Noise-Driven Escape from Metastable Phases explains Grokking in Deep Neural Networks

Ibrahim Talha Ersoy TALHA.ERSOY@UNI-POTSDAM.DE
Karoline Wiesner KAROLINE.WIESNER@UNI-POTSDAM.DE
Complexity Science Group, Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany

Abstract

Deep neural networks (DNNs) exhibit first order phase transitions under variations of the L2 regularization strength, with each transition marking the onset of a new learnable feature. Below a critical regularization strength, all features are in principle learnable, but coexisting metastable states, separated by energy barriers, can trap the network and impede convergence. A strength of DNNs is their ability to generalize. But many open questions remain, among them the origin of so called grokking: the abrupt, delayed onset of generalization after prolonged apparent overfitting. We show for linear DNNs that grokking is consistent with hysteresis in first-order L2 phase transitions: using L2 regularization to engineer deliberate trapping, we demonstrate that a model in a low-accuracy metastable state escapes only when SGD noise drives it across an energy barrier, with escape times following Arrhenius scaling. We reproduce grokking-like delayed convergence across two orders of magnitude in escape time by deliberately trapping models in metastable phases. Using sparse sub-sampling we also reproduce the canonical grokking curve where test error eventually approaches the final training error. Our work suggests that the number of metastable states equals the number of learnable features – one per singular value of the data covariance – the potential for hysteresis grows naturally with task complexity. We provide evidence that the same mechanism likely operates in general nonlinear DNNs. Our results provide routes toward more efficient learning schemes.

1. Introduction

L2 regularization is a cornerstone of modern machine learning, employed to combat overfitting from classical ridge regression [1] to large-scale deep learning [2]. Beyond its practical role, L2 regularization gives rise to rich phenomenology recently understood through statistical physics. Ziyin and Ueda [3] showed that varying regularization strength drives genuine phase transitions in DNNs, identifying first-order transitions in DNNs at the onset of learning (the transition from under- to over-parameterization). Subsequent work extended this picture beyond the onset: Ersoy and Wiesner linked these transitions to curvature drops in the loss landscape [6], Ersoy, Licha, and Wiesner connected them to learnable feature hierarchies [7], and Ladewig, Ersoy, and Wiesner showed that in linear networks each non-zero singular value of the data covariance produces its own rank transition [8], meaning transitions recur throughout training, once per learnable feature.*Grokking*, the sudden transition from near-zero to near-perfect generalization long after the training loss has saturated, has attracted much attention since Power et al. demonstrated it in transformers trained on modular arithmetic [9]. The origin of this phenomenon is yet not fully understood. Liu et al. [10] implicated regularization as the central driver of grokking. Nanda et al. [21] identified interpretable representations at the transition, and Tian [22] derived scaling laws for feature emergence. Rubin

et al. [11] proposed a first-order phase transition analogy involving an entropy barrier. This was then challenged by Zhang et al. [23], who found no entropy barrier and proposed glassy relaxation instead. Solvable linear models have likewise been shown to grok [24], and a distribution shift between training and test data has been identified as a driver of delayed generalization [25]. We offer an alternative account. As the central mechanism we suggest *hysteresis* in analogy to the statistical physics of phase transitions: a model initialized in a low-accuracy phase remains there until SGD noise drives it across an energy (loss) barrier into the globally optimal phase. To show that hysteresis is consistent with the grokking phenomenology, we build on the finding that varying the L2 regularization strength leads to first order phase transitions. we ask: Can activated escape from the resulting metastable states, rather than the transitions themselves, account for the hallmark features of grokking? To answer this, we use L2 regularization as a control tool to engineer metastable trapping. We successfully reproduce grokking behavior, and, furthermore, show that this activated process is governed by Arrhenius-type kinetics [12, 13] with an effective temperature [14] $T_{\text{eff}} \propto \eta_{\text{lr}}/B$. Here η_{lr} is the learning rate and B the batch size, making the escape time exponentially sensitive to hyperparameter choices. Our results, established in Sections 2–3, support three claims: **(1)** First-order L2 phase transitions produce d coexisting metastable states (one per learnable feature) whose energy barriers trap models in low-accuracy phases. **(2)** Escape is analogous to a thermally activated process obeying Arrhenius kinetics with $T_{\text{eff}} \propto \eta_{\text{lr}}/B$; we confirm this numerically with $R^2 = 0.991$. **(3)** Deliberate trapping reproduces hallmarks of grokking, i.e. long delay, abruptness, sensitivity to initialization, and, under sparse sampling, the train/test dissociation, across two orders of magnitude in escape time.

