Shannon Entropy for Early Stopping: A New Criterion for Neural Networks

Javier Herrero Pérez
jherrerop88@gmail.com
August 28, 2025

Abstract

Identificar el punto óptimo de early stopping en el entrenamiento de redes neuronales es un desafío. En este trabajo, se propone un enfoque analítico novedoso, basado en la física estadística. Se estudia la Entropía de Shannon de los pesos de la segunda capa oculta (W_2) de un Perceptrón Multicapa (MLP) entrenado en el conjunto de datos de MNIST. Los resultados demuestran que la Entropía de los pesos de esta capa alcanza un mínimo de incertidumbre justo cuando la red logra la máxima capacidad de generalización. Este mínimo de Entropía establece un equilibrio entre la minimización de la pérdida de validación y la estabilidad interna de los pesos. Concluyendo que la Entropía de los pesos en la capa de salida es un criterio robusto para la detención temprana, proporcionando una visión profunda y teóricamente fundamentada sobre el estado interno del modelo.

Palabras clave: Redes neuronales, Entropía de Shannon, física estadística, sobreajuste, MNIST, aprendizaje automático, early stopping.

1 Introducción

El entrenamiento de redes neuronales es un proceso iterativo cuyo éxito depende de buscar un equilibrio entre maximizar la precisión y asegurarse de la capacidad de generalización del modelo a datos no vistos. La tarea de reconocimiento de dígitos del conjunto de datos MNIST ha servido históricamente como un campo de pruebas fundamental para validar arquitecturas y algoritmos de aprendizaje. En este contexto, la métrica tradicional para evaluar el rendimiento de un modelo es la precisión en un conjunto de validación, que se estudia para evitar sobreajustes, fenómeno en el que el modelo memoriza el conjunto de entrenamiento en detrimento de su capacidad de generalización.

El sobreajuste es un problema recurrente en el entrenamiento de redes neuronales, y el uso de técnicas de detención temprana (early stopping) es una práctica estándar para mitigar sus efectos. Sin embargo, como se ha señalado, la elección del momento óptimo para detener el entrenamiento es una cuestión no trivial y ha sido objeto de una considerable investigación [1]. En este trabajo, proponemos un nuevo criterio de detención temprana.

El uso de conceptos de la teoría de la información para analizar el comportamiento de las redes neuronales ha ganado tracción en la literatura. Por ejemplo, algunos trabajos han explorado la integración de la Entropía de los pesos como un término de regularización en la función de pérdida para mejorar la capacidad de generalización del modelo [2]. En contraste, nuestro estudio adopta un enfoque diferente, utilizando la Entropía de los pesos como una métrica de diagnóstico a posteriori para monitorear la dinámica de convergencia.

Este artículo propone un estudio que va más allá de las métricas convencionales. Desde una perspectiva de la física aplicada, específicamente la física estadística, exploramos la dinámica interna del modelo analizando la **Entropía de Shannon** de la matriz de pesos de la segunda capa oculta durante el entrenamiento.

La Entropía de los pesos ofrece una perspectiva sobre la cantidad de información que tenemos de la matriz de pesos, de tal forma que un mínimo en esta métrica indicaría una baja

incertidumbre, lo que podría correlacionarse con el momento en el que el modelo ha alcanzado su configuración más estable y ordenada. Presento resultados de un modelo Multi-Layer Perceptron (MLP) [3] entrenado en el conjunto de datos MNIST, demostrando que el punto de máxima convergencia interna (mínima Entropía) coincide con la estabilización de la precisión en el conjunto de validación, proporcionando así un indicador alternativo y robusto para la detención temprana del entrenamiento.

2 Metodología

Para llevar a cabo el análisis se implementó una red neuronal simple y se la entrenó sobre el conjunto de datos MNIST, siguiendo una metodología que integra tanto métricas de rendimiento tradicionales como un análisis complementario basado en la Entropía de Shannon.

