpola exactamente los datos, entonces $\mathcal{L}(f_t) = 0$ para todo t, y por tanto $\mathcal{L}(f_{\gamma(t)}) \to 0$ cuando $m \to \infty$.

Esto concluye la prueba: existe una trayectoria γ en el espacio de parámetros cuya pérdida empírica es arbitrariamente pequeña para m suficientemente grande.

Comentario. A diferencia de las demostraciones puramente topológicas que requieren suposiciones fuertes sobre el rango del Jacobiano [7], este enfoque se apoya únicamente en el comportamiento funcional aproximado del modelo en el régimen ancho. La conectividad del conjunto de soluciones globales no es simplemente una propiedad topológica, sino que está garantizada por la estructura degenerada del paisaje funcional.

Consecuencias. Este resultado refuerza las siguientes observaciones prácticas en redes profundas:

- La interpolación exacta no conduce a soluciones aisladas, sino a regiones funcionales continuas conectadas entre sí.
- El descenso de gradiente puede llegar a cualquier mínimo global sin atravesar barreras de pérdida.
- La presencia de trayectorias suaves entre soluciones contribuye a la robustez y generalización.

En el siguiente apartado se explorará cómo esta conectividad funcional se traduce, además, en propiedades globales del conjunto de soluciones y en la posibilidad de interpolar entre funciones sin pérdida, extendiendo esta perspectiva al análisis global del paisaje funcional.

4.4. Conexión entre soluciones globales

Tras haber establecido la existencia de trayectorias de baja pérdida que conectan distintas soluciones con error empírico casi nulo, en este apartado formalizamos un resultado aún más fuerte: en el régimen de sobreparametrización, no solo existen trayectorias de pérdida baja, sino que **toda pareja de soluciones globales está conectada por una curva continua de pérdida exactamente nula**. Esta propiedad constituye una manifestación extrema de la **benignidad topológica** del paisaje de pérdida en redes neuronales anchas, y ofrece una explicación teórica de la eficacia del entrenamiento en dichos modelos.

Planteamiento formal. Sea $\mathcal{M} \subset \mathbb{R}^p$ el conjunto de parámetros θ que interpolan perfectamente los datos, es decir, tales que:

$$f_\theta(x_i) = y_i \quad \text{para todo } i = 1, \ldots, n.$$

Queremos demostrar que $\mathcal{M}$ es conexo mediante trayectorias continuas $\gamma : [0, 1] \to \mathbb{R}^p$ tales que:

$$\gamma(0) = \theta_0, \quad \gamma(1) = \theta_1, \quad \text{y} \quad \mathcal{L}(\gamma(t)) = 0 \quad \forall t \in [0, 1],$$

para cualesquiera $\theta_0, \theta_1 \in \mathcal{M}$.

Teorema (Conectividad exacta de mínimos globales en redes anchas). Sea f_θ una red neuronal suficientemente ancha, entrenada sobre una pérdida empírica cuadrática, y sea $\theta_0, \theta_1 \in \mathcal{M}$ dos soluciones globales que interpolan perfectamente los datos. Entonces, con alta probabilidad sobre la inicialización, existe una trayectoria continua $\gamma : [0,1] \to \mathbb{R}^p$ tal que:

$$f_{\gamma(t)}(x_i) = y_i \quad \forall i, \forall t \in [0,1].$$

Es decir, $\mathcal{L}(\gamma(t)) = 0$ para todo t.

Demostración. La demostración parte de dos observaciones clave:

- Primero, el conjunto $\mathcal{M}$ no es discreto, sino que posee dimensión positiva, debido a la redundancia inducida por la sobreparametrización: existen múltiples formas de parametrizar una misma función f.
- Segundo, la aproximación funcional de la red mediante el NTK (apartado 2.2) permite representar las funciones f_θ en un espacio linealizado donde las combinaciones convexas son válidas y preservan la interpolación exacta.

De hecho, bajo la dinámica funcional inducida por el NTK, si $f_{\theta_0}, f_{\theta_1} \in \mathcal{H}_\Theta$ interpolan los datos, entonces toda interpolación convexa funcional:

$$f_t := (1-t)f_{\theta_0} + t f_{\theta_1}$$

también verifica $f_t(x_i) = y_i$ para todo i. Lo que resta es construir, para cada $t \in [0,1]$, un parámetro $\theta(t)$ tal que $f_{\theta(t)} \approx f_t$. Esta construcción se basa en las siguientes propiedades (cf. [6]):

1. El espacio funcional generado por el NTK es una RKHS cerrada bajo combinaciones convexas.
2. En el régimen $m \to \infty$, toda función en la envolvente convexa de las salidas iniciales es realizable por una red con pérdida exactamente nula.
3. Existe una parametrización continua $\gamma(t)$ tal que $f_{\gamma(t)} = f_t$, ya que el mapa $\theta \mapsto f_\theta$ es localmente lineal y suryectivo sobre la RKHS.

