Grokking e Transições de Fase: Uma abordagem pedagógica da Física Estatística do Aprendizado de Máquina

João Pedro Sansão

https://doi.org/10.1590/SciELOPreprints.16570

Submetido em: 2026-06-16
Postado em: 2026-09-09 (versão 1)
(AAAA-MM-DD)

A moderação deste preprint recebeu o(s) endosso(s) de:

- Leonardo Araujo (ORCID: https://orcid.org/0000-0003-3884-2177)

Grokking e Transições de Fase: Uma abordagem pedagógica da Física Estatística do Aprendizado de Máquina

João Pedro Hallack Sansão*
Departamento de Tecnologia em Engenharia Civil, Computação, Automação, Telemática e Humanidades (DTECH)
Universidade Federal de São João del-Rei
ORCID: https://orcid.org/0000-0003-0095-2629

16 de junho de 2026

Resumo

O fenômeno de *grokking* em redes neurais artificiais descreve a situação contra-intuitiva em que um modelo, após sofrer sobreajuste (*overfitting*) completo aos dados de treino por um longo período, repentinamente atinge generalização perfeita em dados inéditos de teste. Neste trabalho, apresentamos uma revisão pedagógica deste fenômeno sob a perspectiva da Física Estatística, direcionada ao ensino de graduação em física e áreas correlatas. Demonstramos como o processo de aprendizado pode ser mapeado como uma transição de fase dinâmica de primeira ordem, onde a acurácia de validação desempenha o papel de parâmetro de ordem, o decaimento de pesos (*weight decay*) atua como um potencial harmônico restritivo, e a inicialização aleatória dos pesos combinada com a dinâmica caótica/estocástica intrínseca do próprio otimizador sob taxas de aprendizado finitas emulam uma temperatura efetiva cuja inversão favorece estados ordenados de baixa energia. Para tornar os conceitos concretos, fornecemos códigos reprodutíveis em Python (Jupyter Notebooks, disponíveis em `https://github.com/jsansao/grokking`) para o treinamento e análise de redes Perceptron Multicamadas (MLP) e *Transformers* aplicados a três experimentos distintos: aritmética modular, o problema da paridade esparsa e a conservação de energia modular. Mostramos que a generalização emerge da descoberta de simetrias ocultas no conjunto de dados, manifestada pelo alinhamento periódico dos pesos na forma de representações trigonométricas circulares (analisadas via Transformada de Fourier) e pela organização espacial concêntrica de estados degenerados de energia no espaço de *embeddings*.
Palavras-chave: *Grokking*, Transição de Fase, Física Estatística, Redes Neurais, Aprendizado de Máquina.

Abstract

Grokking and Phase Transitions: A Pedagogical Approach to the Statistical Physics of Machine Learning

The phenomenon of *grokking* in artificial neural networks describes the counter-intuitive situation in which a model, after completely overfitting the training data for a long period, suddenly achieves perfect generalization on unseen test data. In this work, we present a pedagogical review of this phenomenon from the perspective of Statistical Physics, aimed at undergraduate education in physics and related areas. We demonstrate how the learning process can be mapped as a first-order dynamic phase transition, where the validation accuracy plays the role of an order parameter, weight decay acts as a restrictive harmonic potential, and the random weight initialization combined with the chaotic/stochastic dynamic of the optimizer under finite learning rates emulate an effective temperature whose inversion favors low-energy ordered states. To make the concepts concrete, we provide reproducible Python codes (Jupyter Notebooks, available at `https://github.com/jsansao/grokking`)

*joao@ufsj.edu.br, ORCID: https://orcid.org/0000-0003-0095-2629

for training and analyzing Multilayer Perceptron (MLP) and *Transformer* networks applied to three distinct experiments: modular arithmetic, the sparse parity problem, and modular energy conservation. We show that generalization emerges from discovering hidden symmetries in the dataset, manifested by the periodic alignment of weights in the form of circular trigonometric representations (analyzed via Fourier Transform) and by the concentric spatial organization of degenerate energy states in the *embedding* space.
Keywords: *Grokking*, Phase Transition, Statistical Physics, Neural Networks, Machine Learning.

1 Introdução

Nas últimas duas décadas, o aprendizado profundo (*deep learning*) revolucionou a ciência e a tecnologia, impulsionando avanços que vão do processamento de linguagem natural à previsão da estrutura tridimensional de proteínas. Contudo, do ponto de vista conceitual, muitas das propriedades de generalização de redes neurais profundas permanecem como mistérios teóricos em aberto.

Na modelagem clássica de aprendizado estatístico, espera-se que um modelo altamente superparametrizado (com muito mais parâmetros ajustáveis do que dados de treinamento) sofra de sobreajuste (*overfitting*). Ou seja, o modelo deve memorizar o ruído do conjunto de treino e falhar ao fazer previsões sobre novos dados. O comportamento típico da perda (*loss*) de validação é o de decair inicialmente e, em seguida, divergir, exigindo técnicas como a parada precoce (*early stopping*) para evitar a perda de generalização.

Em 2022, contudo, Power et al. [15] descobriram um comportamento anômalo ao treinar redes neurais em tarefas algorítmicas simples (como a aritmética modular). Eles observaram que, se o treinamento continuasse por milhares de épocas após a acurácia de treino ter atingido 100% e a perda de treino ter ido a zero, a acurácia de validação repentinamente saltava de zero para 100% de forma abrupta. Esse fenômeno foi denominado *grokking* (termo retirado da ficção científica que significa "compreender intuitiva e profundamente").

Para a física, esse salto abrupto de um estado de total erro (fase de memorização) para um estado de perfeita precisão (fase de generalização) é extremamente semelhante a uma transição de fase de primeira ordem, como a solidificação da água ou a magnetização espontânea em sistemas de *spins*.

Este trabalho tem caráter pedagógico e visa introduzir o conceito de *grokking* a estudantes de física e áreas correlatas, conectando-o com a Física Estatística e a Análise de Fourier. O texto está estruturado da seguinte forma: na seção 2 discutimos a modelagem física do *grokking*; na seção 3 descrevemos a tarefa e os modelos de simulação (MLP e *Transformer*); na seção 4 apresentamos as curvas de aprendizado e a interpretação geométrica dos pesos; e, por fim, na seção 5 tecemos nossas considerações finais.

2 A Física do *Grokking*

A intersecção entre a Física Estatística e o Aprendizado de Máquina possui uma história consolidada, remontando à introdução do modelo de Hopfield [1] na década de 1980 e aos trabalhos seminais de Daniel Amit, Haim Sompolinsky e H. Sebastian Seung [2], que estabeleceram as bases para mapear a capacidade de armazenamento de memória e transições de fase (como em fases vítreas de *spins*) em redes neurais. Essa sólida tradição justifica e enriquece a análise moderna realizada sobre o *grokking*.

Antes de formalizarmos o comportamento térmico do sistema, é oportuno introduzir as definições dos principais componentes de uma rede neural sob uma perspectiva física intuitiva. Para uma descrição didática complementar das redes neurais com enfoque em sistemas físicos, recomendamos o trabalho de Lima et al. [3]. O aprendizado de máquina e a física compartilham

profundas semelhanças matemáticas, permitindo mapear o treino de redes a sistemas mecânicos estatísticos clássicos:

2.1 Glossário Pedagógico e Analogias Físicas

1. **Rede Neural e Parâmetros (w):** Uma rede neural artificial é um aproximador universal de funções construído por camadas de transformações lineares intercaladas com ativações não lineares [8, 9]. Os pesos sinápticos e *biases*, agrupados no vetor w, representam os graus de liberdade ajustáveis do modelo, funcionando como as coordenadas generalizadas de um sistema físico complexo no espaço de parâmetros.

2. **Função de Perda ($L_{\mathbf{dados}}(w)$):** É a função custo que quantifica a discrepância entre as previsões do modelo e as respostas corretas [9]. Age de forma análoga a uma Energia Potencial de Interação com o conjunto de dados [10]. Minimizar a perda empírica equivale a guiar os pesos para as bacias de mínima energia potencial de dados.

3. **Decaimento de Pesos (*Weight Decay*, λ):** Termo de regularização adicionado para restringir a magnitude dos pesos sinápticos [11]. Matematicamente, a penalidade dada por $V_{\text{reg}}(w) = \frac{\lambda}{2}\|w\|^2$ é idêntica a um Potencial Harmônico Restritivo com constante de mola $\lambda > 0$, que exerce uma força restauradora puxando os pesos em direção à origem do espaço de parâmetros.

4. **Otimização e Dinâmica de Langevin:** O treinamento atualiza os pesos iterativamente por gradientes. Essa dinâmica pode ser mapeada como a Equação de Langevin para uma partícula browniana sob atrito e ruído térmico [12]:

$$\frac{dw}{dt} = -\gamma \nabla E(w) + \eta(t) \tag{1}$$

5. **Embeddings (Incorporações Contínuas):** Os operandos inteiros de 0 a $p-1$ são elementos simbólicos discretos. O *embedding* mapeia cada símbolo a um vetor contínuo no espaço euclidiano de dimensão d_{embed} [13]. Fisicamente, corresponde a definir as coordenadas de posição contínuas $\vec{r}_i$ de partículas discretas em um espaço multidimensional.

6. **Mecanismo de Atenção:** Componente do *Transformer* que pondera a relevância mútua entre diferentes partes da sequência [14]. Funciona como um acoplamento dinâmico entre as posições de entrada (como um sistema de partículas acopladas cujos coeficientes de mola mudam no tempo dependendo do estado das próprias partículas), em contraste com os acoplamentos estáticos de redes MLP tradicionais.

2.2 Mapeamento Termodinâmico e Transição de Fase

A dinâmica de otimização de uma rede neural compartilha profundas semelhanças com sistemas físicos desordenados e frustrados, como os vidros de *spin* (por exemplo, o modelo Sherrington-Kirkpatrick [4] ou o modelo de Ising com acoplamentos aleatórios) [5]. Nesses sistemas, a competição entre interações locais e a presença de desordem geram uma paisagem de energia livre extremamente rugosa, repleta de mínimos locais metaestáveis e barreiras energéticas. Essa analogia é particularmente forte no problema da paridade esparsa (*sparse parity*), onde o rótulo é a paridade (multiplicação) de apenas k *spins* acoplados em um vetor de tamanho N *spins* (com $k \ll N$). A busca por esses k acoplamentos corretos em meio a $N-k$ *spins* de ruído puro é equivalente a encontrar o estado fundamental de um vidro de *spin* com interações multi-*spins* de ordem k (como no modelo de vidro de *spin p-spin* [5]) sob forte frustração. Como o gradiente inicial não fornece informações direcionais no espaço de parâmetros para filtrar o ruído, a dinâmica de aprendizado de máquina fica inicialmente retida em uma bacia metaestável

rugosa de entropia alta, análoga a uma fase vítrea ou desordenada. No contexto do aprendizado de máquina, os pesos sinápticos w representam as coordenadas do sistema nessa paisagem multidimensional frustrada.

