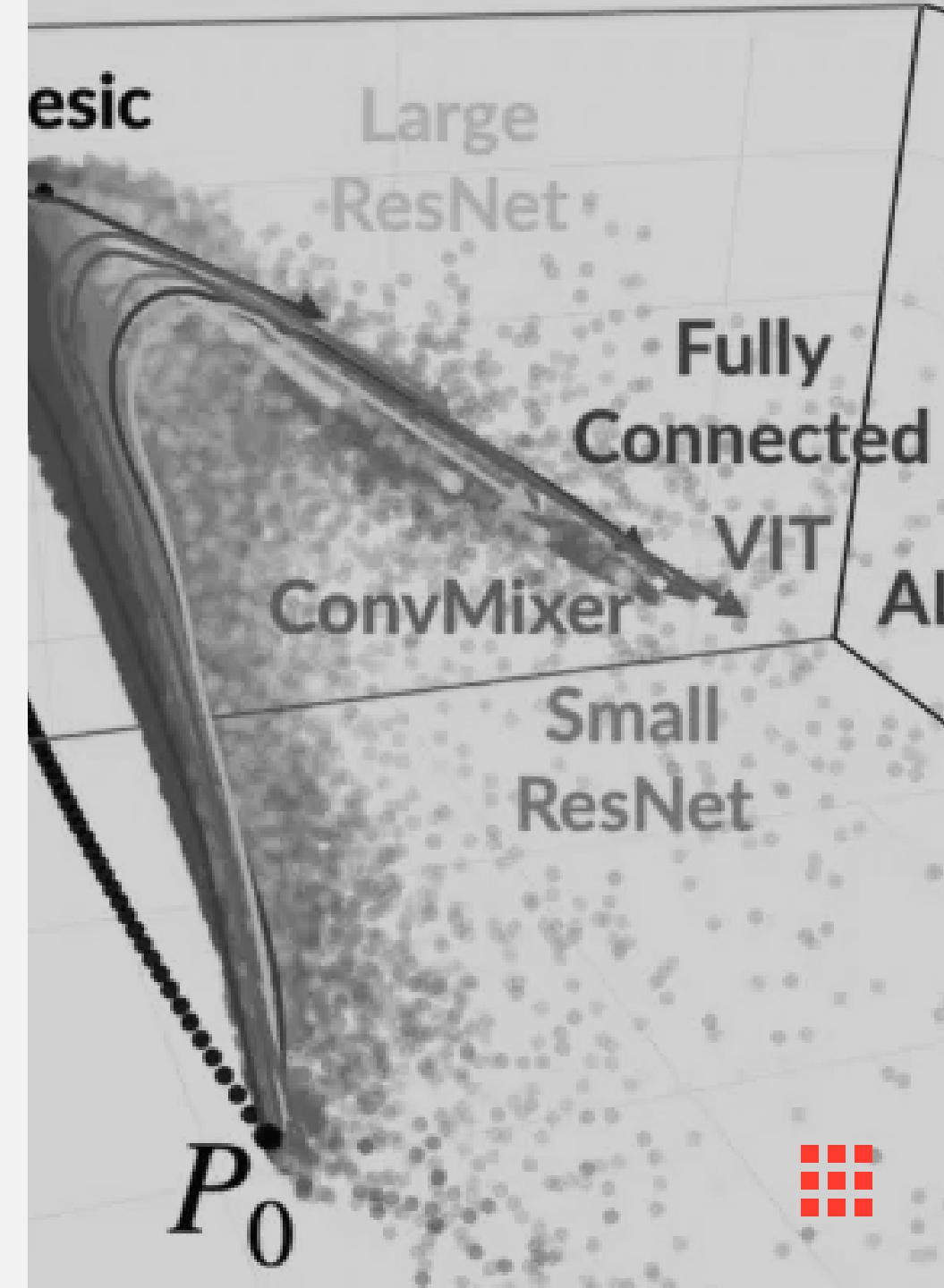


La Geometría de Trayectorias Determina la Estructura Algorítmica

Un Cambio de Paradigma en el Entrenamiento
de Redes Neuronales

POR GRISUNO



El Cambio de Paradigma Central

Vista Tradicional: La estructura algorítmica emerge pasivamente de la optimización.