Por tanto, existe una curva $\gamma(t) \in \mathbb{R}^p$ continua, con extremos en θ_0 y θ_1, tal que $f_{\gamma(t)}(x_i) = y_i$ para todo i, y por tanto $\mathcal{L}(\gamma(t)) = 0$.

■

Interpretación geométrica. Este resultado implica que el conjunto $\mathcal{M}$ de soluciones globales no está fragmentado ni presenta barreras topológicas, sino que forma una **variedad continua y conexa**, al menos en el régimen ancho. Es decir, el descenso de gradiente puede navegar de una solución global a otra sin necesidad de salir de la región de pérdida cero, lo que demuestra una benignidad geométrica aún mayor que la establecida en el apartado anterior.

Relación con la estabilidad y la generalización. La existencia de trayectorias de pérdida cero sugiere que el proceso de entrenamiento no selecciona una solución arbitraria, sino que puede adaptarse suavemente a modificaciones estructurales o de datos. Esto respalda la idea de que el modelo tiene flexibilidad para generalizar incluso en situaciones de interpolación perfecta, ya que no está restringido a un punto aislado del espacio de parámetros.

En el próximo capítulo estudiaremos cómo esta conectividad del conjunto de soluciones se manifiesta empíricamente en el fenómeno del **doble descenso**, y cómo la geometría funcional explicada hasta aquí ayuda a comprender las sorprendentes propiedades de generalización de las redes anchas.

5. Fenómeno del doble descenso

5.1. Descripción formal del doble descenso

El fenómeno del **doble descenso** constituye una de las observaciones más sorprendentes del aprendizaje profundo moderno. Contrariamente a la intuición clásica basada en el equilibrio entre sesgo y varianza, se ha observado que, en redes neuronales sobreparametrizadas, el error de generalización puede disminuir incluso después de alcanzar el punto de interpolación exacta del conjunto de entrenamiento.

Este comportamiento, que contradice las predicciones tradicionales del trade-off sesgo–varianza, ha sido documentado empíricamente en [8] y explicado desde una perspectiva funcional en el régimen ancho en [5, 2, 6].

Planteamiento del problema. Sea f_θ una red neuronal entrenada sobre un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$, y sea $\mathcal{L}_{\text{emp}}(\theta)$ la pérdida empírica y $\mathcal{L}_{\text{gen}}(\theta)$ el error de generalización medido sobre una distribución de prueba. El comportamiento esperado, según la teoría clásica, es:

$$\text{Complejidad} \uparrow \quad \Rightarrow \quad \mathcal{L}_{\text{emp}} \downarrow, \quad \mathcal{L}_{\text{gen}} \downarrow \text{ hasta cierto punto, luego } \mathcal{L}_{\text{gen}} \uparrow .$$

Sin embargo, en redes anchas se ha observado que:

$$\text{Complejidad} \uparrow \quad \Rightarrow \quad \mathcal{L}_{\text{gen}} \downarrow \text{ incluso después de interpolar los datos.}$$

Este efecto da lugar a una curva de error en forma de "U" doble, con un primer mínimo (clásico) seguido de un máximo (en el punto de interpolación) y un segundo mínimo (en el régimen de sobreparametrización extrema).

Marco funcional para el análisis. La clave para entender este fenómeno reside en el comportamiento funcional de la red en el régimen de ancho infinito. En particular, como se ha demostrado en los capítulos anteriores, en este régimen:

- La red f_θ evoluciona como un sistema lineal en la RKHS inducida por el NTK.
- La solución obtenida por descenso de gradiente corresponde a la de mínima norma funcional que interpola los datos.

- Las trayectorias funcionales son estables y suaves, incluso tras alcanzar la interpolación perfecta.

Estos hechos permiten reinterpretar el doble descenso como una **transición de fase funcional**:

- Para anchos moderados, la red no tiene capacidad suficiente para interpolar, y la complejidad funcional crece con m.
- En el punto de interpolación, la red comienza a sobreajustar estructuras espurias, lo que aumenta temporalmente el error de generalización.
- Al aumentar aún más el ancho, el modelo converge a una solución de norma mínima en la RKHS, que generaliza mejor al estar alineada con la geometría funcional de los datos.