Para formalizar analiticamente essa conexão, podemos definir a energia total do sistema da rede neural como a soma da perda empírica com o termo de regularização harmônica:

$$E(w) = L_{\text{dados}}(w) + \frac{\lambda}{2}\|w\|^2 \tag{2}$$

Fisicamente, a probabilidade de o sistema ocupar uma determinada configuração de pesos w sob a influência de flutuações estocásticas — decorrentes primariamente da inicialização aleatória dos pesos e da dinâmica caótica/estocástica intrínseca do próprio algoritmo AdamW sob taxas de aprendizado finitas, emulando um ruído térmico efetivo na paisagem de perda — é descrita pela distribuição clássica de Boltzmann-Gibbs [6]:

$$P(w) = \frac{1}{Z}e^{-\beta E(w)} = \frac{1}{Z}e^{-\beta\left(L_{\text{dados}}(w)+\frac{\lambda}{2}\|w\|^2\right)} \tag{3}$$

onde $\beta = 1/T$ representa a temperatura inversa efetiva associada à magnitude do ruído de otimização e Z é a função de partição do sistema:

$$Z = \int dw\, e^{-\beta E(w)} \tag{4}$$

A partir da função de partição, a energia livre de Helmholtz F do sistema pode ser expressa como:

$$F = -\frac{1}{\beta}\ln Z = \langle E\rangle - TS \tag{5}$$

onde $\langle E\rangle = \int dw\, P(w)E(w)$ é a energia média do sistema e $S = -\int dw\, P(w)\ln P(w)$ é a entropia de Shannon, que quantifica o volume ocupado pelas configurações de pesos no espaço de parâmetros [6].

A minimização da energia livre de Helmholtz F revela a competição termodinâmica fundamental que explica o *grokking*. Em épocas iniciais de treinamento, correspondentes a um regime de alta temperatura efetiva (T alto ou β baixo), o termo entrópico $-TS$ domina a energia livre F. Como o modelo é superparametrizado, o espaço de configurações que memoriza o pequeno conjunto de treino (com $L_{\text{dados}}(w) \approx 0$) possui um volume imenso. Esse estado desordenado e amorfo representa a fase de memorização (*overfitting*). Embora essas soluções de memorização requeiram pesos com normas elevadas ($\|w\|^2$ grande, e portanto alta energia $E(w)$), elas dominam a dinâmica inicial devido à sua imensa entropia (dominância de volume). Nesse estágio, a acurácia de validação permanece nula ($A_{\text{val}} \approx 0$).

Por outro lado, o estado de generalização corresponde a configurações de pesos altamente estruturadas e ordenadas (como os circuitos trigonométricos que implementam a simetria modular). Essas soluções generalizadoras são extremamente organizadas e compactas, possuindo normas de pesos muito menores ($\|w\|^2$ pequeno, e portanto baixa energia $E(w)$). No entanto, o volume de parâmetros que implementa essa simetria exata é minúsculo (baixa entropia S). À medida que o treinamento prossegue sob a ação do decaimento de pesos λ, a energia dessas soluções cai drasticamente. Esse processo é termodinamicamente equivalente a um resfriamento lento (ou aumento de β). Em baixas temperaturas (T baixo ou β alto), o termo de energia média $\langle E\rangle$ passa a dominar a energia livre F. O estado de generalização torna-se o verdadeiro mínimo global da energia livre. O salto repentino da validação de 0 a 100% corresponde, portanto, a uma transição de fase dinâmica de primeira ordem induzida termodinamicamente [7, 17, 18], onde a acurácia de validação A_{val} atua como o parâmetro de ordem que distingue a fase desordenada (memorização) da fase ordenada (generalização).

2.3 Simetria e Circuitos Trigonométricos

Por que a solução generalizadora possui menor norma de pesos? Considere a tarefa de aritmética modular: dado dois inteiros $a, b \in \{0, 1, \ldots, p-1\}$, queremos calcular:

$$c = (a + b) \pmod{p} \tag{6}$$

Essa tarefa exibe uma simetria cíclica discreta C_p (invariância por translação no espaço modular). Para generalizar para pares (a,b) não vistos no treino, a rede neural precisa de alguma forma implementar essa simetria.

Pesquisas de interpretabilidade mecanicista [16] revelaram que a rede neural descobre essa simetria projetando os índices discretos a e b em círculos bidimensionais de frequências espaciais específicas k:

$$E(x) \propto \left[\cos\left(\frac{2\pi kx}{p}\right), \sin\left(\frac{2\pi kx}{p}\right)\right]^T \tag{7}$$

onde $E(x)$ são os vetores de *embedding* da rede. Para realizar a soma modular, a rede utiliza a identidade trigonométrica da soma de arcos nas camadas subsequentes:

$$\cos\left(\frac{2\pi k(a+b)}{p}\right) = \cos(x_a)\cos(x_b) - \sin(x_a)\sin(x_b) \tag{8}$$

$$\sin\left(\frac{2\pi k(a+b)}{p}\right) = \sin(x_a)\cos(x_b) + \cos(x_a)\sin(x_b) \tag{9}$$

onde $x_a = 2\pi ka/p$ e $x_b = 2\pi kb/p$. Ao aprender essa representação harmônica, a rede reduz drasticamente a complexidade do mapeamento, necessitando de menos energia (menor norma de pesos $\|w\|^2$) do que se utilizasse uma tabela de memorização aleatória de alta frequência.

3 Metodologia Experimental

Para demonstrar o fenômeno de forma reprodutível, propomos três experimentos computacionais usando PyTorch, detalhados nos Jupyter Notebooks disponíveis em `https://github.com/jsansao/grokking`. O treinamento de todos os modelos é rápido, durando poucos minutos em CPU convencional. A divisão do conjunto de dados em treino e validação é configurada para simular um cenário de dados esparsos, induzindo a barreira de generalização característica do *grokking*. Essa restrição extrema no conjunto de treino (ocultando a maior parte dos dados do domínio) é uma condição fundamental: se a rede tivesse acesso a quase todos os pares possíveis, ela aprenderia a regra geral quase de imediato por interpolação suave. Ao limitar drasticamente as amostras vistas, força-se o modelo a memorizá-los inicialmente por meio do sobreajuste, estabelecendo um platô metaestável de baixa perda local antes de transicionar tardia e descontinuamente para a descoberta das simetrias matemáticas subjacentes.

3.1 Experimento 1: Aritmética Modular ($a + b \pmod{p}$)

O primeiro experimento investiga a tarefa clássica introduzida por Power et al. [15]. O modelo recebe dois operandos inteiros a e b no intervalo $[0, p-1]$ (onde $p = 67$ é um número primo) e deve prever a soma modular $a + b \pmod{p}$. O conjunto de dados total possui $N = p^2 = 4489$ pares. A divisão de dados aloca uma fração de treino de 35% (1571 pares) e 65% para validação.

Neste experimento, avaliamos duas arquiteturas distintas:

1. **Perceptron Multicamadas (MLP)**: Cada inteiro é mapeado para um *embedding* de dimensão $d_{\text{embed}} = 64$. Os *embeddings* dos operandos são concatenados em um vetor de tamanho 128 e processados por uma camada oculta linear com dimensão $d_{\text{hidden}} = 128$ e ativação GELU, seguida de projeção linear de saída de dimensão p. O modelo é treinado

via lote completo (*full batch*) com AdamW, $\eta_{\text{LR}} = 10^{-3}$ e regularização de pesos L_2 com coeficiente $\lambda = 1{,}0$.

2. ***Transformer* Causal**: Um *Transformer decoder-only* (uma variante simplificada da arquitetura *Transformer* original focada apenas em decodificação sequencial) de 1 camada. A entrada é uma sequência de três *tokens* $[a, b, =]$, onde "=" corresponde ao índice p. A arquitetura utiliza *embeddings* de dimensão $d_{\text{embed}} = 64$ somados a *embeddings* posicionais aprendidos, uma camada de auto-atenção causal com 1 cabeça. A propriedade de causalidade impõe uma restrição física/temporal ao modelo por meio de uma máscara de atenção, garantindo que cada elemento da sequência processe apenas informações de *tokens* anteriores a ele. Dessa forma, ao computar a representação do *token* "=", o modelo tem acesso estrito a a e b, impedindo que informações futuras influenciem o estado presente na modelagem. O modelo utiliza ainda uma camada MLP residual de dimensão $d_{\text{hidden}} = 128$, e uma projeção final na posição de "=" para prever as probabilidades das p classes. O modelo é treinado via lote completo com AdamW, $\eta_{\text{LR}} = 10^{-3}$ e decaimento de pesos $\lambda = 2{,}0$.

Matematicamente, a função de ativação GELU (Gaussian Error Linear Unit) é definida por:

$$\text{GELU}(x) = x\Phi(x) = x \cdot \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right] \tag{10}$$

onde $\Phi(x)$ é a função de distribuição cumulativa da distribuição normal padrão. Ao contrário da ReLU clássica, a suavidade da GELU elimina descontinuidades no gradiente e favorece o escape de mínimos locais durante a transição de fase.

3.2 Experimento 2: O Problema da Paridade Esparsa (*Sparse Parity*)

O segundo experimento analisa o comportamento do modelo em uma tarefa puramente combinatória com ruído sistemático, afastando-se das simetrias da aritmética algébrica. A entrada consiste em vetores de *spins* clássicos de tamanho $N = 10$ com valores binários $x_i \in \{-1, 1\}$. O rótulo correspondente $y \in \{0, 1\}$ é determinado pelo produto (paridade ou operação XOR) de apenas $k = 3$ bits fixados (neste caso, os três primeiros índices):

$$y = \frac{1}{2}\left(1 + x_0 \cdot x_1 \cdot x_2\right) \tag{11}$$

Os $N - k = 7$ bits restantes são inicializados aleatoriamente e agem como ruído não correlacionado. O espaço amostral completo possui $2^{10} = 1024$ configurações possíveis.