2. Methods

2.1. L2 Phase Transitions

We use deep linear networks as our minimal model because their loss landscape is exactly solvable, allowing us to locate all d metastable minima and energy barriers analytically. All qualitative results, saddle-node bifurcations, coexisting phases, Arrhenius escape, survive in nonlinear networks (Appendix G), but the linear case keeps the analysis tractable. In this section we review the key prior results on L2 phase transitions before introducing our escape-time framework. Ladewig et al. [8] described the precise mechanism for linear DNNs, linking the transitions to the singular values of the data covariance: with input x and output y, covariances Σ_{xx}, Σ_{yy}, and cross-covariance Σ_{yx}, the singular values η_i of Σ_{yx} directly characterize the learnable features in the aligned case ($\Sigma_{xx} = \mathbf{I}$). After the network reaches the 0-balanced subspace, the set of weight configurations where all layers contribute equally to the end-to-end map, the L2-regularized loss decouples into independent terms for each singular value λ_i of the end-to-end weight matrix:

$$\mathcal{L}(\{\lambda_i\}, \beta) = \frac{1}{2}\sum_{i=1}^{d}(\lambda_i - \eta_i)^2 + \frac{\beta}{2}\sum_{i=1}^{d}\lambda_i^{2/L}, \tag{1}$$

where L is network depth, d is the number of non-zero modes, and $\beta > 0$ is the regularization strength. For depth $L \geq 3$, the stationarity condition $\partial\mathcal{L}/\partial\lambda_i = 0$ for a single mode reads:

$$\lambda - \eta + \beta\,\lambda^{2/L-1} = 0. \tag{2}$$

As β increases through β_c, the two non-trivial solutions (a stable minimum and an unstable saddle) merge and annihilate in a saddle-node bifurcation (see Appendix D), leaving only $\lambda = 0$. Below β_c,

zero and non-zero rank solutions coexist, separated by a finite energy barrier (see Appendix C for the bifurcation diagram). The critical regularization strength is:

$$\beta_c^{(i)} = \frac{1}{1-k}\left(\eta_i \frac{1-k}{2-k}\right)^{2-k}, \qquad k = \frac{2}{L}. \tag{3}$$

For $\beta > \beta_c^{(i)}$ the metastable minimum vanishes. The equal-loss point β_i^* (where the lower and higher rank solutions have equal regularized loss) lies strictly below $\beta_c^{(i)}$ for $L \geq 3$, producing hysteresis: a model remains trapped in the lower rank phase even when the higher rank phase is energetically preferred. The ordering $\eta_1 > \cdots > \eta_d > 0$ gives d distinct bifurcations, one metastable state per learnable feature.

2.2. SGD as Langevin Dynamics and Arrhenius Escape

To model escape from metastable states, we note that SGD mini-batch noise injects stochastic fluctuations into the gradient with covariance scaling as η_{lr}/B, where η_{lr} is the learning rate and B the batch size. In the overdamped limit, the dynamics maps onto Langevin dynamics with an effective temperature [14]:

$$T_{\text{eff}} \propto \frac{\eta_{\text{lr}}}{B}. \tag{4}$$

The mean escape time from a metastable state then follows the Kramers–Arrhenius law [12, 13]:

$$\ln \tau = \ln \tau_0 + \frac{\Delta E_{\text{eff}}}{T_{\text{eff}}}, \tag{5}$$

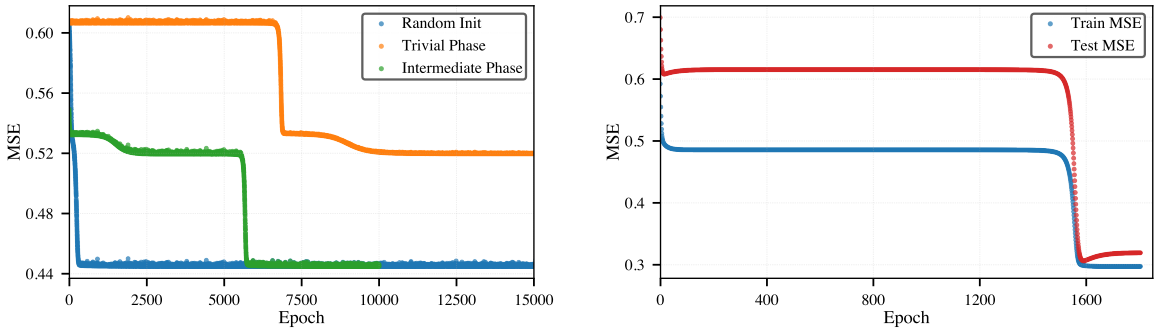
where ΔE_{eff} is an effective barrier height absorbing high-dimensional curvature and entropic corrections (see Appendix F). This yields the falsifiable prediction that $\ln \tau$ is linear in B/η_{lr}.