2.1 Conjunto de datos y preprocesamiento

El estudio se realizó sobre el conjunto de datos de MNIST de dígitos [4]. El dataset se dividió en dos:

- **Entrenamiento:** 37000 muestras, utilizadas para entrenar el modelo.
- **validación:** 5000 muestras, utilizadas para evaluar el rendimiento del modelo durante el entrenamiento.

Cada imagen en el conjunto de datos consiste en una cuadrícula de 28×28 píxeles. Las intensidadades de los píxeles varían entre 0 y 255, fueron normalizadas al rango $[0, 1]$ para optimizar el entrenamiento.

2.2 Algoritmo del estudio

Se empleó una red neuronal artificial de tipo Perceptrón Multicapa (MLP) con una arquitectura simple de dos capas ocultas, utilizando la propagación hacia atrás [5]. La configuración del

modelo fue la siguiente:

- **Capa de entrada:** $28 \times 28 = 784$ neuronas.
- **Capa oculta:** 256 neuronas con la función de activación ReLU definida como $f(z) = max(0, z)$ [6].

  Se optó por una capa oculta con $256 = 2^8$ neuronas. Esta elección no solo se justificó por la necesidad de una red con suficiente capacidad para aprender las complejas características del conjunto de datos MNIST, sino también para asegurar un número de parámetros lo suficientemente grande para que el análisis de la Entropía de Shannon fuera estadísticamente significativo. Con una matriz de pesos lo suficientemente grande, la distribución de los valores de los pesos se vuelve más estable, permitiendo que la Entropía refleje de manera más fiel la dinámica interna del sistema.
- **Capa de salida:** 10 neuronas (una para cada dígito del 0 al 9) con la función Softmax [6] que convierte las salidas en una distribución de probabilidad sobre las clases.

El modelo se entrenó utilizando el algoritmo de descenso de gradiente [5] para minimizar la función de pérdida. Se definieron los siguientes hiperparámetros de entrenamiento:

- **Función de pérdida:** Entropía cruzada categórica.
- **Tasa de aprendizaje (α):** Se fijó en 0.1.
- **Iteraciones:** El entrenamiento se ejecutó para $5 \cdot 10^4$ iteraciones.

La red fue entrenada para minimizar la función de pérdida (Entropía cruzada), un método estándar para problemas de clasificación multiclase. Esta función mide la diferencia entre la distribución de probabilidad de las etiquetas verdaderas y las predicciones del modelo [7].

Durante el entrenamiento, se monitorearon la pérdida y la precisión tanto en el conjunto de entrenamiento como en el de validación en cada iteración. La **pérdida de validación** se va a utilizar como métrica para detectar el sobreajuste, se calculó utilizando la función de pérdida de Entropía cruzada sobre el conjunto de validación.

2.3 Análisis de la Entropía de los pesos

Como parte central de este estudio, se extendió el proceso de entrenamiento para registrar la evolución de los parámetros del modelo. Los pesos (W_1, W_2) y los sesgos (b_1, b_2) se guardaron cada 200 iteraciones.

Posteriormente, se realizó un análisis de la dinámica interna del modelo a través de la Entropía de Shannon de los pesos. La Entropía de Shannon, un concepto fundamental en la teoría de la información y la física estadística [8], mide la incertidumbre o la cantidad de información que tenemos de un sistema. Para una variable discreta, se define como:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2(p(x_i)) \tag{1}$$

donde $p(x_i)$ es la probabilidad de que el resultado sea x_i y n es el número total de resultados posibles. En este estudio, aplicamos esta métrica a la distribución de los valores de los pesos de las capas ocultas.

Para calcular la probabilidad $p(x_i)$ de los valores de los pesos, que son una variable continua y pueden ser negativos, no es posible aplicar la fórmula directamente. En su lugar, se procedió a aproximar su distribución de probabilidad empírica. Se construyó un histograma de los valores de la matriz de pesos, dividiendo el rango de valores en un número fijo de intervalos (bins). La probabilidad $p(x_i)$ de un valor se definió como la frecuencia normalizada de su intervalo en el histograma.