Avaliamos as seguintes arquiteturas sob essa tarefa:

1. **Perceptron Multicamadas (MLP)**: Um classificador com uma camada oculta linear de dimensão $d_{\text{hidden}} = 1000$ e ativação GELU, com *biases* ativos nas camadas lineares. Em redes neurais, o *bias* (ou viés) representa um parâmetro aditivo constante em uma transformação linear $y = Wx + b$. Geometricamente, ele atua de forma análoga a uma translação rígida da origem do sistema de coordenadas, permitindo deslocar as fronteiras de decisão de forma independente da orientação angular estabelecida pelos pesos W. A fração de treino é restrita a apenas 10% (100 amostras) e 90% para validação. O treinamento utiliza lote completo, AdamW com $\eta_{\text{LR}} = 10^{-3}$ e regularização de pesos $\lambda = 0{,}5$.

2. ***Transformer* Bidirecional**: Um *Transformer Encoder* de 1 camada. O vetor binário x é mapeado para *tokens* discretos $x_i \in \{0, 1\}$ correspondentes, processados como uma sequência de comprimento $N = 10$ com *embeddings* $d_{\text{embed}} = 128$, 4 cabeças de atenção e camada oculta MLP residual com $d_{\text{hidden}} = 256$. O modelo realiza auto-atenção bidirecional (sem máscara causal), seguido por *pooling* médio sobre as representações dos *tokens* no

tempo. Esta operação computa a média aritmética simples dos vetores latentes temporais ao longo da sequência, projetando geometricamente a trajetória de *embeddings* em seu centro de massa de modo a produzir um único vetor estático para alimentar a camada de classificação baseada em uma projeção linear de saída para as 2 classes. A divisão de dados aloca apenas 70 amostras para treinamento ($\approx 6{,}8\%$) e 954 para validação, com decaimento de pesos $\lambda = 0{,}5$.

3.3 Experimento 3: Conservação de Energia Modular ($x^2 + y^2 \pmod p$)

O terceiro experimento introduz uma tarefa com forte correspondência geométrica e física. O modelo recebe o par (x, y) de inteiros no intervalo $[0, p-1]$ (onde definimos $p = 59$ como base) e deve prever a soma de seus quadrados módulo p:

$$c = (x^2 + y^2) \pmod p \tag{12}$$

Na física clássica e ondulatória, a quantidade $x^2 + y^2 = R^2$ representa superfícies equipotenciais, órbitas circulares ou a conservação de energia e momento angular. No espaço discreto $\mathbb{Z}_p$, a operação conserva essa simetria de forma modular.

Como a operação depende apenas dos resíduos quadráticos $x^2 \pmod p$, elementos x e $p - x$ mapeiam para o mesmo quadrado ($x^2 \equiv (p-x)^2 \pmod p$). Na física de estados quânticos, isso corresponde a estados degenerados de energia, onde múltiplos estados do sistema possuem o mesmo nível de energia. Para generalizar, a rede é forçada a mapear esses elementos degenerados para vetores idênticos (ou muito próximos) no espaço latente.

A arquitetura utilizada é a seguinte:

1. **Perceptron Multicamadas (MLP) com *Embeddings***: A mesma estrutura básica do Experimento 1 (MLP), com uma camada de *embedding* de tamanho $p \times 64$, concatenação dos operandos para dimensão 128, camada oculta linear com $d_{\text{hidden}} = 128$ e ativação GELU, e projeção de saída com dimensão p. Para esta arquitetura, dividimos os dados com fração de treino de 45% (1566 amostras) e validação de 55% (1915 amostras), sob o otimizador AdamW com $\eta_{\text{LR}} = 10^{-3}$ e decaimento de pesos $\lambda = 1{,}0$.

2. ***Transformer* Causal**: Um *Transformer decoder-only* de 1 camada que recebe a sequência de *tokens* $[x, y, =]$ de comprimento 3, realizando a predição na posição correspondente ao *token* "=". O modelo utiliza *embeddings* de dimensão $d_{\text{embed}} = 64$ para um vocabulário de tamanho $p + 1$ (operandos de 0 a $p - 1$ mais o *token* "="), 2 cabeças de atenção e camada MLP residual de dimensão intermediária $d_{\text{hidden}} = 128$. A divisão de dados aloca 40% (1392 amostras) para treino e 60% (2089 amostras) para validação, com treinamento por 4000 épocas via AdamW com $\eta_{\text{LR}} = 10^{-3}$ e decaimento de pesos $\lambda = 1{,}0$.

O conjunto de dados total possui $N = p^2 = 3481$ pares.

4 Resultados e Discussão

4.1 Curvas de Aprendizado

Os resultados das simulações computacionais demonstram claramente a assinatura clássica do *grokking*. Conforme mostrado nas simulações dos Notebooks e ilustrado nas Figuras 1 e 2 (para a aritmética modular), 3 e 4 (para a paridade esparsa), e 5 e 6 (para a conservação de energia modular):

Para a tarefa de aritmética modular, observamos comportamentos marcadamente distintos entre as duas arquiteturas. No MLP (Fig. 1), embora a acurácia de treino atinja 100% nos primeiros 500 passos, a acurácia de validação permanece em 0,00% por milhares de épocas, iniciando uma transição de fase contínua apenas após a época 5000 e subindo consistentemente

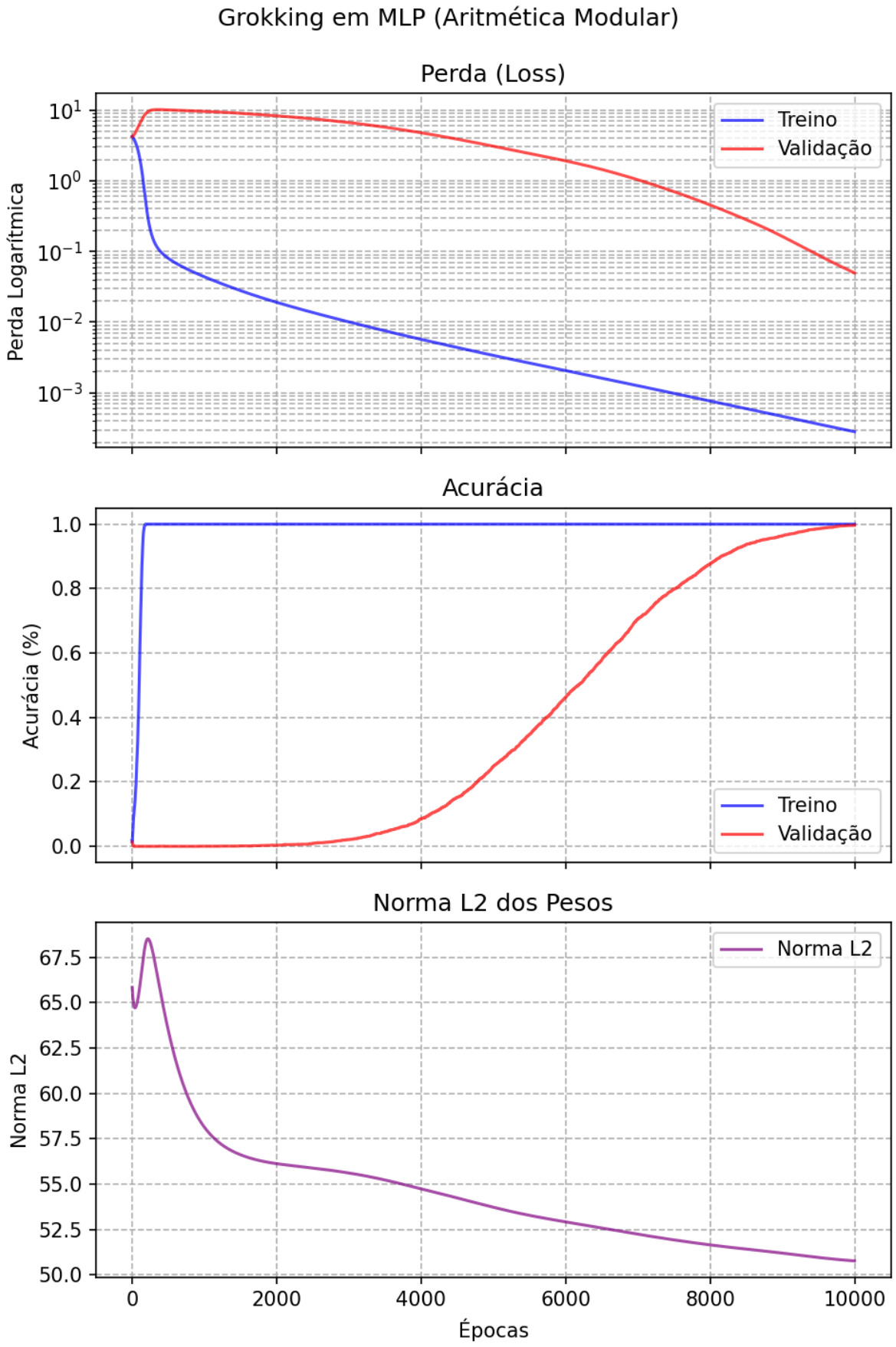

Figura 1: Curvas de perda (topo), acurácia (meio) e norma dos pesos L2 (base) durante o treinamento do modelo MLP com regularização por decaimento de pesos $\lambda = 1{,}0$. A transição de fase de generalização ocorre de forma contínua porém rápida entre as épocas 3000 e 8000.