3. Results

3.1. Hysteresis and Trapping Reproduce Grokking

The coexistence of metastable states implies that the training outcome depends sensitively on initialization. To demonstrate this we operate at $\beta = 0.32$, which lies in the range $\beta < \beta_c^{(1)}$, so the global minimum is rank-2 but a metastable rank-1 minimum also exists. All three experiments use the learning rate $\eta_{\text{lr}} = 0.08$ and batch size $B = 64$. We consider three initialization protocols (Fig. 1): **(i) Random initialization.** *Setup*: standard random weight initialization at $\beta = 0.32$. *Observation*: the model converges rapidly to the rank-2 global minimum (Fig. 1, blue; $\tau \approx 10$ epochs). *Conclusion*: no trapping occurs when the model starts outside any metastable phase. **(ii) Rank-1 trap.** *Setup*: model initialized from a checkpoint trained at $\beta > \beta_c^{(1)}$, placing it in the metastable rank-1 phase. *Observation*: the model remains in the rank-1 state for $\tau \approx 5500$ epochs before abruptly escaping to rank-2 (Fig. 1, green). *Conclusion*: trapping in a metastable phase produces grokking-like delayed convergence. **(iii) Trivial-phase trap.** *Setup*: model initialized from a checkpoint trained at $\beta > \beta_c^{(2)}$, placing it in the rank-0 phase. *Observation*: the model escapes to rank-1 after $\tau > 7000$ epochs, then remains trapped in rank-1 for a further extended period, with total delay $\tau > 10{,}000$ epochs (Fig. 1, orange). *Conclusion*: sequential trapping across multiple metastable phases reproduces staged grokking delays spanning two orders of magnitude. **(iv) Canonical grokking via sparse**

sub-sampling. *Setup*: train and test are drawn from the *same* distribution (weak correlation 0.8, strong correlation 0.9), but the training set is very sparse, only 25 samples (0.5% of a 5 000-sample pool), so the weak feature is poorly determined from the training data while remaining at full strength on test. The model is initialised in the rank-1 trapped state and trained at $\beta = 0.0025$, a small regularisation chosen so that the plateau is long-lived but eventually resolves. *Observation*: After a brief transient in which the strong mode equilibrates, both errors plateau, with train below test (Fig. 1(b)); the model holds in the rank-1 phase for $\approx$1500 epochs, then escapes to rank-2, and test error falls sharply to approach train error to within a small residual gap set by the irreducible noise of the stochastic task. *Conclusion*: the same trapping mechanism reproduces the canonical grokking curve when the weak feature is sufficiently underrepresented in the training sample.



(a) Hysteresis as Initialization Dependence **(b) Train-Test Discrepancy in Hysteresis**

Figure 1: Delayed convergence in deep linear networks. **(a)** Random (blue), rank-1 trap (green), trivial-phase trap (orange) at $\beta = 0.32$, $\eta_{\text{lr}} = 0.08$, $B = 64$. The convergence is strongly delayed when the model is initialized in the local minima of the lower accuracy phases. **(b)** Canonical grokking via sparse sub-sampling (25 training samples, $\beta = 0.0025$). Initialised in the rank-1 phase, train MSE (blue) drops quickly but only to a plateau while test MSE (red) plateaus higher; at $\approx$1500 epochs the model escapes the rank-1 phase and test MSE falls sharply, approaching train MSE to within a small residual gap set by the irreducible noise of the task.

3.2. Energy Barrier Between Metastable States

The trapping mechanism requires a finite energy barrier between the differing rank phases. Fig. 2(a) shows the loss landscape section along the minimal-loss path at $\beta = 0.32$, parameterized by the singular value λ. A clear barrier separates the local minimum at $\lambda = 0$ from the global minimum. Evaluating Eq. (1) numerically at the saddle point $\lambda^{\text{sad}} \approx 0.41$ (obtained from Eq. (2) at $\beta = 0.32$, $\eta = 0.8$, $L = 3$). This gives $\Delta E_{\min} \approx 0.003$, the barrier along the lowest-loss path out of the metastable phase (Fig. 2(a)). As we show next, the effective barrier governing escape times is far larger.

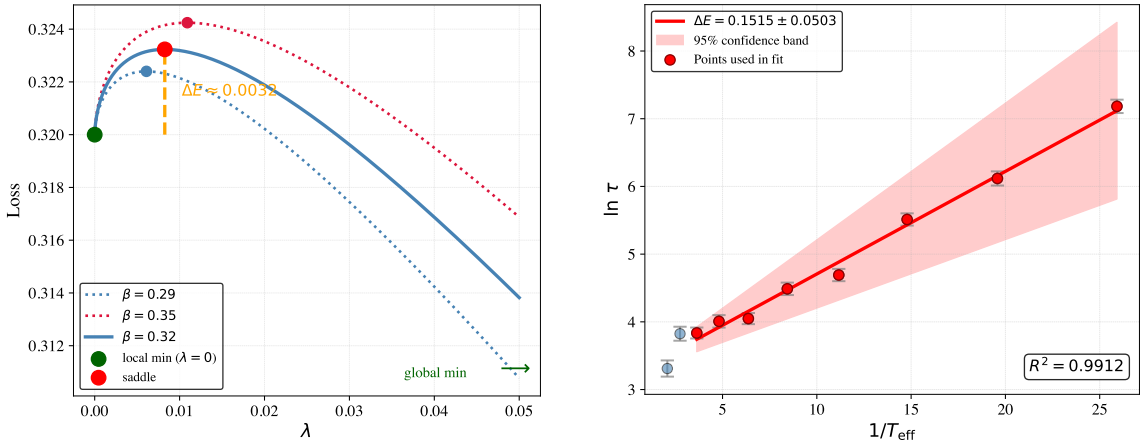
3.3. Arrhenius Scaling Confirms Thermally Activated Escape