Nuevo Paradigma: La estructura algorítmica se **construye activamente** mediante la manipulación precisa de la dinámica del gradiente.

Trayectorias Controladas

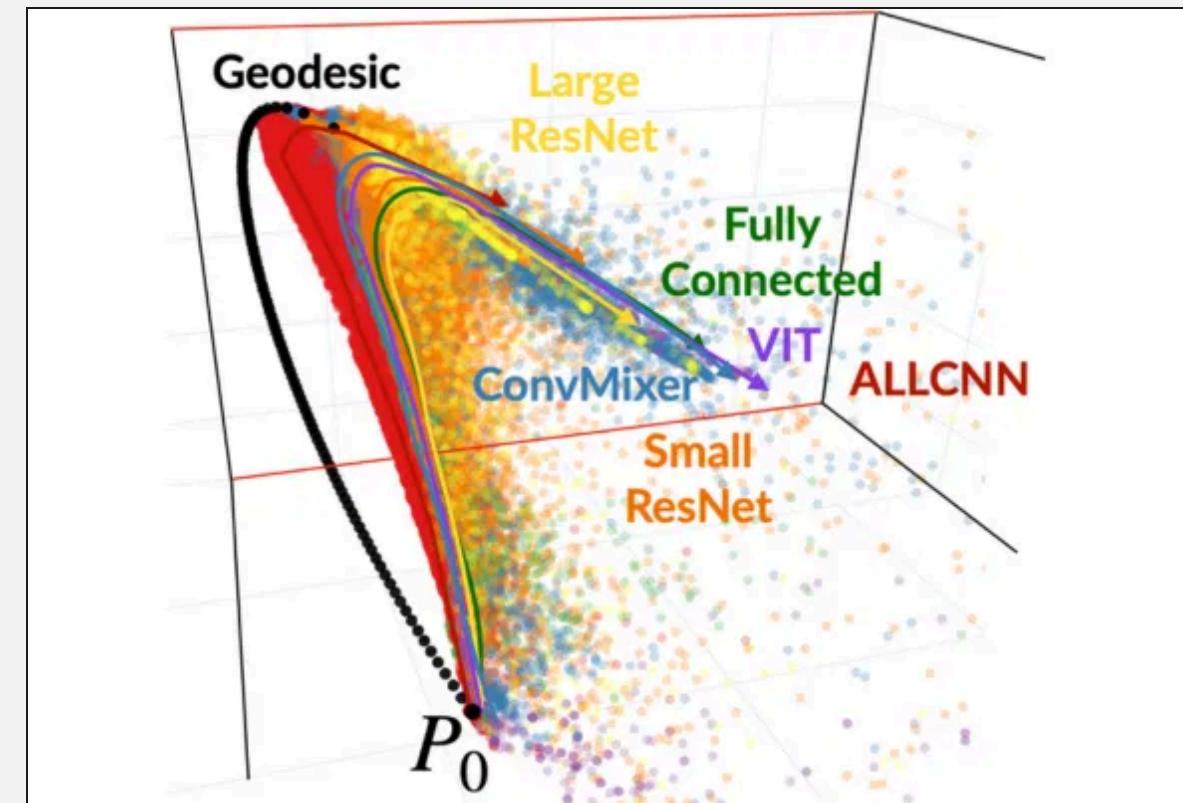
68% Éxito

No Controladas (Mínimos Locales)

32% Fallo

Tamaño de Lote Crítico

[24, 128]



Representación visual de trayectorias de entrenamiento. Valores de pérdida idénticos pueden resultar de diferentes caminos, pero solo geometrías específicas alcanzan soluciones algorítmicas genuinas.