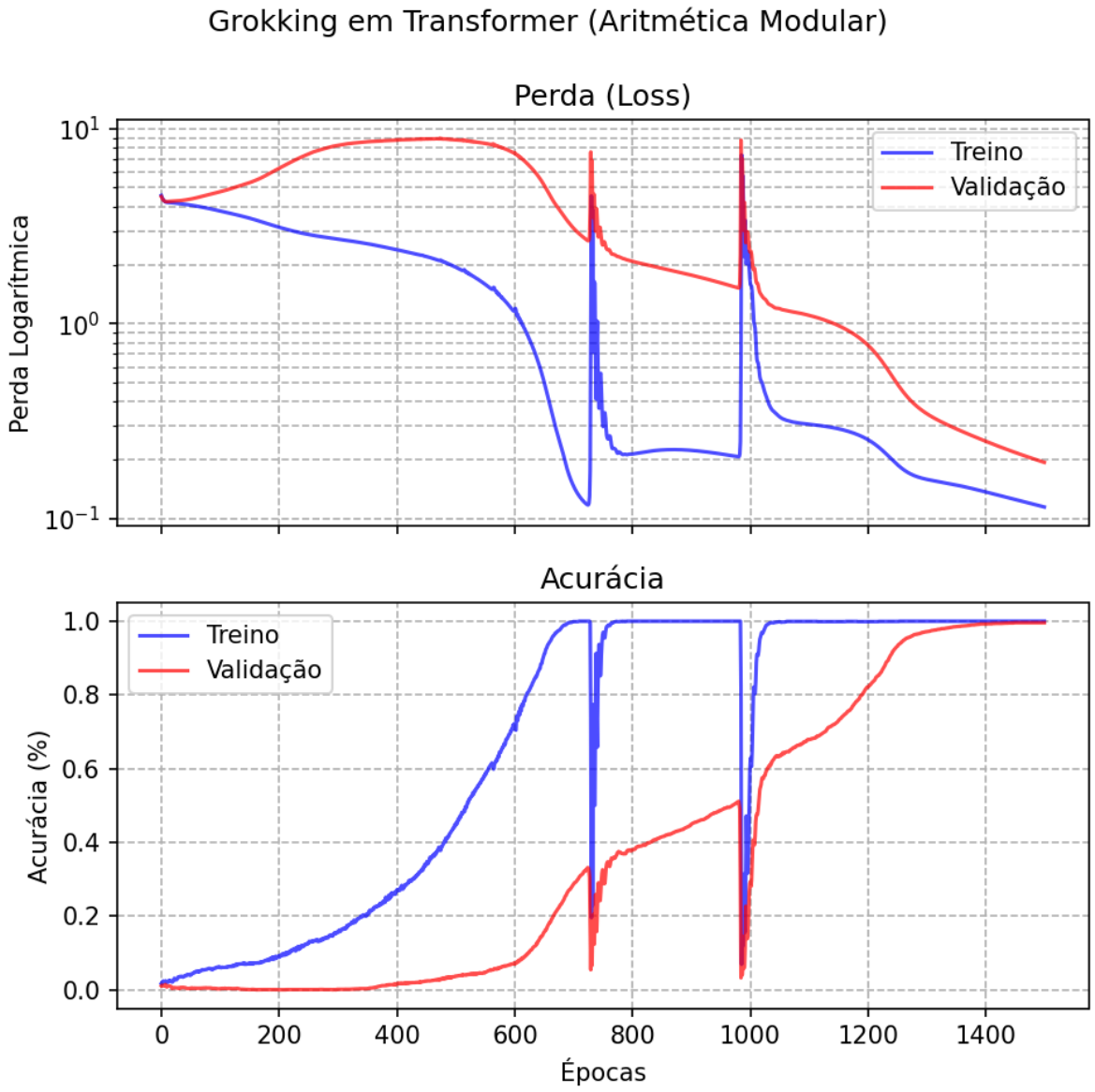

Figura 2: Curvas de perda (topo) e acurácia (base) durante o treinamento do mini-*Transformer* com regularização $\lambda = 2{,}0$. O modelo atinge a transição de fase de forma muito mais abrupta e rápida, por volta de 1500 épocas.

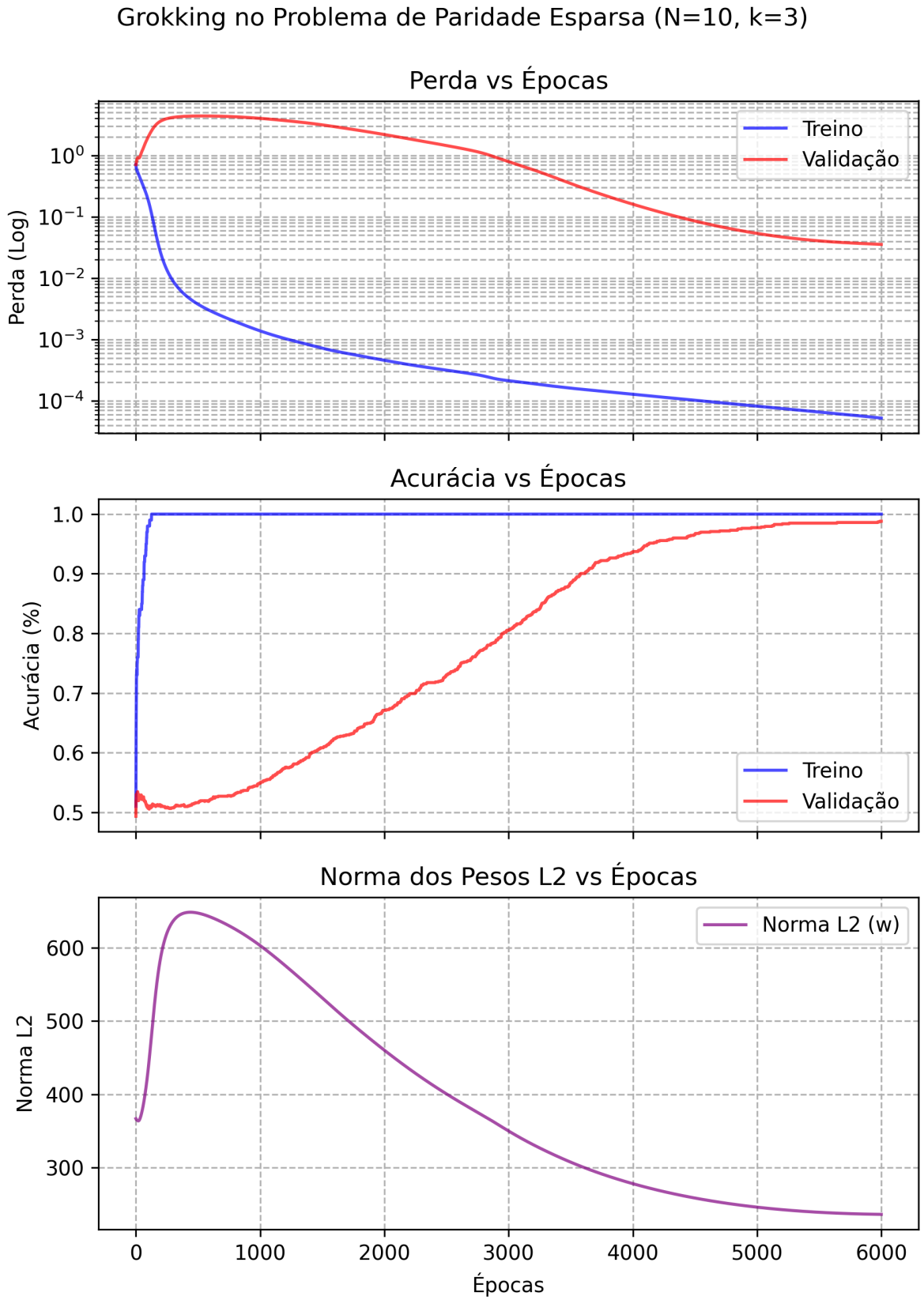

Figura 3: Curvas de perda (topo), acurácia (meio) e norma dos pesos L2 (base) durante o treinamento do Perceptron para o problema de paridade esparsa ($N = 10$, $k = 3$) com decaimento de pesos $\lambda = 0{,}5$. Observa-se um nítido platô metaestável de memorização onde a validação se mantem em 50%, seguido por um encolhimento acentuado na norma dos pesos (base) que culmina na transição descontínua para generalização perfeita.

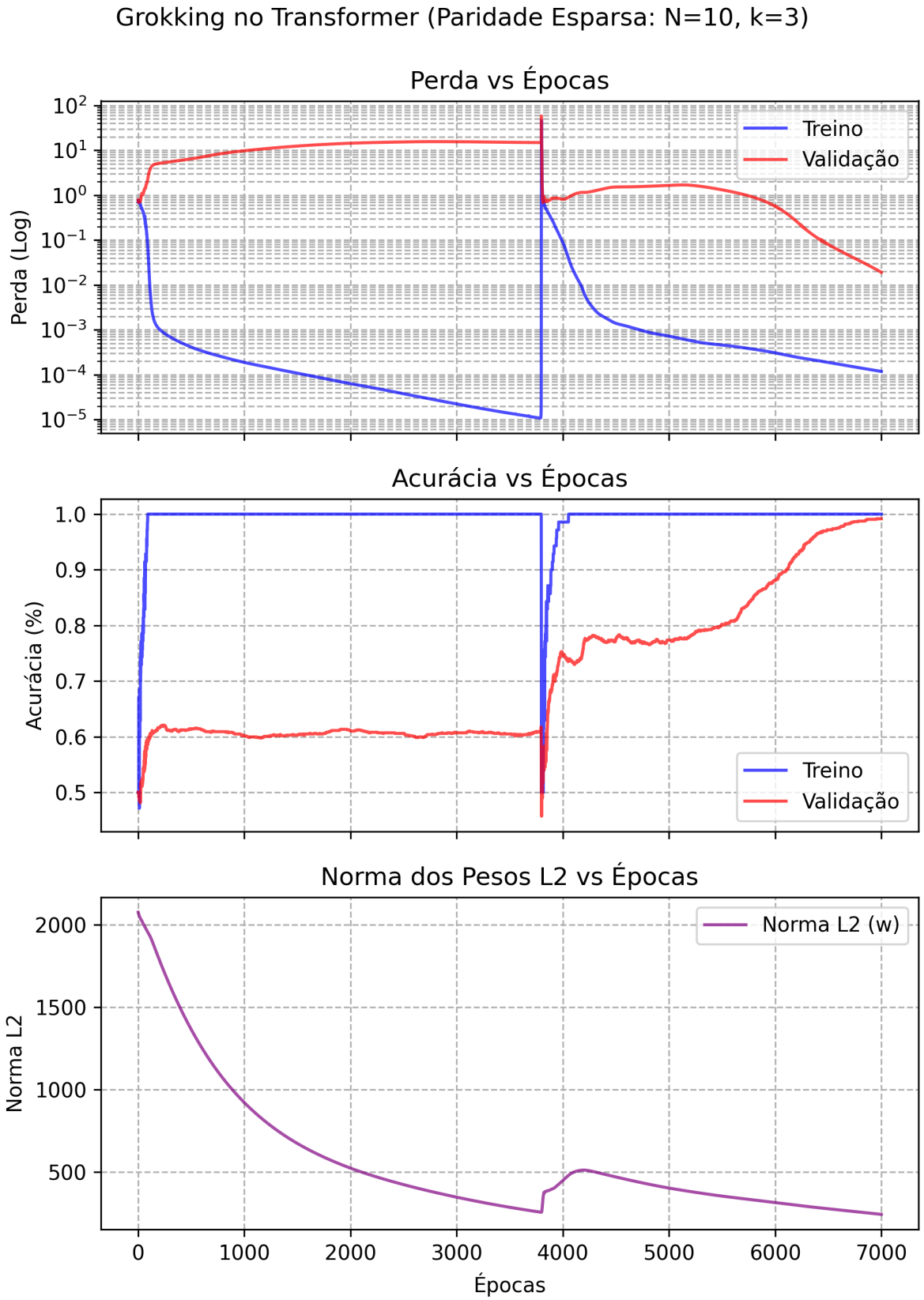

Figura 4: Curvas de perda (topo), acurácia (meio) e norma dos pesos L2 (base) durante o treinamento do *Transformer* para o problema de paridade esparsa ($N = 10$, $k = 3$) com decaimento de pesos $\lambda = 0{,}5$. Observa-se um platô metaestável de memorização onde a validação se mantém em torno de 61%, seguido pelo colapso de pesos (base) e a transição abrupta para generalização perfeita.