To test whether the escape is governed by the activated barrier crossing as predicted by Eq. (5), we measured escape times from the rank-1 trapped state at $\beta = 0.32$ while varying $\eta_{\text{lr}} \in [5 \times 10^{-4}, 5 \times 10^{-3}]$ with batch size fixed at $B = 64$. With $T_{\text{eff}} \propto \eta_{\text{lr}}/B$, Eq. (5) predicts $\ln \tau$ to be linear in B/η_{lr}. Figure 2(b) shows this Arrhenius plot; the linear fit achieves $R^2 = 0.991$,confirming Eq. (5). The

slope yields the effective barrier:

$$\Delta E_{\text{eff}} = 0.15 \pm 0.05. \tag{6}$$

This is far above $\Delta E_{\min} \approx 0.003$ shown in Fig. 2(a). The discrepancy is expected: in a $D = 170$-dimensional parameter space the Kramers–Langer formula adds entropic and geometric corrections from the $D - 1$ transverse directions, driving $\Delta E_{\text{eff}} \gg \Delta E_{\min}$ (Appendix F).



(a)Loss Landscape Barrier **(b) Arrhenius Fit for Escape Times**

Figure 2: (a) Loss landscape section at $\beta = 0.32$, $\eta = 0.8$, $L = 3$; dotted curves show $\beta = 0.29$ and $\beta = 0.35$. The barrier from the local minimum at $\lambda = 0$ to the saddle gives $\Delta E_{\min} \approx 0.003$. **(b)** Arrhenius fit: $\ln \tau$ versus $1/T_{\text{eff}}$. The linear fit ($R^2 = 0.991$) confirms thermally activated barrier crossing; the slope gives $\Delta E_{\text{eff}} = 0.15 \pm 0.05 \gg \Delta E_{\min}$, with the discrepancy explained by entropic and curvature corrections in $D = 170$ dimensions (Appendix F).

4. Discussion

4.1. A Mechanistic Account of Grokking

Our results establish three claims. First, first-order L2 phase transitions produce coexisting metastable states whose barriers trap models in low-accuracy phases. Second, escape from these states is an activated process governed by Arrhenius kinetics with $T_{\text{eff}} \propto \eta_{\text{lr}}/B$. Third, deliberate trapping reproduces the hallmark features of grokking, long delay, abruptness, sensitivity to initialization, and the train/test dissociation under sparse sampling, in an analytically tractable minimal model. This framework makes concrete, falsifiable predictions:

1. *Staged grokking*: In tasks with d learnable features, grokking should proceed in up to d discrete stages, one per singular value of Σ_{yx}.

2. *Depth dependence*: Deeper networks (L larger) produce higher barriers and thus longer grokking delays, since the critical strength $\beta_c^{(i)}$ decreases with L while the equal-loss crossing β_i^* remains finite [8].

3. *Hyperparameter control*: Escape time obeys $\ln \tau \propto B/\eta_{\text{lr}}$, providing a direct lever to accelerate or suppress grokking.

Predictions (2) and (3) are immediately testable by varying L, η_{lr}, and B in existing grokking benchmarks. Preliminary experiments and work on nonlinear networks [7] with sigmoid and tanh activations reproduce qualitatively identical first order phase transition behavior, strongly suggesting that the same mechanism applies beyond the linear framework.

4.2. Relation to Memorization and Generalization

In the dense-data experiments of Fig. 1(a) the training error converges monotonically, and the large train/test gap of canonical grokking [9] does not appear, because with $d = 2$ well-sampled modes the model has few directions along which train and test can diverge. Under sparse sub-sampling this gap does emerge: Fig. 1(b) shows train error settling onto a plateau while test error remains substantially higher, until both fall abruptly as the model escapes the rank-1 phase. Because the network is linear and the task is stochastic, the model cannot memorise the training sample: train error plateaus at the best rank-1 linear fit and, after escape, settles near the best rank-2 fit, which is bounded below by the irreducible (Bayes) error of the Gaussian task. Train and test therefore approach one another and the noise floor but cannot coincide or fall to zero. The grokking signature here is thus the delayed, abrupt closing of a train/test gap down to the task's noise floor, reproduced *without any memorising solution* — consistent with the view that memorisation is not a necessary ingredient of the phenomenon. When d is large, the optimization trajectory reduces training error by learning some singular directions while stagnating in others, giving rise to an apparent memorization phase. In this view, memorization and generalization need not be primitive concepts but instead emerge as descriptions of partial versus complete progress through a cascade of rank transitions. Moreover, what is currently labeled "grokking" possibly encompasses several mechanistically distinct phenomena with a superficial resemblance; our framework offers a principled basis for distinguishing them.