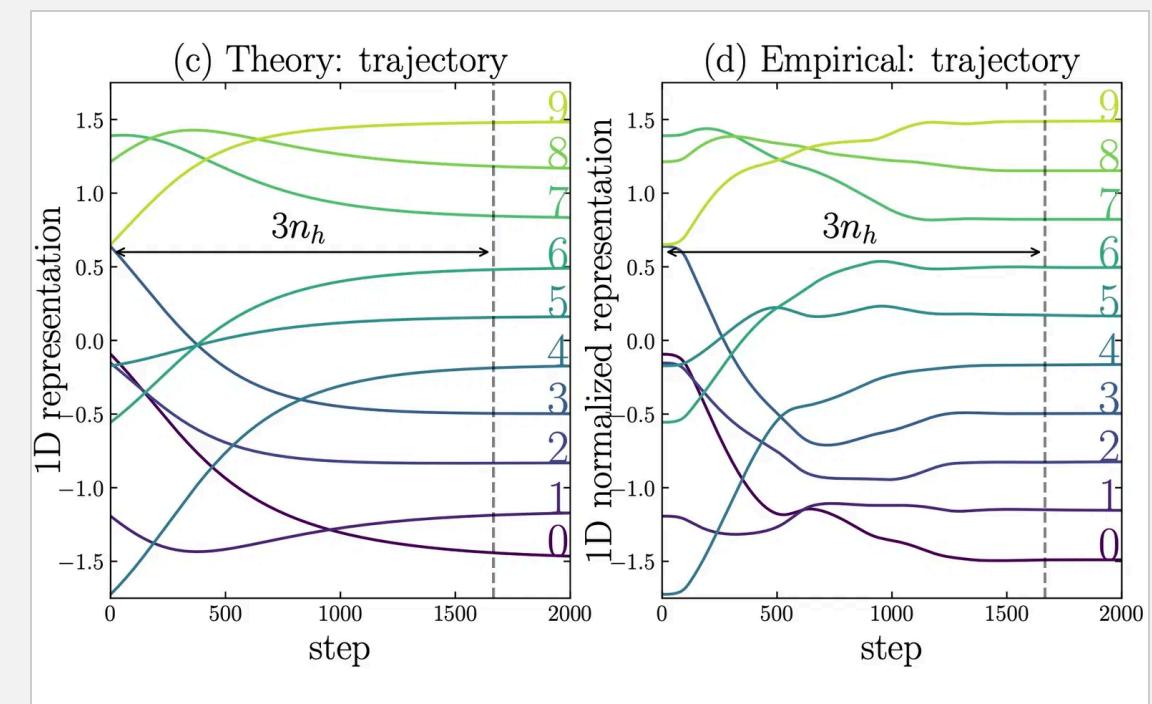
El Problema del Grokking y Verificación

Cuando una red hace grokking, ¿ha aprendido el algoritmo, o ha encontrado un mínimo local que resulta generalizar?

MARCO DE VERIFICACIÓN

- **Esparsificación:** Podar a estructura exacta de rango 7 (Strassen usa 7 multiplicaciones vs. 8 ingenuas).
- **Discretización:** Proyectar pesos a coeficientes enteros en $\{-1, 0, 1\}$.
- **Transferencia Cero Disparos:** Verificar corrección de 2×2 a 64×64 sin reentrenamiento.

Solo las soluciones que pasan los tres pasos representan cristalización algorítmica genuina.



Evolución de representaciones durante el entrenamiento. El marco de verificación distingue el aprendizaje estructural verdadero de la mera correlación estadística.

El Protocolo de Dos Fases para Inducción Algorítmica

1 Entrenamiento (Inducción)

- **Tamaño de Lote:** [24, 128] (Crítico para geometría)
- **Optimizador:** AdamW con decaimiento $\geq 1e-4$
- **Duración:** 1000+ épocas (Hasta grokking)
- **Objetivo:** Alcanzar cuenca de atracción

2 Verificación (Cristalización)

- **Esparsificación:** Podar a exactamente 7 ranuras
- **Discretización:** Redondear pesos a $\{-1, 0, 1\}$
- **Validación:** Verificar multiplicación de matrices

PERSPECTIVA CRÍTICA

La Fase 1 sola logró 0% de éxito. La esparsificación explícita (Fase 2) es requerida para cristalizar la solución aproximada en el algoritmo exacto.

$$\begin{aligned} p1 &= a(f - h) \\ p3 &= (c + d)e \\ p5 &= (a + d)(e + h) \\ p7 &= (a - c)(e + f) \end{aligned}$$

$$\begin{aligned} p2 &= (a + b)h \\ p4 &= d(g - e) \\ p6 &= (b - d)(g + h) \end{aligned}$$