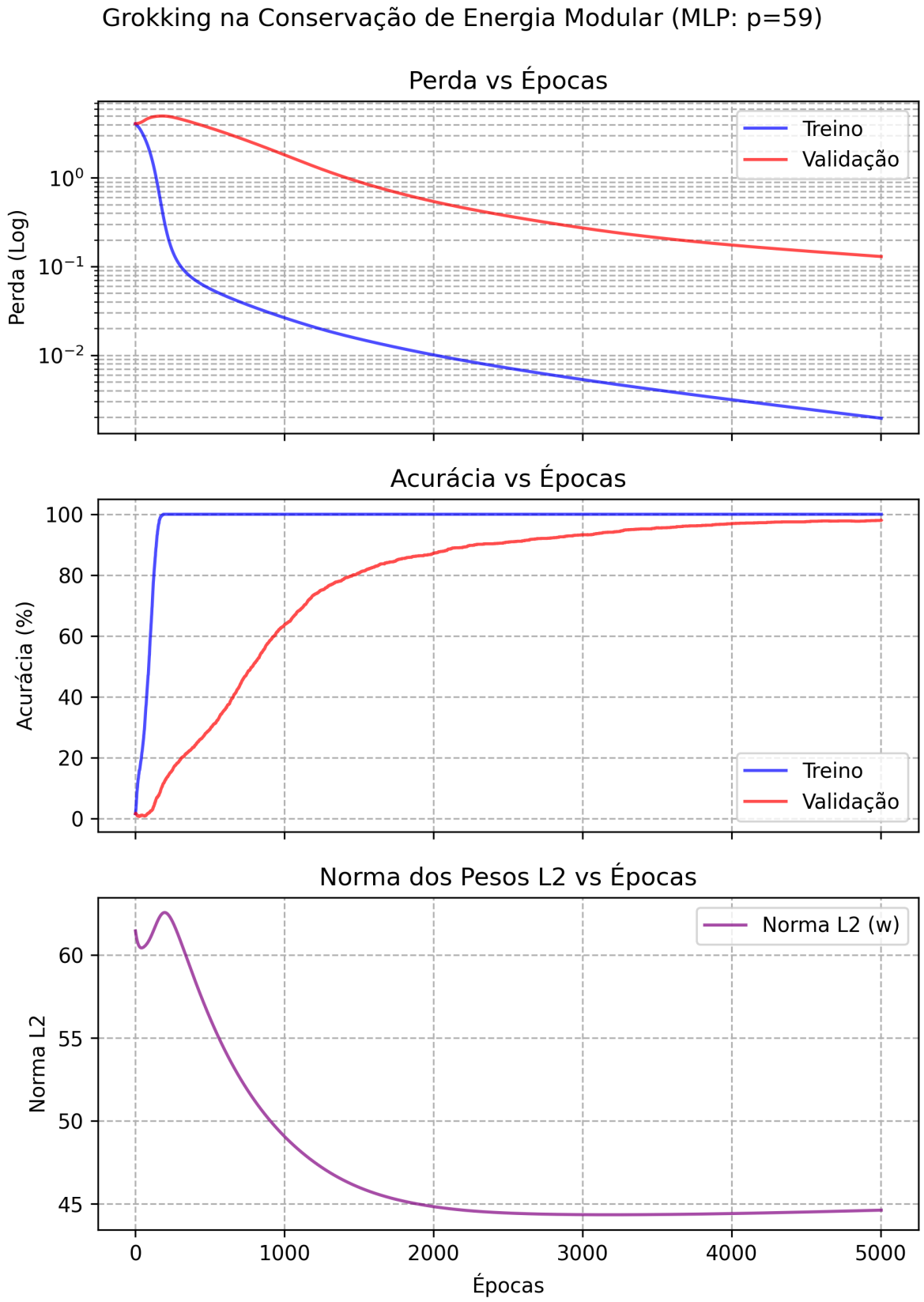

Figura 5: Curvas de perda (topo), acurácia (meio) e norma dos pesos L2 (base) para a tarefa de Conservação de Energia Modular ($x^2 + y^2 \pmod{59}$) com regularização $\lambda = 1{,}0$ e fração de treino de 45%. A rede exibe um platô metaestável com acurácia de validação estagnada em $\sim 63{,}6\%$ antes de transicionar abruptamente para generalização de 100% por volta da época 6000.

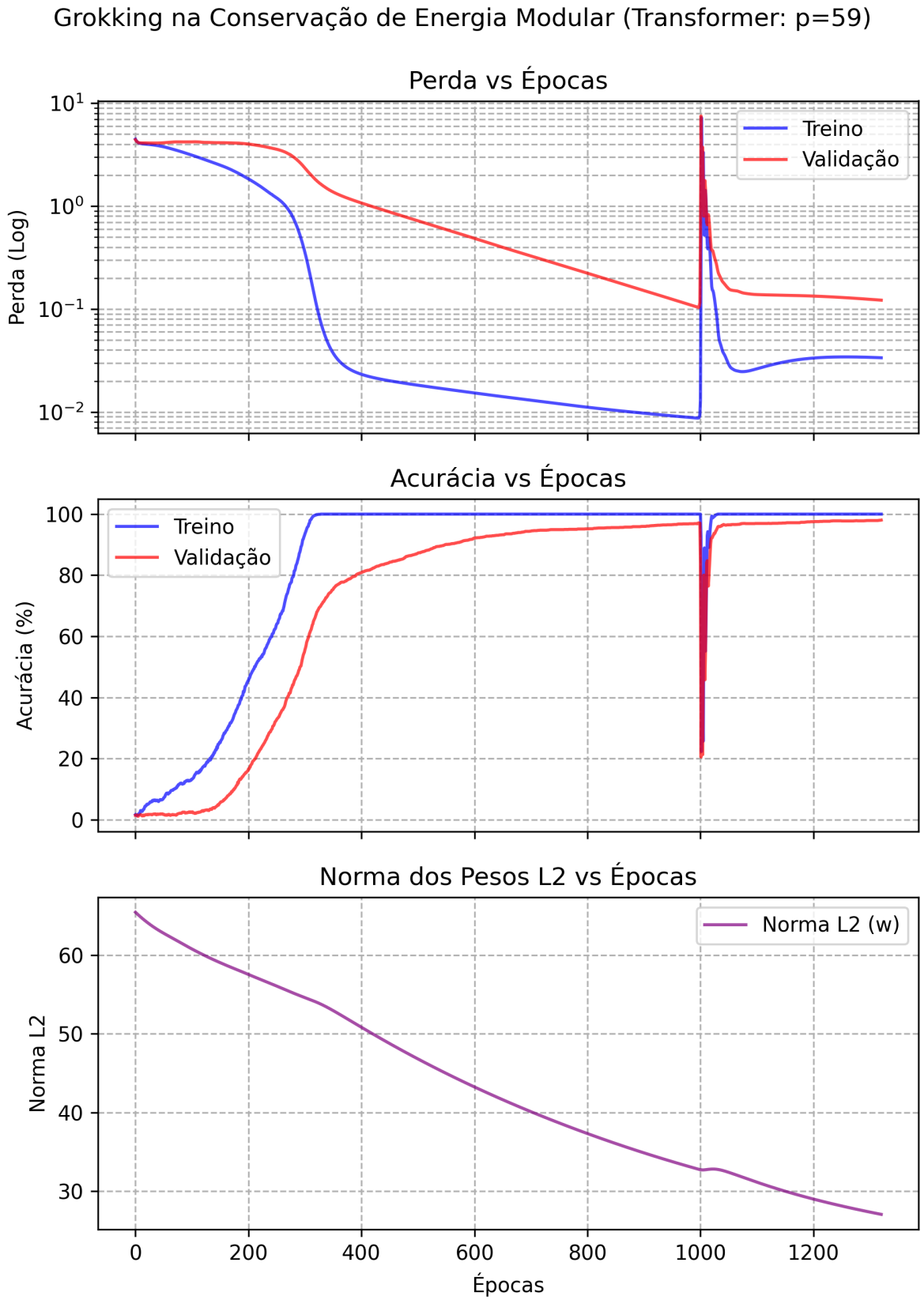

Figura 6: Curvas de perda (topo), acurácia (meio) e norma dos pesos L2 (base) para a tarefa de Conservação de Energia Modular ($x^2 + y^2 \pmod{59}$) usando o *Transformer* Causal com decaimento de pesos $\lambda = 1{,}0$ e fração de treino de 40%. Observa-se a clássica transição abrupta de generalização (*grokking*) por volta da época 2000, onde a acurácia de validação sobe de 63,6% para $> 98\%$.

até atingir 99,66% de validação na época 10000. Por outro lado, o mini-*Transformer* (Fig. 2) demonstra um aprendizado de treino inicial mais gradual, contudo a sua acurácia de validação exibe um salto descontínuo extremamente acentuado entre as épocas 1000 e 1500, alcançando a generalização perfeita de 99,59% na época 1500, o que caracteriza uma transição de fase dinâmica muito mais rápida.

No problema da paridade esparsa, sob um regime de dados fortemente limitados, a acurácia de treino de ambos os modelos converge rapidamente para 100% (antes da época 200). Entretanto, a acurácia de validação fica inicialmente retida em um platô metaestável de memorização desordenada, em torno de 51% para o MLP (Fig. 3) e 61% para o *Transformer* (Fig. 4). A transição para o estado generalizador ocorre a partir da época 1000 para o MLP (com a validação convergindo para 98,81% na época 6000) e após a época 3000 para o *Transformer* (atingindo 99,16% na época 7000), ambos os processos sendo acompanhados pelo encolhimento sistemático e acentuado da norma dos pesos sob a ação da penalidade harmônica.