5. Conclusion

We propose a candidate mechanism for grokking: noise-activated escape from metastable states created by first-order phase transitions in L2-regularized deep networks. Using deep linear networks, i.e. the minimal setting in which the loss landscape is exactly solvable [8], we demonstrate three results: **(i)** deliberate trapping in metastable phases reproduces grokking-like delayed convergence; **(ii)** escape times obey Arrhenius scaling $\ln \tau \propto 1/T_{\mathrm{eff}}$ with $T_{\mathrm{eff}} \propto \eta_{\mathrm{lr}}/B$; and **(iii)** the extracted barrier $\Delta E_{\mathrm{eff}} = 0.15 \pm 0.05$ exceeds the minimum-path barrier $\Delta E_{\mathrm{min}} \approx 0.003$ by the factor predicted from entropic and curvature corrections in $D = 170$ dimensions. The framework connects grokking to the established physics of noise-activated escape from metastable states [12, 13, 19], and offers a direct practical handle: since $\ln \tau \propto B/\eta_{\mathrm{lr}}$, grokking delays can in principle be accelerated or suppressed by hyperparameter choice alone.

Acknowledgments

We thank Björn Ladewig, and members of the Complexity Science Group for stimulating discussions and inspiring contributions.

References

[1] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics* **12**, 55 (1970).

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2017).

[3] Ziyin, L. & Ueda, M. Zeroth, first, and second-order phase transitions in deep neural networks. *Physical Review Research*. **5**, 043243 (2023)

[4] S. Amari, *Information Geometry and Its Applications*, Vol. 194 (Springer, Tokyo, 2016).

[5] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*, Vol. 25 (Cambridge University Press, Cambridge, 2009).

[6] I. T. Ersoy and K. Wiesner, "Exploring L2-phase transitions on error landscapes," *Workshop on High-Dimensional Learning Dynamics* (2025).

[7] I. T. Ersoy, A. F. C. Licha, and K. Wiesner, "Phase transitions reveal hierarchical structure in deep neural networks," arXiv:2512.11866 (2025).

[8] B. Ladewig, I. T. Ersoy, and K. Wiesner, "Cascading through the Hierarchy: Regularizer-induced Feature Detection as Phase Transitions in Deep Linear Neural Networks," *(in preparation)* (2026).

[9] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization beyond overfitting on small algorithmic datasets," arXiv:2201.02177 (2022).

[10] Z. Liu, E. J. Michaud, and M. Tegmark, "Omnigrok: Grokking beyond algorithmic data," arXiv:2210.01117 (2022).

[11] N. Rubin, I. Seroussi, and Z. Ringel, "Grokking as a first order phase transition in two layer networks," *International Conference on Learning Representations* (2024).

[12] H. A. Kramers, "Brownian motion in a field of force and the diffusion model of chemical reactions," *Physica* **7**, 284 (1940).

[13] P. Hänggi, P. Talkner, and M. Borkovec, "Reaction-rate theory: fifty years after Kramers," *Rev. Mod. Phys.* **62**, 251 (1990).

[14] S. Mandt, M. D. Hoffman, and D. M. Blei, "Stochastic gradient descent as approximate Bayesian inference," *J. Mach. Learn. Res.* **18**, 4873 (2017).

[15] Smith, S., Kindermans, P., Ying, C. & Le, Q. Don't decay the learning rate, increase the batch size. *ArXiv Preprint ArXiv:1711.00489*. (2017)

[16] Welling, M. & Teh, Y. Bayesian learning via stochastic gradient Langevin dynamics. *Proceedings Of The 28th International Conference On Machine Learning (ICML-11)*. pp. 681-688 (2011)

[17] Saxe, A., McClelland, J. & Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *ArXiv Preprint ArXiv:1312.6120*. (2013)

[18] Wang, Z. & Jacot, A. Implicit bias of SGD in L2-regularized linear DNNs: One-way jumps from high to low rank. *ArXiv Preprint ArXiv:2305.16038*. (2023)

[19] T. Mori, L. Ziyin, K. Liu, and M. Ueda, "Power-law escape rate of SGD," *Proceedings of the 39th International Conference on Machine Learning*, PMLR **162**, 15959 (2022).

[20] Draxler, F., Veschgini, K., Salmhofer, M. & Hamprecht, F. Essentially No Barriers in Neural Network Energy Landscape. (2019)

[21] Nanda, N., Chan, L., Lieberum, T., Smith, J. & Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *ArXiv Preprint ArXiv:2301.05217*. (2023)

[22] Y. Tian, "Provable scaling laws of feature emergence from learning dynamics of grokking," arXiv:2509.21519 (2025).

[23] Zhang, X., Shang, Y., Yang, E. & Zhang, G. Is Grokking a Computational Glass Relaxation?. (2026), https://arxiv.org/abs/2505.11411

[24] N. Levi, A. Beck, and Y. Bar-Sinai, "Grokking in linear estimators—a solvable model that groks without understanding," *International Conference on Learning Representations* (2024), arXiv:2310.16441.

[25] K. Lyu, J. Jin, Z. Li, S. S. Du, J. D. Lee, and W. Hu, "Dichotomy of early and late phase implicit biases can provably induce grokking," *International Conference on Learning Representations* (2024), arXiv:2311.18817.

Appendix A. Experimental Setup

All experiments use a deep linear network of depth $L = 3$ (two hidden layers) with hidden width $w = 10$ and output dimension $p = 2$, giving $D = 170$ total parameters. No nonlinear activation is applied; the end-to-end map is a single matrix product. The optimizer is SGD throughout.