The $A \times B$ can be calculated using above seven multiplications.
Following are values of four sub-matrices of result C

$$\left[\begin{array}{c|c} a & b \\ \hline c & d \end{array} \right] X \left[\begin{array}{c|c} e & f \\ \hline g & h \end{array} \right] = \left[\begin{array}{c|c} p5 + p4 - p2 + p6 & p1 + p2 \\ \hline p3 + p4 & p1 + p5 - p3 - p7 \end{array} \right]$$

X , Y and C are square metrices of size $N \times N$
 a, b, c and d are submatrices of A , of size $N/2 \times N/2$
 e, f, g and h are submatrices of B , of size $N/2 \times N/2$
 $p1, p2, p3, p4, p5, p6$ and $p7$ are submatrices of size $N/2 \times N/2$

Característica	Tipo
Restricción Rango-7	Ingenierizado
Coeficientes Enteros	Ingenierizado
Transferencia Cero	Emergente
Valores Discretos	Mixto

La Investigación del Tamaño de Lote

01 OBSERVACIÓN

Tamaños de lote en [24, 128] logran 68% de éxito. Otros valores fallan mayoritariamente.

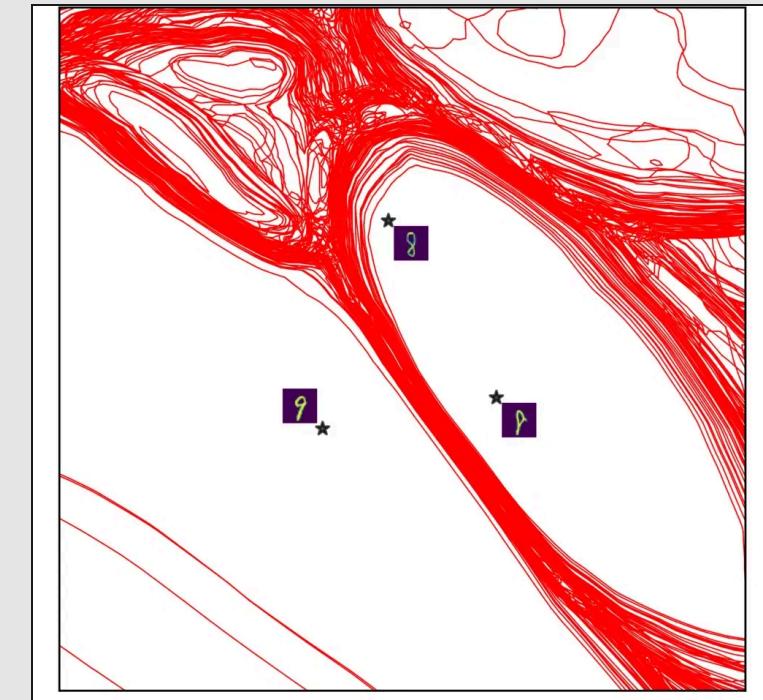
02 HIPÓTESIS (RECHAZADA)

¿Efectos de caché de hardware?

Evidencia en Contra: El análisis de memoria muestra que incluso $B=1024$ cabe cómodamente en caché L3 (321 KB vs $\geq 1\text{MB}$).

03 MECANISMO VERDADERO

Geometría de Covarianza del Gradiente. El tamaño de lote óptimo estabiliza el número de condición, creando trayectorias donde el ruido equilibra exploración y estabilidad.



Demostración del Paradigma: Soluciones en $B=32$ y $B=512$ pueden tener pérdida idéntica, pero solo geometrías de trayectoria específicas alcanzan la cuenca estrecha que contiene el algoritmo.

Fragilidad Extrema Revela Cuencas Estrechas

Las soluciones algorítmicas ocupan **cuencas extremadamente estrechas** en el espacio de pesos.

Los pesos deben converger precisamente a valores enteros. Incluso desviaciones microscópicas impiden la cristalización estructural.

IMPLICACIONES MÁS AMPLIAS

Fallos de reproducibilidad pueden reflejar cuencas perdidas, no fallos arquitectónicos.

Se requiere "Ingeniería de Trayectorias" para dirigir la dinámica hacia estas regiones estrechas.

El ajuste estándar de hiperparámetros es insuficiente sin control geométrico.