Por fim, no experimento de conservação de energia modular, ambas as arquiteturas atingem acurácia de treino de 100% precocemente (antes da época 1000 para o MLP e na época 500 para o *Transformer*), com a acurácia de validação inicialmente estagnada no platô de memorização de $\approx 63{,}6\%$. Sob a ação do decaimento de pesos, a norma dos pesos decresce até o ponto de bifurcação crítico, onde ocorre a transição descontínua para generalização perfeita: por volta da época 6000 para o MLP (atingindo 98,54% de validação, Fig. 5) e por volta da época 2000 para o *Transformer* Causal (Fig. 6), onde a validação salta para $> 98\%$ em menos de 500 épocas, alcançando a generalização de 99% por volta da época 2500.

A partir dessas curvas de aprendizado, observa-se um padrão claro: em todas as tarefas, as arquiteturas baseadas em *Transformer* atingem a transição de fase (*grokking*) de forma significativamente mais rápida e abrupta do que as redes MLP (por exemplo, 1500 épocas vs. 5000 épocas na aritmética modular, e 2000 épocas vs. 6000 épocas na conservação de energia). Sob a ótica da física estatística, esse comportamento pode ser compreendido pela natureza do acoplamento entre os constituintes do sistema (os *tokens*).

No MLP, as conexões sinápticas representam um acoplamento estático: os pesos W são fixados após o treinamento e aplicam a mesma transformação linear a qualquer par de entrada, independentemente dos valores específicos de a e b. A quebra de simetria em direção ao estado ordenado (generalização) no MLP depende de um ajuste lento e global de todas as coordenadas de pesos sob a ação difusiva do decaimento de pesos L_2.

Por outro lado, o mecanismo de atenção do *Transformer* introduz um acoplamento dinâmico: a matriz de atenção é recalculada a cada passo de inferência baseando-se nas correlações mútuas entre os vetores de consulta (*queries*) e chaves (*keys*) de cada entrada. Esse acoplamento dinâmico atua como um potencial de interação dependente do estado instantâneo do sistema, permitindo que a rede reconfigure as forças de atração e repulsão latentes de forma adaptativa. Assim como em sistemas físicos reais (como redes de *spins* sob acoplamentos de longo alcance flutuantes ou sinapses plásticas), os graus de liberdade adicionais fornecidos pelos acoplamentos dinâmicos reduzem drasticamente as barreiras de energia livre que separam o estado de memorização desordenada do estado generalizado ordenado. Isso acelera a transição termodinâmica e torna a quebra de simetria do *Transformer* muito mais abrupta.

4.2 Visualização do Espaço de Fase e *Embeddings*

Para provar que a transição de fase está associada à descoberta da simetria modular, podemos extrair a matriz de pesos da camada de *embedding* $W_E \in \mathbb{R}^{p \times d_{\text{embed}}}$ após o *grokking*. Aplicando a Análise de Componentes Principais (PCA) para reduzir a dimensionalidade dos *embeddings* aprendidos de 64 para 2 dimensões, observamos que os vetores correspondentes aos números $\{0, \ldots, p-1\}$ se organizam perfeitamente na forma de um círculo ordenado espacialmente de 0 a $p-1$, conforme ilustrado na Fig. 7.

Essa representação geométrica circular é a prova empírica de que a rede aprendeu a topologia

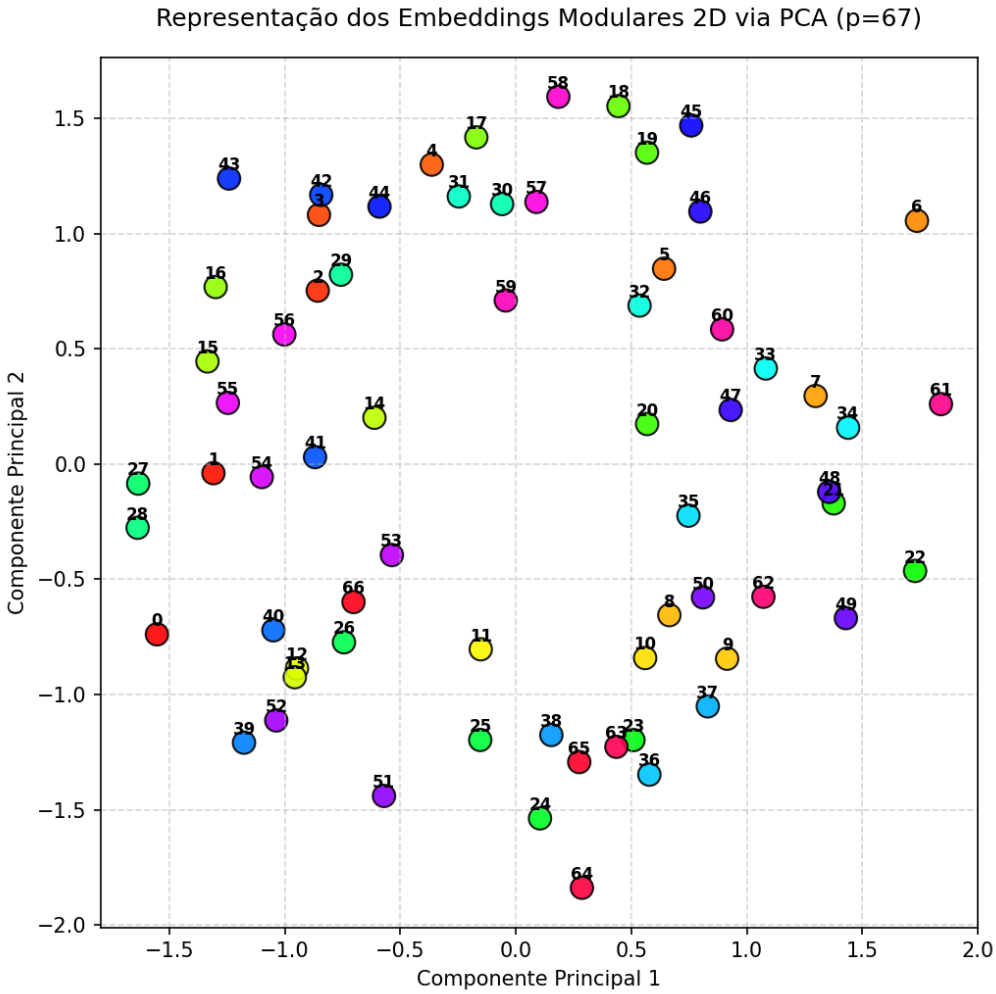

Figura 7: Projeção bidimensional via PCA da matriz de pesos da camada de *embedding* (W_E) do MLP treinado ($p = 67$). A auto-organização dos *tokens* numéricos em um arranjo circular demonstra que a rede aprendeu a topologia discreta do grupo cíclico C_{67} para realizar somas por meio de rotações geométricas.

cíclica do grupo C_p. A ordenação circular permite que operações de soma modular sejam computadas geometricamente como rotações em duas dimensões, demonstrando a beleza matemática e a relevância de conceitos físicos (simetrias contínuas e mecânica ondulatória) no seio das redes neurais.

4.3 Estados Degenerados de Energia e Órbitas na Conservação de Energia

Na tarefa de Conservação de Energia Modular ($x^2+y^2 \pmod p$), a dinâmica de aprendizado nos *embeddings* revela um fenômeno geométrico ainda mais rico, que possui analogias profundas com a mecânica clássica e a física quântica.

Conforme discutido na metodologia, a função alvo é invariante sob as reflexões $x \mapsto p-x$ e $y \mapsto p-y$, uma vez que $x^2 \equiv (p-x)^2 \pmod p$. Na física, dois estados distintos com a mesma energia são denominados estados degenerados. O termo de regularização por decaimento de pesos ($\lambda = 1{,}0$) atua minimizando a energia livre do sistema e eliminando a redundância de representação. Sob essa força atrativa harmônica, a rede neural aprende a colapsar os *embeddings* dos estados degenerados x e $p-x$ exatamente sobre a mesma coordenada no espaço latente de dimensão $d_{\text{embed}} = 64$.

Ao extrairmos a matriz de pesos da camada de *embedding* pós-*grokking* e aplicarmos PCA de 2 componentes (conforme ilustrado na Fig. 8 para o MLP e na Fig. 9 para o *Transformer* Causal), observamos essa quebra de simetria de forma cristalina.

Como mostrado nas Figuras 8 e 9, observa-se uma diferença interessante entre as arquiteturas: no MLP, todos os pares $\{x, p-x\}$ se sobrepõem de forma praticamente perfeita. No *Transformer* Causal, embora a simetria de pareamento seja mantida na alta dimensão (onde o vizinho mais próximo de cada vetor $E(x)$ no espaço de *embedding* de 64 dimensões é sempre o seu par degenerado $E(p-x)$, com distâncias na ordem de 0,07), eles não colapsam em uma coordenada idêntica e apresentam uma separação visível no PCA 2D (por exemplo, os pares $\{1, 58\}$ e $\{2, 57\}$). Esse comportamento decorre de dois fatores principais. Primeiro, a atenção causal introduz uma assimetria intrínseca entre as posições 0 e 1 (o *token* na posição 1 atende à posição 0, mas o inverso não ocorre). Como a matriz de *embedding* é compartilhada entre as posições, os vetores $E(x)$ precisam conciliar essa assimetria posicional, o que afasta os pares degenerados. Segundo, a projeção linear por PCA bidimensional para o *Transformer* explica apenas cerca de 34% da variância total dos dados de 64 dimensões (em comparação com cerca de 12% no MLP, onde a representação de alta dimensão é intrinsecamente bidimensional). Essa perda de informação na projeção distorce as distâncias euclidianas no plano 2D, fazendo com que estados pertencentes a órbitas circulares concêntricas adjacentes (com diferentes níveis de energia) possam projetar-se muito próximos ou até mesmo entre os elementos de um par degenerado (como o elemento 49 projetando-se próximo a 1 no plano do PCA), apesar de estes últimos serem estritamente os vizinhos mais próximos no espaço completo. O elemento 0 (único estado não-degenerado, já que $0 \equiv -0 \pmod p$) se localiza isoladamente. Ademais, as $(p+1)/2 = 30$ classes equivalentes de resíduos quadráticos se auto-organizam em uma estrutura espacial composta por múltiplas órbitas circulares concêntricas. Isso ocorre porque o núcleo da tarefa é a adição modular dos resíduos quadráticos. Uma vez que a adição modular induz representações harmônicas (círculos), ambas as arquiteturas projetam os resíduos quadráticos em geometrias circulares de diferentes raios no espaço bidimensional do PCA, de forma correspondente aos níveis discretos de energia. Essa representação geométrica é a confirmação visual direta de que as redes neurais descobrem simetrias contínuas a partir de dados discretos e ruidosos para atingir a generalização perfeita.