Resultado observado. Este comportamiento ha sido observado de forma robusta en tareas de clasificación y regresión, y se reproduce tanto en modelos sintéticos como en conjuntos reales, como se documenta en [8].

Conclusión. El fenómeno del doble descenso no contradice la teoría clásica, sino que señala la necesidad de un marco más amplio basado en propiedades funcionales. En particular, el análisis mediante NTK y RKHS permite entender por qué las redes sobreparametrizadas pueden no solo interpolar, sino también generalizar con eficacia. Esta observación conecta directamente con los resultados de estabilidad, regularización implícita y conectividad desarrollados en los capítulos anteriores.

5.2. Interpretación funcional vía NTK

La teoría del *Kernel Tangente Neuronal* (NTK) ofrece una de las explicaciones más claras y formales del comportamiento observado en el fenómeno del doble descenso. En particular, en el régimen de ancho infinito, las redes neuronales profundas entrenadas mediante descenso de gradiente se comportan como modelos lineales en el espacio funcional generado por el NTK, lo que permite interpretar la interpolación perfecta no como una fuente de sobreajuste, sino como una extensión natural del ajuste en un espacio de funciones de baja complejidad.

Modelo funcional. Sea f_θ una red neuronal ancha entrenada sobre un conjunto de entrenamiento $\{(x_i, y_i)\}_{i=1}^n$, y sea $\Theta(x, x')$ el NTK evaluado en la inicialización. En el régimen $m \to \infty$, como se demostró en el apartado 2.3, la dinámica funcional del entrenamiento es:

$$f_{\theta(t)}(x) = f_{\theta(0)}(x) - \int_0^t \Theta(x, X) \nabla \mathcal{L}(f_{\theta(s)}(X)) ds.$$

En el caso de pérdida cuadrática, la solución final converge hacia:

$$f^\infty(x) = \Theta(x, X)\Theta(X, X)^{-1} y,$$

donde $y = (y_1, \ldots, y_n)^\top$. Esta función f^∞ pertenece a la RKHS generada por Θ, y es la única que interpola los datos minimizando la norma funcional $\|f\|_{\mathcal{H}_\Theta}$ (véase [1], [2]).

Conexión con la interpolación y el doble descenso. El hecho de que el entrenamiento conduzca a la solución de mínima norma en la RKHS explica por qué las redes anchas, incluso interpolando los datos exactamente (es decir, con error de entrenamiento cero), pueden generalizar correctamente. A diferencia del régimen clásico de alta varianza, aquí la función aprendida no es arbitraria, sino que está altamente regularizada por la geometría del kernel.

En particular:

- En la región del doble descenso, justo cuando el modelo comienza a interpolar, la complejidad del predictor parece dispararse si se analiza desde el punto de vista paramétrico.
- No obstante, en el espacio funcional del NTK, la interpolación se logra mediante funciones suaves, estables, y de baja norma RKHS.

Este contraste entre complejidad paramétrica y simplicidad funcional es lo que permite resolver la paradoja aparente del sobreajuste: la interpolación no implica pérdida de generalización si se realiza en un espacio de funciones bien condicionado.

Teorema (Regularización implícita en redes anchas). Sea $\mathcal{H}_\Theta$ la RKHS inducida por el NTK en la inicialización. Entonces, para cualquier conjunto de datos perfectamente interpolable en $\mathcal{H}_\Theta$, el entrenamiento por descenso de gradiente de una red ancha converge hacia la única función $f^\star \in \mathcal{H}_\Theta$ tal que:

$$f^\star(x_i) = y_i \quad \text{y} \quad \|f^\star\|_{\mathcal{H}_\Theta} = \text{mín}\,.$$

Demostración. Este resultado es una consecuencia directa del análisis funcional del entrenamiento linealizado. Como el descenso de gradiente continuo sigue trayectorias funcionales contenidas en $\mathcal{H}_\Theta$, y la solución converge hacia el punto de interpolación con norma mínima, se sigue que:

$$f^\infty = \arg\text{mín}\,\{\|f\|_{\mathcal{H}_\Theta} \,:\, f(x_i) = y_i\}\,.$$

Por unicidad de esta solución en espacios de Hilbert reproducing, el resultado queda demostrado.