Data generation. Training and test data each consist of $N = 512$ samples drawn from a zero-mean multivariate Gaussian over the joint input–output space. The data covariance has $d = 2$ non-zero singular values $\eta_1 = 0.9$, $\eta_2 = 0.8$ with $\Sigma_{xx} = \mathbf{I}$. Samples are generated via Cholesky decomposition: given $\Sigma = LL^\top$, each sample is drawn as $z = L\epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbf{I})$; the first p components are input and the remaining p components are output. An independent test set of the same size is held fixed throughout.

Initialization and annealing. All trapping experiments use *checkpoint initialization*: the model is first trained to convergence $\beta_{\text{init}} = 0$ to get the full rank solution. The resulting weights are then used as the starting point for continued training. This is analogous to annealing. The random initialization baseline uses fresh weights drawn from $\mathcal{N}(0, 1/\sqrt{w})$ instead.

Convergence is assessed by monitoring test MSE; escape from a metastable phase is identified as the epoch at which test MSE drops below a threshold midway between the metastable plateau value and the global minimum value.

Phase diagram (Fig. 3, Appendix B): the model is first trained to convergence at $\beta = 0$, then β is increased quasi-statically in small increments, re-training to convergence at each step.

Trapping experiments (Fig. 1): all three protocols use $\beta = 0.32$, $\eta_{\text{lr}} = 0.08$, $B = 64$.

- *Random initialization*: weights drawn from a standard normal distribution scaled by $1/\sqrt{w}$; training run for 2×10^4 epochs.
- *Rank-1 trap*: model initialized from a checkpoint trained to convergence at $\beta = 0.45 > \beta_c^{(1)} = 0.378$, placing the end-to-end weight matrix in the rank-1 phase; training then continued at $\beta = 0.32$ for 2×10^4 epochs.
- *Trivial-phase trap*: model initialized from a checkpoint trained to convergence at $\beta = 0.55 > \beta_c^{(2)} = 0.298$, placing the network in the rank-0 phase; training continued at $\beta = 0.32$ for 2×10^4 epochs.

Each experiment is repeated across 50 random seeds (seeds 42–91); escape times vary across seeds due to stochastic SGD noise, which is the activated escape mechanism under study. **Arrhenius sweep** (Fig. 2(b)): starting from the rank-1 trap initialization above, escape times are measured at $\beta = 0.32$ with $B = 64$ fixed while varying $\eta_{\text{lr}} \in \{5 \times 10^{-4}, 1 \times 10^{-3}, 2 \times 10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}\}$.

Each point is averaged over 50 seeds. The proportionality constant C in $T_{\text{eff}} = \eta_{\text{lr}}/(BC)$ is estimated using the trace of the covariance of the gradients. The linearity ($R^2 = 0.991$) is a key result. With $N = 512$ and $B = 64$, each epoch consists of $\lfloor N/B \rfloor = 8$ SGD steps; this factor enters $\ln \tau_0$ but not the slope $\Delta E_{\text{eff}}/T_{\text{eff}}$, leaving the extracted barrier unchanged.

Sparse sub-sampling experiment (Fig. 1(b)): train and test are drawn from a single zero-mean Gaussian with weak correlation 0.8 and strong correlation 0.9 ($\Sigma_{xx} = \mathbf{I}$). The training set is a random 25-sample subset of a 5 000-sample pool (0.5%), so the weak feature is poorly determined from the training data while remaining at full strength on the (disjoint, 10 000-sample) test set. The model is initialised in the rank-1 trapped state and trained at $\beta = 0.0025$, $\eta_{\text{lr}} = 0.1$, $B = 64$. Escape is identified as the epoch at which the second singular value of the end-to-end map rises above a

small threshold ($\sigma_2 > 10^{-3}$). Because the network is linear, the training error cannot reach zero: it plateaus at the best rank-1 linear fit and, after escape, settles near the best rank-2 fit, both bounded below by the irreducible (Bayes) error of the stochastic task.

Appendix B. Phase Diagram

Figure 3 shows the test MSE as a function of β for a deep linear network with $d = 2$ singular values ($\eta_1 = 0.9$, $\eta_2 = 0.8$, $\Sigma_{xx} = \mathbf{I}$, $L = 3$), obtained by training to convergence and then quasi-statically decreasing β in steps of 0.01. Two sharp drops mark the transitions at $\beta_c^{(1)} = 0.378$ and $\beta_c^{(2)} = 0.298$, in close agreement with Eq. (3). At each transition the effective rank of the end-to-end weight matrix increases by one, corresponding to the model learning an additional feature. The small offset from the theoretical values reflects the bias of the output distribution.