La Entropía se calculó a partir de la distribución de los valores de los pesos, sirviendo como una medida de la incertidumbre o dispersión de los mismos en un momento dado del entrenamiento. Con el fin de normalizar la medida, la Entropía calculada se dividió por $\log_2(N)$, donde N es el número total de elementos en la matriz de pesos. Este análisis nos permitió identificar un punto de mínima Entropía que se correlaciona con la estabilización de la precisión de validación.

3 Resultados

El resultado del comportamiento de la Entropía de W_2 frente a la pérdida de validación se puede observar en la Figura 1, en esta Figura se han destacado tres puntos distintitvos:

- Fase 1: Máxima Entropía (H_2^{max}).
- Fase 2: Mínima Entropía (H_2^{min}).
- Fase 3: Mínimo en la pérdida de validación.

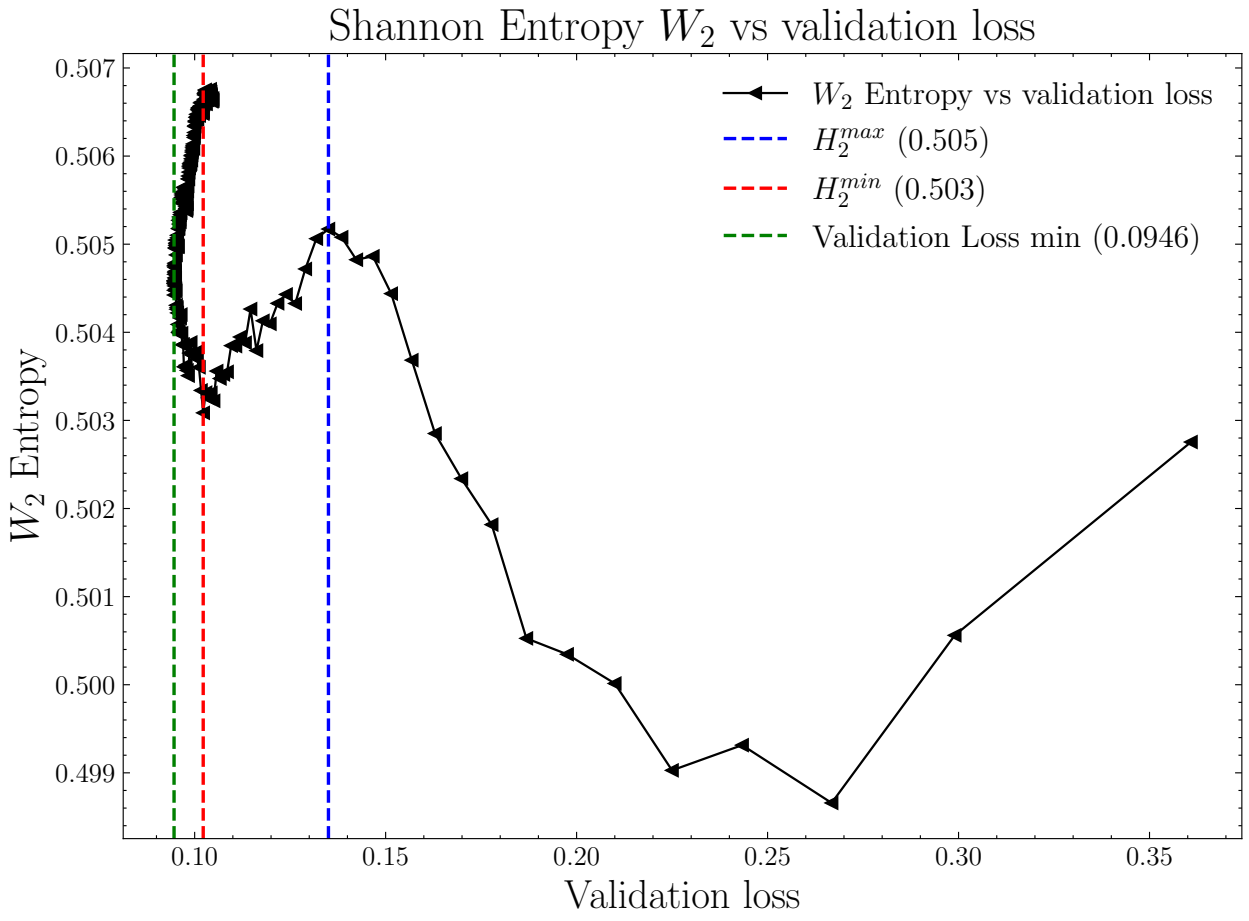

Figure 1: Evolución de la Entropía de los pesos de la capa (W_2) y la pérdida de validación a lo largo del entrenamiento. El gráfico muestra el máximo de Entropía (H_2^{max}) en la fase de exploración y el mínimo de Entropía (H_2^{min}), correlacionado con el inicio del aumento de la pérdida de validación. También se muestra la mínima pérdida de validación.

3.1 Fase 1: Exploración y máxima incertidumbre

En las primeras etapas del entrenamiento, el modelo se encuentra en una fase de exploración. La Entropía de Shannon para los pesos de W_2 aumenta hasta alcanzar un máximo en H_2^{max} alrededor de la iteración 3400, donde la precisión de entrenamiento se sitúa en un 97.73%. Este punto representa un estado de máxima incertidumbre en el sistema de pesos, ya que el modelo ha explorado un amplio abanico de posibles configuraciones antes de encontrar una dirección clara para la convergencia.

En la Tabla 1 se presentan los valores para este punto.

Iterations	Train accuracy	Validation loss	W_2 Entropy	W_1 Entropy
3400	0.9773	0.1350	0.5052	0.3006