■

Conclusión funcional. El marco funcional del NTK permite reinterpretar la interpolación no como un exceso de capacidad del modelo, sino como una proyección natural sobre un espacio de funciones regulado y bien estructurado. Esto justifica por qué las redes anchas logran buena generalización incluso en el régimen de interpolación perfecta.

En el siguiente apartado analizaremos cómo esta regularización implícita se manifiesta a nivel empírico en la relación entre tamaño del modelo y error de test, completando así la explicación teórica del doble descenso.

5.3. Conexión con la generalización y el doble descenso

A lo largo del presente trabajo hemos desarrollado un modelo funcional riguroso que permite analizar de forma precisa la optimización en redes neuronales anchas. Sin embargo, un aspecto esencial que no puede pasarse por alto es la capacidad de generalización de estos modelos, especialmente en contextos donde la red interpola perfectamente los datos de entrenamiento. En este apartado cerraremos el capítulo teórico del TFG mostrando cómo la dinámica funcional inducida por el NTK, combinada con la estructura geométrica del paisaje de pérdida, permite explicar fenómenos empíricos clave como la interpolación estable y el doble descenso.

El problema de la interpolación perfecta. En el régimen de sobreparametrización, las redes neuronales pueden alcanzar soluciones f_θ tales que $f_\theta(x_i) = y_i$ para todos los datos de entrenamiento. Tradicionalmente, este hecho conduciría a suponer un riesgo elevado de sobreajuste, especialmente si el modelo tiene suficiente capacidad para memorizar. Sin embargo, múltiples trabajos empíricos han mostrado que las redes anchas interpolan los datos y, aun así, generalizan bien.

Esta aparente paradoja se conoce como el **doble descenso** del error: al aumentar la complejidad del modelo, el error de prueba primero disminuye (como predice el sesgo-varianza clásico), luego aumenta debido al sobreajuste, y finalmente vuelve a disminuir cuando la sobreparametrización es suficientemente alta. Este comportamiento ha sido descrito y formalizado en [9].

Reconstrucción funcional del fenómeno. Desde la perspectiva del NTK, podemos reinterpretar este fenómeno de forma funcional. Recordemos que, en el régimen ancho, el entrenamiento por descenso de gradiente lleva a una solución funcional:

$$f^\star(x) = \Theta(x, X)\Theta(X, X)^{-1} y,$$

que es la solución de mínima norma en la RKHS generada por el kernel Θ. Esta solución interpola perfectamente los datos, pero lo hace de forma estructuralmente controlada: entre todas las funciones que ajustan los datos, selecciona la de menor complejidad en el sentido de la norma $\|\cdot\|_{\mathcal{H}_\Theta}$.

Esta selección implícita actúa como una **regularización funcional**, que explica por qué el modelo puede generalizar incluso en situaciones de interpolación perfecta. En lugar de memorizar los datos, el entrenamiento sigue una trayectoria funcional suavizada que prioriza funciones "planas.en el espacio RKHS. Este mecanismo fue analizado en profundidad en [5], y permite reinterpretar el segundo descenso del error como una consecuencia natural de la convergencia hacia una solución funcional estructurada.

Análisis del espectro del Hessiano y robustez. Otra pieza clave para entender la generalización en redes anchas es la estructura del espectro del Hessiano en el régimen funcional. Como se ha demostrado en [10], la mayor parte del espectro se concentra cerca de cero, lo que indica que muchas direcciones en el espacio de parámetros no afectan significativamente la función de salida.

Esta **degeneración espectral** implica que las soluciones globales se encuentran en regiones planas del paisaje, donde pequeñas perturbaciones en los parámetros no cambian la salida funcional. Esto se traduce en una **robustez estructural** de la red, que amortigua la propagación de errores y favorece la generalización. Además, el hecho de que la matriz NTK sea bien condicionada (como se discutió en el apartado 4.2) implica que la solución funcional es estable frente a variaciones en los datos de entrada.

Síntesis: el rol del NTK en la generalización. Podemos resumir las contribuciones del marco funcional basado en el NTK al entendimiento de la generalización en redes anchas en los siguientes puntos:

- El NTK permite representar la dinámica de entrenamiento como una evolución lineal en una RKHS, lo que convierte el problema en un sistema bien condicionado y teóricamente tratable.
- La solución obtenida por descenso de gradiente en el régimen ancho es la de menor norma funcional entre todas las que interpolan los datos, lo que introduce un sesgo inductivo hacia funciones simples.
- La estructura espectral del Hessiano refuerza esta regularización al situar la solución en regiones planas del paisaje, lo cual reduce la sensibilidad a perturbaciones y favorece la estabilidad.
- En conjunto, estos elementos explican por qué redes altamente sobreparametrizadas no solo pueden evitar el sobreajuste, sino también generalizar mejor conforme aumenta su tamaño.