Prueba de Estabilidad al Ruido

RUIDO (Σ)	TASA DE ÉXITO
0.001	0%
0.005	0%
0.010	0%
0.050	0%

Umbral Crítico

$\sigma < 0.001$

Validación Estadística y Robustez Empírica

El análisis de 195 ejecuciones confirma que el tamaño de lote es un parámetro de control dominante para la cristalización algorítmica.

La Transferencia de Cero Disparos Valida el Aprendizaje Algorítmico Genuino

TAMAÑO MATRIZ	ERROR MÁXIMO	ESTADO
2 × 2	2.38e-07	PASA
4 × 4	1.91e-06	PASA
8 × 8	6.20e-06	PASA
16 × 16	2.15e-05	PASA
32 × 32	8.13e-05	PASA
64 × 64	2.94e-04	PASA

POR QUÉ ESTO IMPORTA

La transferencia de cero disparos es la **prueba definitiva**. Una red que simplemente memorizó patrones o encontró un mínimo local conveniente fallaría catastróficamente al expandirse.

CONTRASTE CON FALLOS

El 32% de las ejecuciones que fallan logran baja pérdida en 2×2 pero **no pueden transferir**. Aprendieron mínimos locales, no el algoritmo Strassen.

Implicaciones para la Reproducibilidad

Reformulando la reproducibilidad como un problema de ingeniería de trayectorias.

EL DESAFÍO

Muchos resultados son difíciles de reproducir incluso con hiperparámetros idénticos.

Las explicaciones estándar como detalles de implementación o variaciones de datos son a menudo **insuficientes** para explicar por qué una ejecución tiene éxito y otra falla.

LA PERSPECTIVA

La **Geometría de Trayectoria** es el factor faltante.

La inicialización aleatoria o el ruido de hardware conducen a diferentes trayectorias. Si la solución objetivo está en una **cuenca estrecha**, una trayectoria puede alcanzarla mientras otra se asienta en un mínimo local cercano.

EL CAMBIO

Debemos pasar del "Ajuste de Hiperparámetros" a la **Ingeniería de Trayectorias**.

Es necesario entender qué parámetros controlan la geometría y cómo **dirigir las trayectorias** hacia las cuencas objetivo.

Conclusión

Construcción Activa Sobre Emergencia Pasiva

CAMBIO DE PARADIGMA CENTRAL

VISTA TRADICIONAL

Entrenar redes, analizar soluciones finales, reportar propiedades emergentes.



NUEVA VISTA

Ingenierizar trayectorias, **verificar** estructura, entender **por qué** tienen éxito.

CONTRIBUCIONES CLAVE

- 02 **Marco de Verificación:** Distingue aprendizaje genuino de mínimos locales.
- 04 **Evidencia Empírica:** 68% de éxito bajo condiciones controladas.
- 06 **Perspectiva Mecanística:** Efectos del tamaño de lote reflejan geometría de covarianza.
- 08 **Fragilidad:** 0% de éxito con ruido $\sigma \geq 0.001$ revela cuencas estrechas.
- 010 **Reproducibilidad:** Nuevo marco basado en ingeniería de trayectorias.

La geometría de la trayectoria determina si las redes alcanzan soluciones algorítmicas. La tasa de fallo del 32% refleja la estrechez fundamental de las cuencas.

Referencias y Reproducibilidad

REFERENCIAS CLAVE

Deep Networks Always Grok and Here is Why

Humayun et al. (2024) • arXiv:2402.15555

Superposition as Lossy Compression

Bereska et al. (2024) • arXiv 2024

Algorithmic Induction via Structural Weight Transfer

grisun0 (2025) • Zenodo DOI: 10.5281/zenodo.18263654

REPRODUCCIÓN

[github . com/grisuno/strass_strassen](https://github.com/grisuno/strass_strassen)

```
git clone https://github.com/grisuno/strass_strassen
cd strass_strassen
pip install -r requirements.txt
python app.py
```

CRITERIOS DE ÉXITO (CHECKPOINTS)

- ✓ $\delta = 0$ (Pesos exactamente enteros en $\{-1, 0, 1\}$)
- ✓ Error Máx $< 1e-5$ (Multiplicación correcta)
- ✓ $S(\theta) = 1$ (Cristalización exitosa)