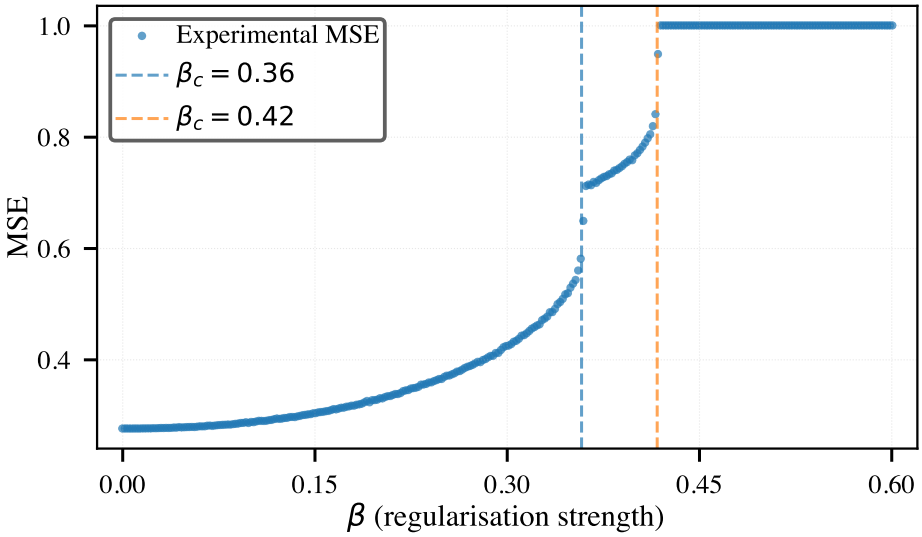


Figure 3: Phase transitions in a deep linear network ($L = 3$, $d = 2$, $\eta_1 = 0.9$, $\eta_2 = 0.8$, $\Sigma_{xx} = \mathbf{I}$). Test MSE versus regularization strength β shows two sharp drops at $\beta_c^{(1)} = 0.378$ and $\beta_c^{(2)} = 0.298$; at each transition the effective rank increases by one. Quasi-static sweep; small offset from theory reflects output distribution bias.

Appendix C. Bifurcation Diagram

The bifurcation structure of the loss landscape is shown in Fig. 4. For a single mode with signal strength η, the stationarity condition Eq. (2) has three solutions below β_c: the trivial minimum at $\lambda = 0$, an unstable saddle at λ^{sad}, and the global minimum at $\lambda^{**} > \lambda^{\text{sad}}$. As β increases through β_c, the saddle and the global minimum merge and annihilate in a saddle-node bifurcation, leaving only $\lambda = 0$. A model on the upper (rank-1) branch remains metastable for all $\beta < \beta_c$; a model on the lower (rank-0) branch at $\lambda = 0$ is trapped there even when the rank-1 phase is energetically preferred (i.e., for $\beta < \beta^*$). The height of the barrier between the two, and thus the mean escape time, grows as β increases toward β_c.

Note that for $L = 2$ no such bifurcation occurs: the stationarity condition is linear in λ and has a unique positive solution for all $\beta > 0$, yielding only second-order behavior [3].

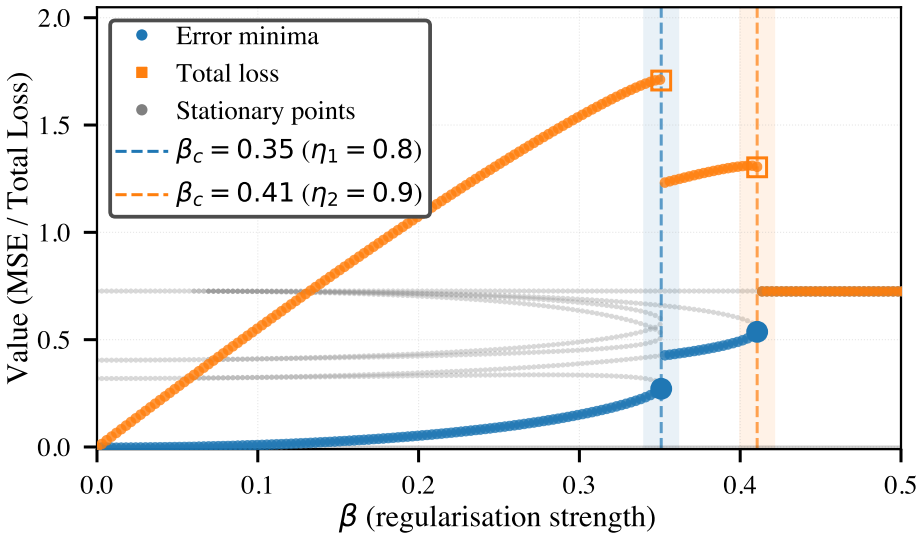


Figure 4: Bifurcation diagram for modes ($\eta_1 = 0.9$, $\eta_2 = 0.8$, $L = 3$). Solid lines: stable stationary points; dashed line: unstable saddle. The two non-trivial branches annihilate at β_c in a saddle-node bifurcation, leaving only $\lambda = 0$ for $\beta > \beta_c$. A model on the rank-1 branch (dotted continuation) remains metastable until stochastic noise drives it across the barrier into the rank-2 global minimum.