Table 1: Métricas del modelo en la interación correspondiente al máximo de Entropía W_2.

3.2 Fase 2: Convergencia y mínimo de Entropía

A partir de la iteración 3400, la Entropía de los pesos comienza a decrecer de manera constante. Esta fase de descenso de la Entropía indica que el modelo ha encontrado una configuración de pesos más estable y ordenada, reduciendo la incertidumbre del sistema. La Entropía alcanza su valor mínimo, denotado como H_2^{min}, en la iteración 7800.

La Tabla 2 presenta las métricas claves del modelo en la interacción donde se alcanza el mínimo de Entropía. En este punto, el modelo ha alncanzado una precisión del 99.23%.

Iterations	Train accuracy	Validation loss	W_2 Entropy	W_1 Entropy
7800	0.9923	0.1022	0.5031	0.2890

Table 2: Métricas del modelo en la interación correspondiente al mínimo de Entropía W_2.

3.3 Fase 3: Mínimo en la pérdida de validación

Tras la iteración 7800, el modelo continua su entrenamiento y la Entropía de los pesos comienza a aumentar, lo que indica un ligero incremento en la aleatoriedad de su configuración interna.

A pesar de esto, el rendimiento externo del modelo mejora, alcanzando su mínimo de pérdida de validación en la iteración 15800. Como se puede observar en la Tabla 3, en este punto la pérdida de validación es de 0.0946, con una precisión de entrenamiento de 0.9992. Este momento representa el punto de detención temprana tradicional, basado únicamente en la minimización del error. Es crucial destacar que, en esta interación, la Entropía es superior a la mínima. Este hecho sugiere que la mejora en el rendimiento externo se ha logrado a expensas de la estabilidad interna del modelo, un indicio de sobreajuste leve.

Iterations	Train accuracy	Validation loss	W_2 Entropy	W_1 Entropy
15800	0.9992	0.0946	0.5044	0.2826

Table 3: Métricas del modelo en la iteración correspondiente al mínimo de la pérdida de validación.

