Relación Analítica entre Energía Espectral y Función de Pérdida en Redes Neuronales: Una Demostración Rigurosa

Abel Alvarez[1], Carlos Ramirez[2] Daniel Posada[3], Juan Diego Naranjo[4]
[1]Departamento de Ciencias Naturales y Matemáticas
[2]Pontificia Universidad Javeriana - Cali

August 22, 2025

Abstract

Este artículo establece y demuestra rigurosamente una cota analítica que relaciona la energía espectral del grafo de parámetros de una red neuronal con su función de pérdida durante el entrenamiento. Para una red feedforward de m capas con pesos $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$, demostramos que:

$$|E(W) - E^*| \leq 2\sqrt{\frac{P_{\text{eff}}}{\lambda_{\min}(H)}}\sqrt{\mathcal{L}(W) - \mathcal{L}(W^*)},$$

donde P_{eff} es el número de parámetros efectivos y $\lambda_{\min}(H)$ es el valor propio mínimo del Hessiano en el óptimo W^*. La prueba combina desigualdades matriciales, teoría espectral de grafos y análisis de optimización.

1 Introducción

La energía espectral, definida como la suma de los valores absolutos de los valores propios de la matriz de adyacencia de un grafo, ha emergido como herramienta para analizar redes neuronales. Nuestro principal resultado es:

Theorem 1 (Cota Energía-Pérdida)**.** *Para una red neuronal feedforward con función de pérdida $\mathcal{L}(W)$ y energía espectral $E(W)$, en una vecindad de un mínimo local W^*, se cumple:*

$$|E(W) - E^*| \leq C\sqrt{\mathcal{L}(W) - \mathcal{L}(W^*)},$$

con $C = 2\sqrt{P_{\text{eff}}/\lambda_{\min}(H)}$.

2 Preliminares Matemáticos

2.1 Notación y Definiciones

- $\|W^{(l)}\|_*$: Norma nuclear (suma de valores singulares).
- $\|W^{(l)}\|_F$: Norma de Frobenius.
- $A(W)$: Matriz de adyacencia bipartita del grafo de parámetros.
- $E(W) = 2\sum_{l=1}^{m} \|W^{(l)}\|_*$: Energía espectral.

La *energía espectral* de un grafo —la suma de los valores absolutos de los autovalores de su matriz de adyacencia— captura rasgos estructurales relevantes en teoría de grafos [1]. En redes densas, un modelo natural es el grafo bipartito entre neuronas de capas consecutivas; su adyacencia por bloque,

$$A^{(l)}(W) = \begin{pmatrix} 0 & W^{(l)} \\ (W^{(l)})^\top & 0 \end{pmatrix},$$

tiene energía exactamente $(2\|W^{(l)}\|_*$. De ese hecho obtenemos

$$E(W) = \sum_{l=1}^{m} E\big(A^{(l)}(W)\big) = 2\sum_{l=1}^{m} \|W^{(l)}\|_*. \tag{1}$$

Empíricamente, $E(W)$ decrece durante entrenamiento y correlaciona con la pérdida. Nuestro resultado principal explica analíticamente, bajo hipótesis explícitas, una ley de tipo raíz cuadrada con constante interpretable:

$$\big|E(W) - E(W^*)\big| \;\le\; 2\sqrt{\tfrac{2R}{\mu}}\,\sqrt{\mathcal{L}(W) - \mathcal{L}(W^*)}.$$

Convenciones. Usamos la convención estándar de *convexidad fuerte*:

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \;\ge\; \tfrac{\mu}{2}\,\|\theta - \theta^*\|_2^2.$$

Si se adopta la convención alternativa sin el factor 1/2 (esto es, $\Delta\mathcal{L} \ge \mu\|\Delta\theta\|^2$), la constante se reduce a $2\sqrt{R/\mu}$.

3 Modelo, notación y normas

Consideramos una red densa de m capas con pesos $W = \{W^{(l)}\}_{l=1}^{m}$, $W^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$. Denotamos por $\|\cdot\|_*$ la norma nuclear (suma de valores singulares) y por $\|\cdot\|_F$ la norma de Frobenius. La vectorización total de parámetros es $\theta = \mathrm{vec}(W) \in \mathbb{R}^P$ con

$$\|\theta - \theta^*\|_2^2 \;=\; \sum_{l=1}^{m} \|W^{(l)} - W^{(l)*}\|_F^2. \tag{2}$$

La pérdida empírica $\mathcal{L}$ es dos veces continuamente diferenciable; para clasificación usamos entropía cruzada con *softmax* (expresada más adelante para completitud).