Appendix D. Critical Regularization Strength

The stationary condition for mode i under the loss of Eq. (1) is

$$\frac{\partial \mathcal{L}}{\partial \lambda_i} = \lambda_i - \eta_i + \beta\, \lambda_i^{2/L-1} = 0. \tag{7}$$

D.1. Saddle-node bifurcation for DNNs

Setting $k = 2/L$, the stationarity condition reads

$$\lambda_i - \eta_i + \beta\, \lambda_i^{k-1} = 0. \tag{8}$$

For a range of β, this equation admits two distinct positive solutions: a larger root λ^{**} (global minimum) and a smaller root λ^{sad} (unstable saddle). The two roots coalesce at $\beta_c^{(i)}$, requiring both Eq. (8) and its derivative to vanish simultaneously:

$$1 + \beta(k-1)\lambda_i^{k-2} = 0. \tag{9}$$

Solving Eq. (9) gives the critical singular value $\lambda_c^{(i)} = \eta_i(1-k)/(2-k)$, and substituting back into Eq. (8) yields Eq. (3) of the main text.

D.2. Number of transitions

Because the modes decouple, each non-zero singular value η_i of Σ_{yx} has its own $\beta_c^{(i)}$. The ordering $\eta_1 > \cdots > \eta_d > 0$ implies $\beta_c^{(1)} > \cdots > \beta_c^{(d)}$, giving d separate saddle-node bifurcations as β is decreased from a large value.

Appendix E. Theoretical Barrier Heights

For a single mode at $\beta < \beta_c^{(i)}$, the barrier for escape from the metastable lower rank phase is the loss difference between the saddle and the trivial minimum:

$$\Delta E = \mathcal{L}(\lambda^{\text{sad}}) - \mathcal{L}(0) = \tfrac{1}{2}(\lambda^{\text{sad}} - \eta)^2 + \frac{L\beta}{2}(\lambda^{\text{sad}})^{2/L} - \tfrac{1}{2}\eta^2, \tag{10}$$

where λ^{sad} is the smaller positive root of Eq. (8), obtained numerically. At $\beta = 0.32$, $\eta_2 = 0.8$, $L = 3$: $\lambda^{\text{sad}} \approx 0.41$ and $\Delta E \approx 0.003$. This is the minimal barrier along the lowest-loss path out of the metastable phase. The large discrepancy with the experimentally extracted $\Delta E_{\text{eff}} = 0.15 \pm 0.05$ is explained by the high-dimensional corrections derived in Appendix F.

Appendix F. Effective Barrier Height in High-Dimensional Landscapes

F.1. Kramers–Langer Theory

In a parameter space of dimension D, the mean first-passage time from a metastable minimum to a saddle point under Langevin dynamics with noise strength T_{eff} is given by the Kramers–Langer formula [13]:

$$\tau = \frac{2\pi}{|\omega_1^{\text{sad}}|}\left(\frac{\prod_{j=1}^{D}|\omega_j^{\text{sad}}|}{\prod_{j=1}^{D-1}\omega_j^{\min}}\right)^{1/2}\exp\left(\frac{\Delta E}{T_{\text{eff}}}\right), \tag{11}$$

where $\omega_j^{\min} > 0$ are Hessian eigenvalues at the metastable minimum and ω_j^{sad} those at the saddle, with exactly one negative eigenvalue $\omega_1^{\text{sad}} < 0$.

F.2. Effective barrier and entropic contributions

Taking the logarithm of Eq. (11) and defining

$$\Delta E_{\text{eff}} \equiv \Delta E - \frac{T_{\text{eff}}}{2}\sum_{j=1}^{D-1}\ln\frac{\omega_j^{\text{sad}}}{\omega_j^{\min}}, \tag{12}$$

the Arrhenius form is recovered exactly: $\ln\tau = \text{const} + \Delta E_{\text{eff}}/T_{\text{eff}}$. In our setting $D = 170$. The metastable minimum is strongly confining in all directions, while the saddle has broader curvature in the $D-1$ transverse directions, so $\omega_j^{\min} \gg \omega_j^{\text{sad}}$ for most j. The resulting entropic contribution drives $\Delta E_{\text{eff}} \gg \Delta E_{\min}$, explaining the observed factor of ~ 50.

F.3. Correction from multiplicative SGD noise

Mori et al. [19] showed that for MSE loss, the multiplicative nature of SGD noise modifies the relevant barrier from the linear loss difference to a logarithmized quantity. For the approximately quadratic landscape near our metastable minimum, this correction does not alter the linear Arrhenius relationship, consistent with $R^2 = 0.991$.

F.4. Escape time and choice of time unit

Escape times τ are reported in epochs. Since B, η_{lr}, and N are fixed within each sweep, the conversion between epochs and SGD steps is a constant (see Appendix A), leaving the extracted barrier height unchanged.

Appendix G. Generalization to Nonlinear Networks

While our quantitative results (barrier heights, critical β values) are specific to linear networks, the qualitative mechanism—metastable trapping due to saddle-node bifurcations, followed by noise-activated escape—applies generally. Prior work has established that deep nonlinear networks exhibit identical first-order L2-driven phase transitions [3, 6], with hierarchical feature learning [7]. The main ingredient for proposing the hysteresis is the underlying first order phase transition that exists beyond the linear setup. The linear case thus serves as a minimal model that preserves the essential bifurcation structure while remaining analytically tractable.