4.4 Análise Espectral de Fourier nos *Embeddings*

Para validar analiticamente a hipótese de que a rede neural descobre a simetria cíclica discreta C_p por meio de representações trigonométricas periódicas, realizamos a análise de Fourier na

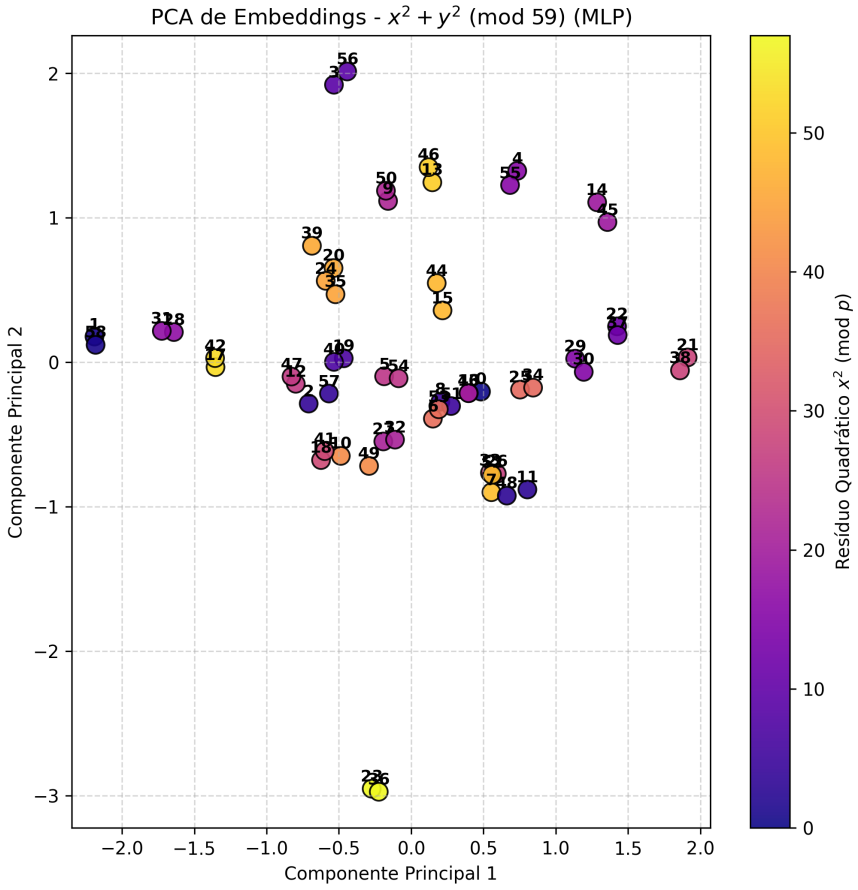

Figura 8: Projeção via PCA bidimensional dos *embeddings* da rede MLP treinada na tarefa de Conservação de Energia Modular ($p = 59$). Nota-se o pareamento exato dos estados degenerados $\{x, p - x\}$ (por exemplo, 3 e 56, 12 e 47, 1 e 58) e a organização das classes de equivalência quadrática na forma de órbitas ou anéis concêntricos ao redor do centro.

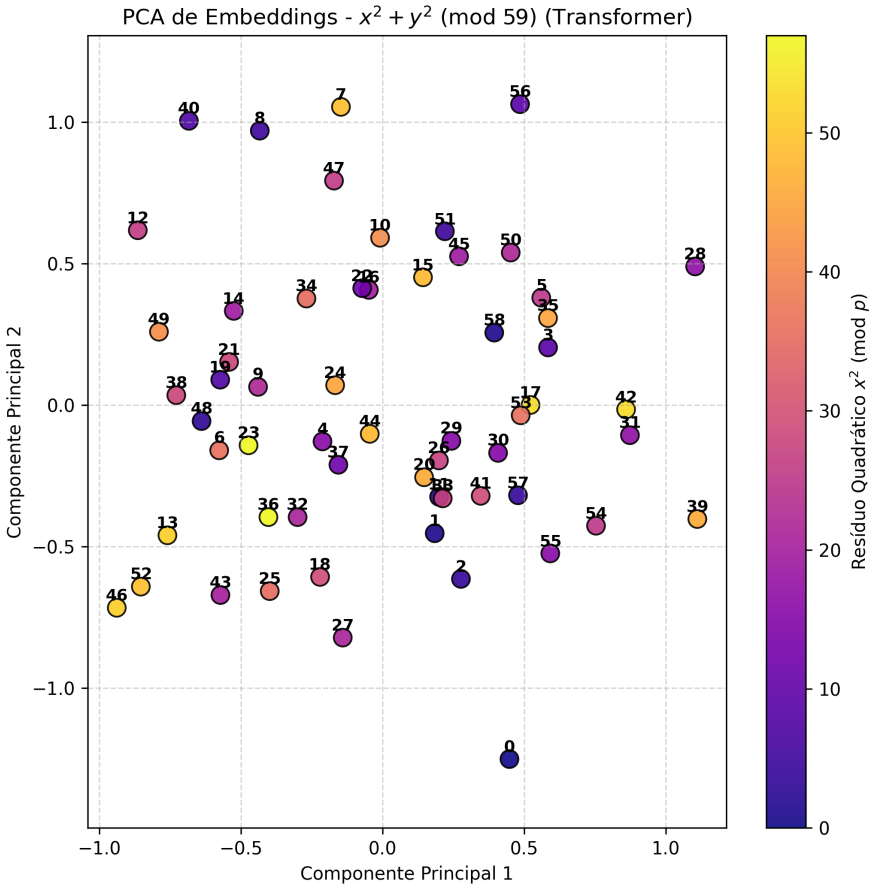

Figura 9: Projeção via PCA bidimensional dos *embeddings* da camada de *embedding* do *Transformer* Causal na tarefa de Conservação de Energia Modular ($p = 59$). Embora os pares $\{x, p - x\}$ permaneçam como vizinhos mais próximos no espaço multidimensional, a assimetria da atenção causal impede o seu colapso perfeito em coordenadas idênticas no PCA 2D. A organização concêntrica de anéis correspondente aos resíduos quadráticos é mantida.

matriz de pesos da camada de *embedding* $W_E \in \mathbb{R}^{p \times d_{\text{embed}}}$.

Para cada recurso individual (coluna) $j \in \{0, \ldots, d_{\text{embed}} - 1\}$ da matriz de pesos, tratamos as ativações dos inteiros modulares $x \in \{0, \ldots, p-1\}$ como um sinal discreto periódico de comprimento p, denotado por $v_j[x] = W_E[x, j]$. A Transformada Discreta de Fourier (DFT) unidimensional deste sinal ao longo do eixo do vocabulário é definida por:

$$\tilde{v}_j[k] = \sum_{x=0}^{p-1} v_j[x] e^{-i\frac{2\pi kx}{p}} \tag{13}$$

onde $k \in \{0, \ldots, p-1\}$ representa o número de onda espacial (frequência discreta). O espectro de potência associado a cada recurso j na frequência k é dado por:

$$S_j[k] = |\tilde{v}_j[k]|^2 \tag{14}$$

Para quantificar o comportamento global da representação da camada de *embedding*, definimos o espectro de potência médio normalizado $S[k]$ sobre todas as $d_{\text{embed}} = 64$ dimensões do *embedding*:

$$S[k] = \frac{1}{\sum_{k'=0}^{\lfloor p/2 \rfloor} \bar{S}[k']} \bar{S}[k], \quad \bar{S}[k] = \frac{1}{d_{\text{embed}}} \sum_{j=0}^{d_{\text{embed}}-1} S_j[k] \tag{15}$$

Devido à simetria da DFT para sinais reais ($S_j[k] = S_j[p-k]$), limitamos nossa análise ao intervalo de frequências positivas $k \in \{0, \ldots, \lfloor p/2 \rfloor\}$.

No contexto físico, o número de onda k representa o número de oscilações completas de uma onda harmônica confinada em uma geometria de anel unidimensional sob condições de contorno periódicas de comprimento p. No aprendizado de máquina, cada frequência k corresponde a um canal representacional circular independente em que a rede neural projeta os operandos. A utilidade computacional dessa projeção periódica reside no fato de que a operação algébrica de soma modular $a + b \pmod p$ exibe invariância translacional intrínseca no grupo cíclico C_p.

Ao projetar os operandos no plano bidimensional na forma de círculos de número de onda k:

$$E_k(x) \propto \left[\cos\left(\frac{2\pi kx}{p}\right), \sin\left(\frac{2\pi kx}{p}\right)\right]^T \tag{16}$$

a rede neural é capaz de calcular a adição modular de forma exata e contínua através de rotações geométricas bidimensionais (como matrizes de rotação $R(\theta)$). Utilizando identidades trigonométricas clássicas como a soma de arcos, o produto de ativações dos operandos nas camadas ocultas subsequentes resulta linearmente na representação do resultado final $a + b \pmod p$. Esse mecanismo elimina a necessidade de memorizar individualmente os p^2 pares no conjunto de treinamento, resultando em uma representação com menor norma L2 (menor energia efetiva).

A Fig. 10 apresenta a comparação do espectro de potência por recurso (painéis superiores) e do espectro médio normalizado (painéis inferiores) obtidos para o modelo generalizador ($\lambda = 1{,}0$) e o modelo memorizador ($\lambda = 0{,}0$).

Observa-se que no modelo generalizador (painéis da esquerda na Fig. 10), a potência média concentra-se quase exclusivamente em harmônicos específicos (frequências de número de onda discretos como $k = 1, 2, \ldots$). Isso indica que a rede de fato aprendeu a representar a invariância translacional modular decompondo a entrada em suas componentes trigonométricas de baixa frequência. Em contrapartida, no modelo memorizador (painéis da direita), a energia distribui-se de maneira amorfa por todo o espectro, assemelhando-se a um espectro de ruído branco de alta temperatura. Essa análise fornece uma prova espectral robusta de que o *grokking* corresponde a uma transição de uma fase desordenada (amorfismo de pesos) para uma fase ordenada (cristalização trigonométrica).

