Survival of the Fittest Representation: A Case Study with Modular Addition

Xiaoman Delores Ding [*1] **Zifan Carl Guo** [*1] **Eric J. Michaud** [1] **Ziming Liu** [1 2] **Max Tegmark** [1 2]

Abstract

When a neural network can learn multiple distinct algorithms to solve a task, how does it "choose" between them during training? To approach this question, we take inspiration from ecology: when multiple species coexist, they eventually reach an equilibrium where some survive while others die out. Analogously, we suggest that a neural network at initialization contains many solutions (representations and algorithms), which compete with each other under pressure from resource constraints, with the "fittest" ultimately prevailing. To investigate this *Survival of the Fittest hypothesis*, we conduct a case study on neural networks performing modular addition, and find that these networks' multiple circular representations at different Fourier frequencies undergo such competitive dynamics, with only a few circles surviving at the end. We find that the frequencies with high initial signals and gradients, the "fittest," are more likely to survive. By increasing the embedding dimension, we also observe more surviving frequencies. Inspired by the Lotka-Volterra equations describing the dynamics between species, we find that the dynamics of the circles can be nicely characterized by a set of linear differential equations. Our results with modular addition show that it is possible to decompose complicated representations into simpler components, along with their basic interactions, to offer insight on the training dynamics of representations.

1. Introduction

The field of *mechanistic interpretability* attempts to reverse-engineer the algorithms that neural networks learn. This involves understanding the representations (features) networks learn (Liu et al., 2022; Zou et al., 2023; Cunningham et al., 2023; Bricken et al., 2023; Marks & Tegmark, 2023; Gurnee & Tegmark, 2024) and how these play a role in larger circuits (Olah et al., 2020; Elhage et al., 2021; Olsson et al., 2022; Nanda et al., 2023; Marks et al., 2024). While most such work studies networks as static objects, some have recently begun to study how network representations and circuits form over training (Liu et al., 2022; Olsson et al., 2022; Nanda et al., 2023; Hoogland et al., 2024; Chen et al., 2023; Singh et al., 2024). In studying training dynamics, we hope to understand not just *what* neural networks learn, but *how* and ultimately *why* they learn the algorithms that they learn. This understanding may eventually be useful for training models with the properties we desire, such as improved efficiency and safety.

One broad question in mechanistic interpretability regards *universality* (Olah et al., 2020): can models consistently learn the same algorithms across different seeds and scales? While some work has found evidence of universality (Olsson et al., 2022; Gould et al., 2024; Gurnee et al., 2024), in other cases there seems to be some variability in algorithms and representations networks learn to solve particular tasks (McCoy et al., 2019; Zhong et al., 2023; Lampinen et al., 2024). In this work, we ask: when networks have a choice between different representations, how do they choose which one to learn?

The question above is hard to answer since representations in the general case are high-dimensional and difficult to disentangle, let alone to understand the dynamics between multiple representations. Therefore, our study focuses on the toy models performing modular addition $a+b=c \pmod p$, a mathematically defined and well studied problem. Prior works have shown that circular representations in the embedding, akin to how numbers are arranged around a clock, are key for modular addition models to generalize (Liu et al., 2022; Power et al., 2022; Liu et al., 2023; Nanda et al., 2023; Zhong et al., 2023), and that the model learns a few such circles that nicely correspond to different Fourier frequen-

[*]Equal contribution [1]Massachusetts Institute of Technology [2]Equal advising. Our code is available at https://github.com/carlguo866/circle-survival. Correspondence to: Zifan Carl Guo <carlguo@mit.edu>.

Accepted to 1st Mechanistic Interpretability Workshop at the International Conference on Machine Learning, Vienna, Austria. 2024. Copyright 2024 by the author(s).