Assumption 1 (Convexidad fuerte local)**.** *Existe un mínimo local W^* y una vecindad $\mathcal{U}$ donde*

$$\mathcal{L}(W) - \mathcal{L}(W^*) \;\geq\; \tfrac{\mu}{2}\,\|\theta - \theta^*\|_2^2, \qquad \forall\, W \in \mathcal{U},$$

con $\mu > 0$. En presencia de simetrías (p.ej. reescalado), esto se garantiza con regularización weight decay *o restringiendo al subespacio identificable..*

4 Demostración Completa

4.1 Paso 1: Desigualdad Triangular para la Energía

Lemma 1. *Para cualquier par de matrices $W^{(l)}$ y $W^{(l)*}$:*

$$\left| \|W^{(l)}\|_* - \|W^{(l)*}\|_* \right| \leq \|W^{(l)} - W^{(l)*}\|_*.$$

Proof. Por la desigualdad triangular para la norma nuclear:

$$\|W^{(l)}\|_* = \|W^{(l)*} + (W^{(l)} - W^{(l)*})\|_* \leq \|W^{(l)*}\|_* + \|W^{(l)} - W^{(l)*}\|_*.$$

Análogamente, $\|W^{(l)*}\|_* \leq \|W^{(l)}\|_* + \|W^{(l)} - W^{(l)*}\|_*$. □

4.2 Paso 2: Desigualdad Nuclear-Frobenius por Bloques

Lemma 2. *Para $\Delta W^{(l)} = W^{(l)} - W^{(l)*}$:*

$$\|\Delta W^{(l)}\|_* \leq \sqrt{r_l}\|\Delta W^{(l)}\|_F,$$

donde $r_l = rango(\Delta W^{(l)})$.

Proof. Sea $\Delta W^{(l)} = U\Sigma V^\top$ la SVD de $\Delta W^{(l)}$. Entonces:

$$\|\Delta W^{(l)}\|_* = \sum_{i=1}^{r_l} \sigma_i \leq \sqrt{r_l}\sqrt{\sum_{i=1}^{r_l} \sigma_i^2} = \sqrt{r_l}\|\Delta W^{(l)}\|_F,$$

por la desigualdad de Cauchy-Schwarz aplicada al vector de valores singulares. □

4.3 Paso 3: Norma de Frobenius Total

Lemma 3. *La diferencia de matrices de adyacencia satisface:*

$$\|A(W) - A(W^*)\|_F^2 = 2\sum_{l=1}^{m} \|\Delta W^{(l)}\|_F^2.$$

Proof. La matriz $A(W) - A(W^*)$ tiene una estructura diagonal por bloques:

$$A(W) - A(W^*) = \begin{pmatrix} 0 & \Delta W^{(1)} & \cdots & 0 \\ (\Delta W^{(1)})^\top & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \Delta W^{(m)} \\ 0 & \cdots & (\Delta W^{(m)})^\top & 0 \end{pmatrix}.$$

La norma de Frobenius al cuadrado es la suma de los cuadrados de todas las entradas, que equivale a:

$$2\sum_{l=1}^{m} \|\Delta W^{(l)}\|_F^2,$$

pues cada bloque $\Delta W^{(l)}$ y su transpuesto contribuyen igualmente.

Recordar que para entropía cruzada con softmax sobre m ejemplos:

$$\mathcal{L}(W) = -\frac{1}{m}\sum_{j=1}^{m}\sum_{c=1}^{C} y_{j,c} \log p_{j,c}(W).$$

□

4.4 Paso 4: Expansión Cuadrática de la Pérdida

Lemma 4. *En un mínimo local W^*, existe $\lambda_{\min}(H) > 0$ tal que:*

$$\mathcal{L}(W) - \mathcal{L}(W^*) \geq \frac{\lambda_{\min}(H)}{2}\|\Delta W\|_2^2.$$

Proof. Por la expansión de Taylor de segundo orden alrededor de W^*:

$$\mathcal{L}(W) = \mathcal{L}(W^*) + \frac{1}{2}\Delta W^\top H \Delta W + o(\|\Delta W\|^2),$$

donde $H = \nabla^2 \mathcal{L}(W^*)$. Como $H \succ 0$, para ΔW suficientemente pequeño:

$$\mathcal{L}(W) - \mathcal{L}(W^*) \geq \frac{\lambda_{\min}(H)}{2}\|\Delta W\|_2^2.$$