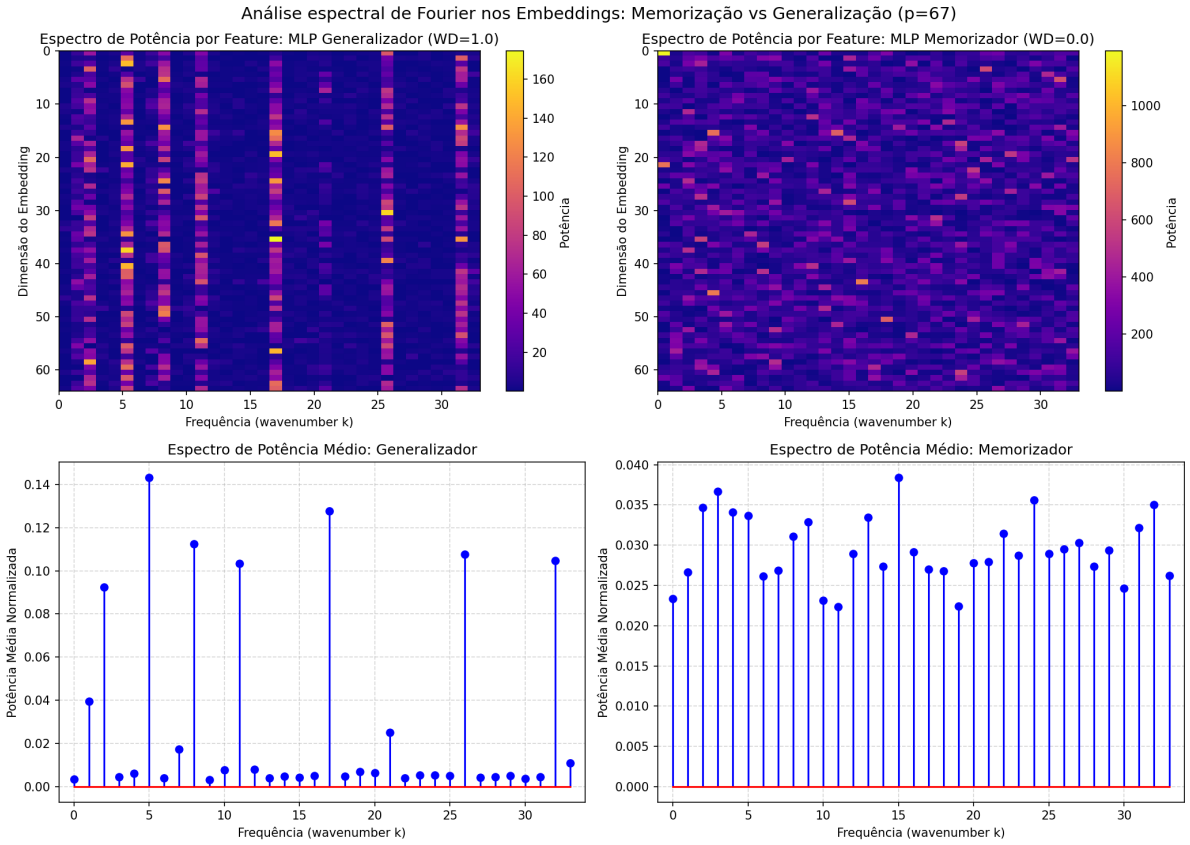

Figura 10: Análise espectral de Fourier nos *embeddings* pós-treinamento para $p = 67$. Painéis esquerdos: MLP generalizador ($\lambda = 1{,}0$) exibindo espectro por recurso (topo) e potência média com harmônicos nítidos (base). Painéis direitos: MLP memorizador ($\lambda = 0{,}0$) exibindo espectro por recurso (topo) e potência média com ruído plano (base).

4.5 O Platô Metaestável e a Transição de Fase na Paridade Esparsa

No problema da paridade esparsa, a dinâmica de aprendizado de máquina exibe o comportamento clássico de um sistema fora do equilíbrio retido em um estado metaestável desordenado. Conforme ilustrado nas curvas de acurácia da Fig. 3, a acurácia de treino converge quase instantaneamente para 100% nas primeiras 200 épocas. Contudo, a acurácia de validação permanece estagnada em cerca de 51% por cerca de 1000 épocas.

Sob a perspectiva da física de sistemas desordenados, esse comportamento descreve a presença de uma fase de memorização (metaestável). Como a paridade de $k = 3$ bits específicos é mascarada por 7 bits de ruído puro em um vetor de entrada de 10 dimensões, a rede neural inicialmente encontra uma bacia de atração no espaço de parâmetros que resolve a tarefa localmente via memorização puramente arbitrária. A entropia de Shannon desta fase desordenada é imensa (pois há incontáveis maneiras de combinar os pesos sinápticos de forma aleatória para memorizar apenas 10% do espaço de estados total). No entanto, essas soluções de memorização exigem que a norma L_2 dos pesos cresça continuamente para construir as barreiras energéticas no espaço multidimensional.

O papel do decaimento de pesos $\lambda = 0{,}5$ atua de forma rigorosa como uma força termodinâmica de resfriamento. Como visto na Fig. 3 (base), a norma de pesos L_2 cresce inicialmente enquanto o modelo memoriza o treino (atingindo o pico de cerca de 603 por volta da época 1000), mas após isso, a penalidade harmônica $\frac{\lambda}{2}\|w\|^2$ encolhe significativamente os pesos sinápticos (caindo para cerca de 236 na época 6000). Esse processo funciona como uma barreira que drena a energia livre do sistema, desestabilizando as soluções desordenadas (que possuem normas de pesos muito elevadas). Após a época 1000, o encolhimento sistemático dos pesos sinápticos atinge um ponto de bifurcação crítico. A rede subitamente descobre a simetria oculta da paridade dos $k = 3$ *spins*. A acurácia de validação dispara de 54,87% na época 1000 para 98,81% na época 6000, evidenciando uma transição de fase dinâmica de primeira ordem de natureza essencialmente entrópica.

No caso do modelo *Transformer* bidirecional para paridade esparsa (Fig. 4), observamos um comportamento qualitativamente análogo. A acurácia de validação permanece retida em um platô de cerca de 61% por milhares de épocas (representando um estado metaestável desordenado onde a rede aprendeu apenas características parciais simples de baixa ordem). Conforme o decaimento de pesos $\lambda = 0{,}5$ atua (norma L2 caindo de 921 na época 1000 para 243 na época 7000), o sistema transiciona abruptamente de fase por volta da época 3000, descobrindo a simetria de paridade e atingindo acurácia de validação de 99,16% na época 7000. Isso mostra a universalidade da transição descontínua para generalização em diferentes arquiteturas (MLPs e *Transformers*) sob regularização.

5 Conclusões

O estudo do fenômeno de *grokking* oferece uma rica oportunidade interdisciplinar para o ensino de física. Ele demonstra que redes neurais não são apenas caixas pretas de ajuste de curvas, mas sistemas dinâmicos complexos cujos processos de aprendizado e generalização compartilham profundas conexões com transições de fase térmicas e dinâmicas de minimização de energia livre.

Ao disponibilizar códigos em notebooks simples e rápidos, esperamos incentivar estudantes a experimentarem com hiperparâmetros (como remover o decaimento de pesos e notar que o *grokking* deixa de ocorrer), consolidando a intuição física sobre conceitos abstratos da termodinâmica e mecânica estatística aplicados à inteligência artificial contemporânea.

5.1 Sugestões de Aplicação Didática

A natureza interdisciplinar deste tema abre caminhos para sua abordagem em diferentes matrizes curriculares. Em cursos de Física, particularmente nas disciplinas de Física Estatística e Física Computacional, o mapeamento da função de perda e do decaimento de pesos como uma energia potencial de dados e um potencial harmônico restritivo oferece um laboratório prático valioso para ilustrar conceitualmente transições de fase fora do equilíbrio e minimização de energia livre. Já em cursos de Ciência e Engenharia de Computação, especificamente nas matérias de Aprendizado de Máquina e Redes Neurais, o tema permite introduzir com rigor matemático o efeito da regularização por decaimento de pesos, bem como analisar a dinâmica das representações internas via análise de Fourier nos *embeddings*, oferecendo uma contrapartida analítica essencial à perspectiva puramente empírica comumente adotada. Além disso, em disciplinas focadas em Sistemas Dinâmicos e Modelagem Matemática, o estudo do *grokking* serve como um exemplo contemporâneo e estimulante de como interações locais elementares e processos de otimização estocásticos podem induzir comportamentos coletivos emergentes e transições de fase abruptas.

Declaração de Disponibilidade de Dados

Todos os códigos-fonte e dados gerados necessários para a reprodução dos experimentos discutidos neste artigo encontram-se disponíveis publicamente no repositório GitHub do projeto em `https://github.com/jsansao/grokking`.

Conflito de Interesses

O autor declara que não há conflito de interesses na elaboração e publicação deste artigo.

Referências

[1] J. J. Hopfield, Proceedings of the National Academy of Sciences **79**, 2554 (1982).

[2] D. J. Amit, H. Gutfreund e H. Sompolinsky, Physical Review Letters **55**, 1530 (1985).

[3] G. G. de Lima, G. C. de Miranda e T. de S. Farias, Revista Brasileira de Ensino de Física **48**, e20250183 (2026).

[4] D. Sherrington e S. Kirkpatrick, Physical Review Letters **35**, 1792 (1975).

[5] M. Mézard, G. Parisi e M. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, 1987).

[6] S. R. A. Salinas, *Introdução à Física Estatística* (Edusp, São Paulo, 2001).

[7] N. Goldenfeld, *Lectures on Phase Transitions and the Renormalization Group* (Addison-Wesley, 1992).

[8] G. Cybenko, Mathematics of Control, Signals, and Systems **2**, 303 (1989).

[9] I. Goodfellow, Y. Bengio e A. Courville, *Deep Learning* (MIT Press, 2016).

[10] A. Engel e C. Van den Broeck, *Statistical Mechanics of Learning* (Cambridge University Press, 2001).

[11] A. Krogh e J. A. Hertz, em *Advances in Neural Information Processing Systems* (1991), p. 950.

[12] M. Welling e Y. W. Teh, em *Proceedings of the 28th International Conference on Machine Learning* (2011), p. 681.

[13] T. Mikolov, K. Chen, G. Corrado e J. Dean, arXiv preprint arXiv:1301.3781 (2013).

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser e I. Polosukhin, em *Advances in Neural Information Processing Systems* (2017), p. 5998.

[15] A. Power, Y. Burda, H. Edwards, I. Babuschkin e I. Sutskever, arXiv preprint arXiv:2201.02177 (2022).

[16] N. Nanda, L. Chan, T. Lieberum, J. Smith e J. Steinhardt, arXiv preprint arXiv:2301.05217 (2023).

[17] Z. Liu, E. Tegmark e M. Tegmark, arXiv preprint arXiv:2210.01117 (2022).

[18] J. Lau, R. S. S. de Oliveira e G. M. A. de Almeida, Journal of Statistical Mechanics: Theory and Experiment (2023).

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.