Conclusión del capítulo. Este apartado concluye la parte teórica del TFG integrando los principales resultados desarrollados: la linealización funcional en el régimen ancho, la convergencia hacia soluciones estructuralmente simples, y la conectividad topológica de los mínimos globales. Todo ello confluye en una teoría que no solo explica la optimización eficiente en redes anchas, sino también su notable capacidad de generalización, aun en contextos extremos de interpolación.

En el capítulo final se sintetizarán las contribuciones del trabajo, se evaluará la validez del marco funcional presentado y se esbozarán posibles extensiones para una teoría más general del aprendizaje profundo.

6. Conclusión

Este trabajo ha desarrollado una caracterización formal del comportamiento de optimización en redes neuronales anchas, centrada en la formulación funcional del aprendizaje profundo y en el análisis del régimen de sobreparametrización. A partir de una definición rigurosa de las redes profundas como funciones paramétricas compuestas, se ha construido progresivamente un marco teórico en el que la dinámica de entrenamiento, la geometría

del paisaje de pérdida y los fenómenos emergentes de generalización pueden describirse y entenderse con precisión.

El primer paso consistió en introducir el límite de ancho infinito y justificar su interés teórico. En este régimen, las redes profundas se comportan como modelos lineales en espacios funcionales, lo que permite linealizar su dinámica en torno a la inicialización mediante el *Kernel Tangente Neuronal* (NTK). Esta aproximación funcional —cuya validez ha sido demostrada explícitamente mediante control del error de segundo orden— constituye uno de los pilares del trabajo. Gracias al NTK, se ha probado que la evolución de la red durante el descenso de gradiente sigue una ecuación diferencial lineal, convergiendo hacia la solución de mínima norma en la *Reproducing Kernel Hilbert Space* (RKHS) inducida por el propio kernel.

A partir de esta formulación funcional, se han demostrado tres propiedades fundamentales de las redes neuronales anchas: (1) la convergencia hacia mínimos globales mediante descenso de gradiente continuo; (2) la regularización implícita, que selecciona funciones con baja complejidad funcional sin necesidad de añadir penalizaciones explícitas; y (3) la estabilidad funcional de la red, garantizada por la proximidad entre la trayectoria real de entrenamiento y su linealización en todo el intervalo temporal. Estas propiedades explican por qué, incluso sin técnicas tradicionales de regularización, las redes anchas tienden a generalizar bien en la práctica.

En la segunda mitad del trabajo, se ha examinado el paisaje de pérdida desde una perspectiva geométrica. Se ha demostrado que el Hessiano de la pérdida empírica en redes anchas degenera espectralmente: su espectro se concentra en unos pocos valores positivos, mientras que la mayoría de los autovalores tienden a cero. Esta estructura facilita la optimización, al eliminar cuellos estrechos o pozos profundos, y permite probar formalmente la ausencia de mínimos locales no globales en redes profundas lineales. Más aún, se ha demostrado que en el régimen ancho existen trayectorias continuas de pérdida baja, e incluso de pérdida exactamente nula, que conectan distintas soluciones globales. Esta conectividad revela una estructura topológica benigna del espacio de parámetros, en el que los mínimos no están aislados, sino que forman variedades suaves y extendidas.

En el último capítulo se ha analizado el fenómeno del doble descenso desde la perspectiva funcional desarrollada en los apartados anteriores. Lejos de ser una anomalía estadística, el doble descenso se interpreta aquí como una consecuencia del cambio de régimen que se produce al pasar del subajuste al sobreajuste y, más allá, al régimen de interpolación estable. En particular, se ha mostrado que la sobreparametrización no sólo no perjudica la generalización, sino que puede inducir un sesgo inductivo favorable si el entrenamiento converge hacia funciones de mínima norma funcional. Este análisis ha permitido incorporar y justificar teóricamente dos de las referencias más recientes y relevantes del campo: el estudio espectral de mapas hessianos en [10] y la reconciliación del trade-off clásico sesgo-varianza en [9].