Después de la iteración 15800, la pérdida de validación comienza a aumentar de manera constante, una señal inequívoca de que el modelo ha empezado a sobreajustarse al conjunto de entrenamiento. La Entropía de los pesos de W_2 continúa su ascenso en esta fase, lo que confirma que el modelo está sacrificando su capacidad de generalización para memorizar los datos de entrenamiento. Este comportamiento divergente, donde la pérdida de validación aumenta mientras que la Entropía de los pesos también lo hace, evidencia la degradaciín del rendimiento del modelo en datos no vistos.

3.4 Fase 2-3: Análisis comparativo y sobreajuste leve

La mejora en la pérdida de validación entre las iteraciones 7800 y 15800 se logra a expensas de un incremento en la Entropía de los pesos. Este fenómeno indica que el modelo ha abandonado su estado de máxima estabilidad interna para memorizar patrones, lo que se refleja con una distribución de pesos más "desordenada". Por lo tanto, el mínimo de Entropía de los pesos de la capa de salida proporciona una señal de detención temprana significativamente más conservadora y robusta, previniendo el sobreajuste leve.

Los datos de la Tabla 4 que comparan las métricas en los puntos de mínima entropía

ΔIterations	ΔTrain accuracy	ΔValidation loss	ΔW_2 Entropy	ΔW_1 Entropy
1.03	$6.95 \cdot 10^{-3}$	0.07	$2.58 \cdot 10^{-3}$	0.022

Table 4: Diferencias relativas en métricas clave del modelo, comparando el punto de mínima pérdida de validación con el de mínima entropía. Los valores se calculan con respecto al punto de mínima entropía como la base $\frac{|\text{fase 3} - \text{fase 2}|}{\text{fase 2}}$.

(fase 2) y y mínima pérdida de validación (fase 3), demuestran una divergencia fundamental. Se observa que el modelo requirió un 103% más de iteraciones para pasar de la fase de mínima entropía a la fase de mínima pérdida de validación. Este esfuerzo adicional de entrenamiento solo resultó en una mejora muy marginal en la precisión de entrenamiento ($6.95 \cdot 10^{-3}$). Estas pequeña ganancia en el rendimiento se produjo a costa de un aumento en la estabilidad interna del modelo, aumentando la entropía de los pesos, lo que indica que la red se volvió más desordenada. Reforzando la idea de que la detención temprana en el punto de mínima entropía no solo es más eficiente en términos de tiempo, sino que también previene el sobreajsute leve al evitar la degradación de la estabilidad interna por una ganancia de rendimiento insignificante.

3.5 Análisis de la Entropía en el peso W_1

Para contrastar el comportamiento de la capa de salida, se analizó también la Entropía de la primera capa oculta (W_1). La Figura 2 muestra la evolución de la Entropía de W_1 frente a la pérdida de validación. A diferencia de W_2, la Entropía de W_1 no exhibe un mínimo. Este comportamiento divergente se atribuye a los diferentes roles funcionales de cada capa. La capa W_1 se encarga de extraer características de bajo nivel de generales a partir de los datos de entrada, mientras que la capa W_2 se especializa en la tarea final de clasificación. Por lo tanto, la Entropía de W_1 refleja una dinámica de aprendizaje de características más graduales. Este hallazgo indica que la Entropía de la capa de salida es el indicador más sensible para monitorear la convergencia y el sobreajuste del modelo.