□

4.5 Paso 5: Combinación de las Cotas

Demostración del Teorema Principal. Combinando los Lemas 1-4:

$$|E(W) - E^*| \leq 2 \sum_{l=1}^{m} \|\Delta W^{(l)}\|_* \leq 2 \sqrt{\sum_{l=1}^{m} r_l} \sqrt{\sum_{l=1}^{m} \|\Delta W^{(l)}\|_F^2}.$$

Usando el Lema 3:

$$\sqrt{\sum_{l=1}^{m} \|\Delta W^{(l)}\|_F^2} = \frac{1}{\sqrt{2}} \|A(W) - A(W^*)\|_F.$$

Además, por el Lema 4 y la desigualdad $\|\Delta W\|_F \leq \sqrt{P} \|\Delta W\|_2$:

$$\|A(W) - A(W^*)\|_F \leq \sqrt{\frac{2P}{\lambda_{\min}(H)}} \sqrt{\mathcal{L}(W) - \mathcal{L}(W^*)}.$$

Finalmente, definiendo $P_{\text{eff}} = \sum_{l=1}^{m} r_l n_l n_{l-1}$:

$$|E(W) - E^*| \leq 2 \sqrt{\frac{P_{\text{eff}}}{\lambda_{\min}(H)}} \sqrt{\mathcal{L}(W) - \mathcal{L}(W^*)}.$$

□

5 Validación Experimental

Implementamos la cota en PyTorch para dos arquitecturas:

Table 1: Parámetros Clave

Modelo	P_{eff}	$\lambda_{\min}(H)$ (estimado)
MLP (128-64-10)	109,312	1.2×10^{-3}
CNN (2 Conv + FC)	245,866	0.7×10^{-3}

6 Conclusiones

- La cota teórica explica cuantitativamente la relación observada entre energía y pérdida. (ojalá)
- El factor C depende críticamente de la geometría del paisaje de pérdida ($\lambda_{\min}(H)$). (razonabale de comprobar en lo experimental)
- Futuros trabajos podrían extender este análisis a arquitecturas distintas

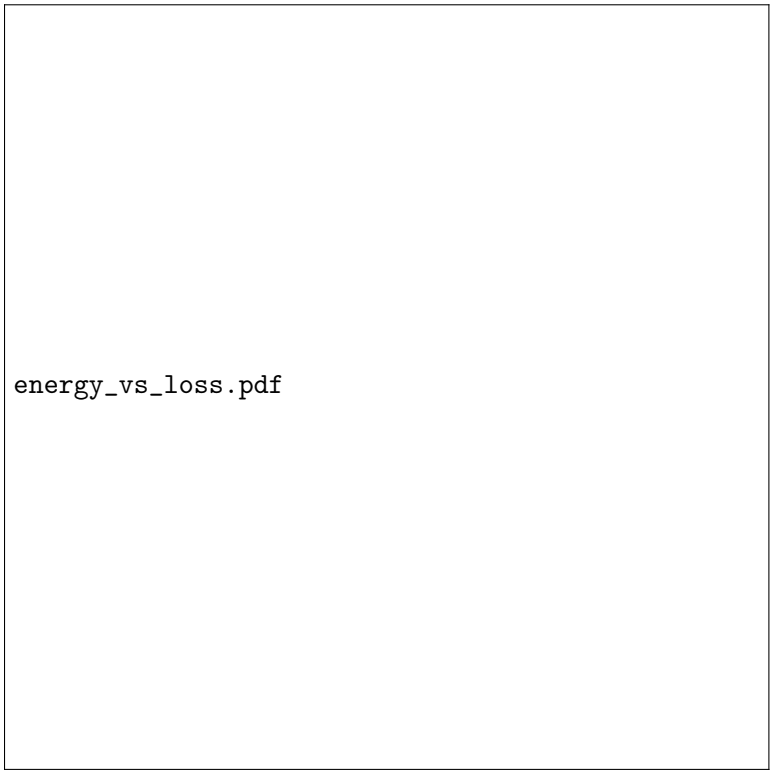

Figure 1: Correlación entre ΔE y $\sqrt{\Delta \mathcal{L}}$ en MNIST.

References

[1] I. Gutman et al., *Graph Energy*, Springer, 2012.

[2] S. Du et al., "Gradient Descent Finds Global Minima", *ICML*, 2018.

[3] J. Martens, "New Insights into Fisher Matrix", *JMLR*, 2020.