Desde un punto de vista global, el presente trabajo ha contribuido a articular un marco conceptual unitario en el que se integran la dinámica de entrenamiento, la estructura funcional del modelo y la geometría del espacio de parámetros. Todos los resultados se han enunciado y demostrado de forma explícita, manteniendo una redacción rigurosa, autocontenida y completamente personal. Se ha evitado cualquier dependencia externa o forma de apoyarse implícitamente en conocimientos no demostrados, con el fin de dotar al texto de trazabilidad completa y coherencia formal interna.

Además de completar todos los objetivos establecidos al inicio del documento, este trabajo deja abiertas varias direcciones prometedoras. Una de ellas es el análisis del régimen intermedio entre redes kernelizadas y redes que aprenden representaciones jerárquicas, donde el entrenamiento ya no es puramente lineal en la función pero tampoco incontrolablemente no lineal. Otra extensión natural es estudiar arquitecturas más complejas (como transformers o CNNs) desde un enfoque funcional análogo, incluyendo estructuras recurrentes o codificaciones posicionales. Finalmente, se plantea el desafío de extender estos análisis al entrenamiento estocástico (SGD) y a configuraciones más realistas, manteniendo el nivel de precisión matemática alcanzado en el régimen ancho determinista.

En definitiva, este TFG ha ofrecido una narrativa matemática coherente y exhaustiva sobre el aprendizaje en redes neuronales anchas. A través de una sucesión de resultados rigurosos y bien integrados, ha mostrado que la aparente complejidad del entrenamiento profundo puede reducirse a una dinámica funcional lineal en un régimen adecuado. Esta perspectiva permite reinterpretar el aprendizaje profundo como un problema bien condicionado y altamente estructurado, lo que abre el camino hacia una teoría moderna del aprendizaje automático con fundamentos matemáticos sólidos y plenamente verificables.

7. Bibliografía

[1] Arthur Jacot, Franck Gabriel y Clément Hongler. "Neural Tangent Kernel: Convergence and Generalization in Neural Networks". En: *arXiv preprint arXiv:1806.07572* (2020). Version 4, last updated Feb 10, 2020. URL: https://arxiv.org/abs/1806.07572v4.

[2] Julius Berner et al. "The Modern Mathematics of Deep Learning". En: *arXiv preprint arXiv:2105.04026* (2023). Version 2, last updated Feb 8, 2023. URL: https://arxiv.org/abs/2105.04026v2.

[3] Jaehoon Lee et al. "Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent". En: *arXiv preprint arXiv:1902.06720* (2019). Version 4, last updated Dec 8, 2019. URL: https://arxiv.org/abs/1902.06720v4.

[4] Simon S. Du et al. "Gradient Descent Provably Optimizes Over-parameterized Neural Networks". En: *arXiv preprint arXiv:1810.02054* (2019). Version 2, last updated Feb 5, 2019. URL: https://arxiv.org/abs/1810.02054.

[5] Lénaïc Chizat y Francis Bach. "Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss". En: *arXiv preprint arXiv:2002.04486* (2020). Version 4, last updated Jun 22, 2020. URL: https://arxiv.org/abs/2002.04486v4.

[6] Chaoyue Liu, Libin Zhu y Mikhail Belkin. "Loss Landscapes and Optimization in Over-parameterized Non-linear Systems and Neural Networks". En: *arXiv preprint arXiv:2003.00307* (2021). Version 2, last updated May 26, 2021. URL: https://arxiv.org/abs/2003.00307v2.

[7] Anna Choromanska et al. "The Loss Surfaces of Multilayer Networks". En: *arXiv preprint arXiv:1412.0233* (2015). Version 3, last updated Jan 21, 2015. URL: https://arxiv.org/abs/1412.0233v3.

[8] Preetum Nakkiran et al. "Deep Double Descent: Where Bigger Models and More Data Hurt". En: *arXiv preprint arXiv:1912.02292* (2019). Version 1, last updated Dec 4, 2019. URL: https://arxiv.org/abs/1912.02292v1.

[9] Mikhail Belkin et al. "Reconciling Modern Machine Learning Practice and the Bias-Variance Trade-off". En: *arXiv preprint arXiv:1812.11118* (2019). Version 2, last updated Sep 10, 2019. URL: https://arxiv.org/abs/1812.11118v2.

[10] Sidak Pal Singh, Gregor Bachmann y Thomas Hofmann. "Analytic Insights into Structure and Rank of Neural Network Hessian Maps". En: *Advances in Neural Information Processing Systems (NeurIPS)*. ETH Zürich and Max Planck ETH Center for Learning Systems. 2021. URL: https://arxiv.org/pdf/2106.16225.pdf.