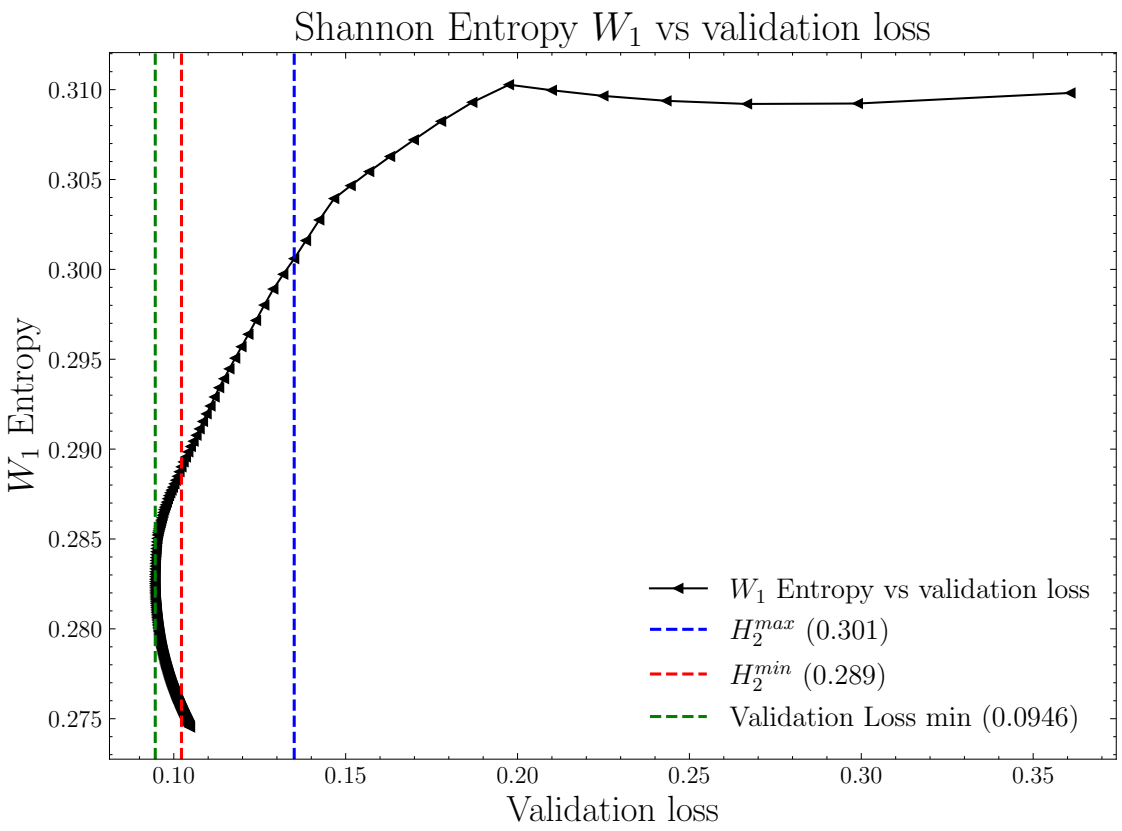

Figure 2: Comportamiento de la Entropía de la matriz de pesos de la capa W_1 frente a la pérdida de validación. La línea roja vertical indica la pérdida de validación en el punto de mínima Entropía de la capa W_2.

4 Discursión

El análisis de la Entropía de Shannon aplicado a los pesos de una red neuronal artificial ha proporcionado una perspectiva novedosa y un complemento valioso a las métricas de rendimiento tradicionales que se basan en la monitorización de la pérdida de validación. En este trabajo, se ha demostrado un enfoque analítico distinto, inspirado en la física estadística, puede proporcionar una visiín más profunda del estado interno de la red neuronal. El análisis de la entropía de Shannon de los pesos de la capa de salida W_2 ha revelado un mínimo en la entropía.

El hallazgo más significativo del estudio es la correlación entre el mínimo de entropía de los pesos y el punto de máxima estabilidad del modelo. Mientras que el criterio de detención

temprano tradicional habría sugerido continuar el entrenamiento hasta la iteración 15800 para alcanzar el mínimo de la pérdida de validación, los resultados indican que la configuración óptima y más estable de los pesos se logró mucho antes, en la iteración 7800. La ganancia en la pérdida de validación entre estos dos puntos fue marginal a cambio de un aumento de la incertidumbre de la matriz de pesos de W_2. Este fenómeno sugiere un sobreajuste leve que las métricas tradicionales son incapaces de detectar.

A diferencia de trabajos previos que utilizan la entropía como función de pérdida, el enfoque que se da a la entropía en este trabajo es de una métrica de estudio, sin modificar el proceso de optimización. Los resultados observados para W_1 y W_2 muestran que esta métrica es más sensible en las capas de salida.

5 Conclusiones

Este trabajo ha explorado la Entropía de Shannon de los pesos de las redes neuronales como un criterio analítico y complementario para la detención temprana del entrenamiento. A través de un estudio en un Perceptrón Multicapa (MLP) entrenado en el conjunto de datos de MNIST, se ha demostrado que el comportamiento de la entropía interna del modelo proporciona una visión fundamental que va más allá de las métricas de rendimiento tradicionales.

Los hallazgos claves son los siguientes:

- La entropía de los pesos de la capa de salida (W_2) exhibe un comportamiento predecible a lo largo del entrenamiento, alcanzando un mínimo de incertidumbre justo cuando el modelo logra su estado de máxima estabilidad y generalización.
- Este mínimo de entropía se produce significativamente antes del punto de mínima pérdida de validación. La continuación del entrenamiento hasta el mínimo de la pérdida de validación resultó en una mejora de rendimiento marginal a expensas de la estabilidad interna del modelo, lo que sugiere la presencia de un sobreajuste leve que la métrica tradicional es incapaz de detectar.

- El uso de la entropía de los pesos como criterio de detención temprana ofrece una soluciín más eficiente y robusta, ya que permite detener el entrenamiento en un punto óptimo de equilibrio entre rendimiento y estabilidad, previniendo el sobreajuste de manera más conservadora.

La clara divergencia entre la entropía de los pesos de la capa de entrada (W_1) y de la capa de salida (W_2) resalta la importancia de monitorizar las capas más especializadas para obtener indicadores fiables del estado del modelo. Este enfoque, inspirado en la física estadística, valida una nueva perspectiva para entender la dinámica interna del aprendizaje automático.

5.1 Futuras líneas de investigación

El presente estudio abre varias vías de investigación prometedoras. Proponemos las siguientes como continuaciones naturales de este trabajo:

- **Aplicación en otro sistema**: Validar este criterio en otras arquitecturas de red neuronal y en conjunto de datos más complejos.
- **Análisis de sesgos en los pesos:** El estudio se ha enfocado en los pesos. Un análisis más profundo podría ser estudiar los sesgos de la distribución de los pesos.

References

[1] Lutz Prechelt. "Early Stopping—But When?" In: *Neural Networks: Tricks of the Trade*. Ed. by Gene B. Orr and Klaus-Robert Müller. Berlin, Heidelberg: Springer, 1998, pp. 55–69.

[2] Ying Zheng et al. "Entropy Based Weight Regularization for Deep Neural Networks". In: *Entropy* 22.11 (2020), p. 1276.

[3] Simon S Haykin. *Neural networks: a comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall PTR, 1999, pp. 178–197.

[4] Kaggle. *Digit Recognizer*. https://www.kaggle.com/competitions/digit-recognizer/data. 2024.

[5] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536.

[6] Jingli Ren and Haiyan Wang. "Chapter 3 - Calculus and optimization". In: *Mathematical Methods in Data Science*. Ed. by Jingli Ren and Haiyan Wang. Elsevier, 2023, pp. 51–89. ISBN: 978-0-443-18679-0. DOI: https://doi.org/10.1016/B978-0-44-318679-0.00009-0. URL: https://www.sciencedirect.com/science/article/pii/B9780443186790000090.

[7] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[8] C E Shannon. "A Mathematical Theory of Communication". en. In: *The Bell system technical journal* (1948). URL: https://pure.mpg.de/rest/items/item_2383162_7/component/file_2456978/content.