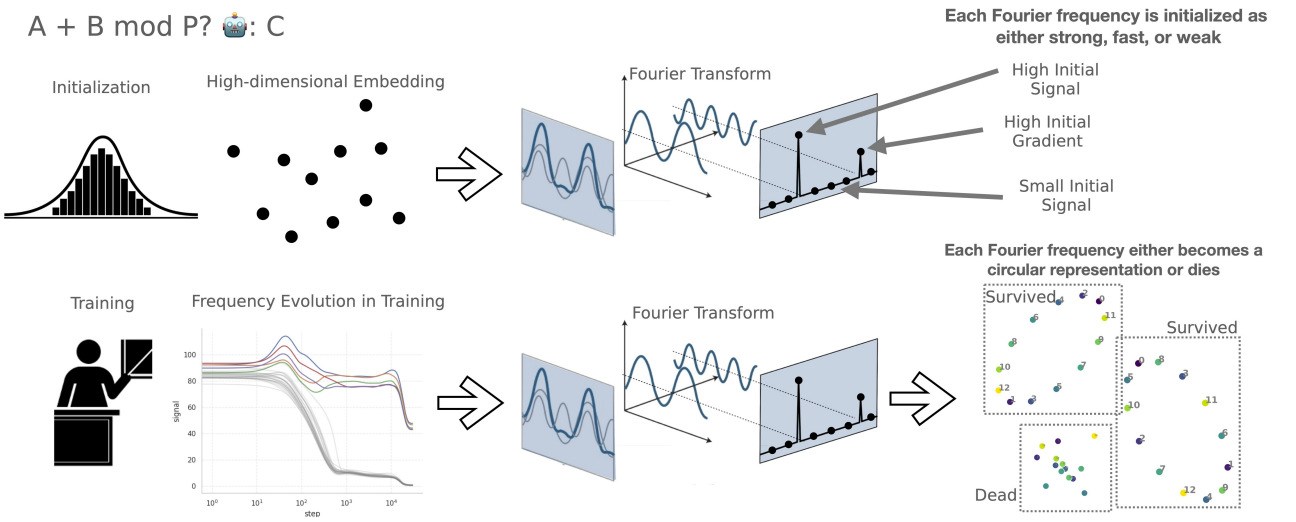

Figure 1. Our experiment method: perform a Fourier transform to the embedding, analyze the initialization, signal evolution, and their effects on the final learned circles over different training runs.

cies (Nanda et al., 2023). Since the embedding is randomly initialized, the projection of the embedding onto the Fourier basis results in roughly similar frequencies, meaning that all possible frequencies have a chance to become the final circle representation. However, only 3-5 circles survive after training, while some frequencies do not form circles, as illustrated in Figure 1. This begs the question: is there any pattern to how these representations form?

To understand the pattern, we propose the *Survival of the Fittest hypothesis* in analogy to ecosystems, where circles of different frequencies can be thought of as "species" competing for a fixed amount of total resources. The two systems are analogous in many aspects: (1) We find surviving circles to have large magnitudes and/or large expanding velocity (gradient) at initialization, supporting the "survival of the fittest" hypothesis. (2) The number of surviving frequencies increases as the resources (embedding dimension) increases. (3) A linear differential equation, inspired by the Lotka–Volterra equations (Alon, 2019) describing interacting dynamics between species, is able to fit the evolution of the magnitude of these circles quite well, even when we push the interaction matrices to be extremely sparse via Lasso regression. Some circle pairs display collaborative behavior, while other pairs are competitive.

Our results show that it is possible to decompose complicated, high-dimensional embeddings into low-dimensional, interpretable representations that also have simple interactions. Our work serves as a proof-of-concept example, demonstrating that this decomposition-style analysis can help understand model training dynamics in more real-world contexts. The paper is organized as follows: In Section 2, we introduce the problem setup, observing that most circles die and only a few survive. In Section 3, we investigate how many circles survive and which circles survive under different circumstances. In Section 4, we show that the dynamics of circles can be well modeled by a linear differential equation. Related works are discussed in Section 5. We conclude in Section 6.

2. Problem Setup

Modular addition We study models performing the task of modular addition in the form of $a + b = c \pmod p$, where $a, b, c = 0, 1, \ldots, p-1$. Our models have an embedding matrix W_E of size (p, d), where every integer $t \in \{0, 1, \cdots, p-1\}$ is treated as a token and has an associated embedding vector $E_t \in \mathbb{R}^d$. The model tokenizes the two inputs a and b, concatenates them, and feeds $[E_a, E_b] \in \mathbb{R}^{2d}$ to a two-layer MLP with two hidden layers of 100 neurons each, producing a categorical output c. We default to $d = 128$ and a large weight decay of 0.5 to make sure the model quickly "groks" (Power et al., 2022) to form the final representation.

Circles and signals Prior work has shown that circular representations (circles) are important for neural networks to perform modular addition (Liu et al., 2022; Nanda et al., 2023; Zhong et al., 2023). However, as shown in Figure 3, neighboring numbers along the circle may have increments other than 1 because there are p equivalent group representations that correspond to circles with different Fourier frequencies $k = 1, 2, \cdots, (p-1)/2$. [1] The circle of frequency k places a token t at $(\cos(2\pi kt/p), \sin(2\pi kt/p))$.

Similar to Nanda et al. (2023), we decompose representa-

[1] Note that frequency k and $p-k$ refer to the same circle, which accounts for the factor 2.

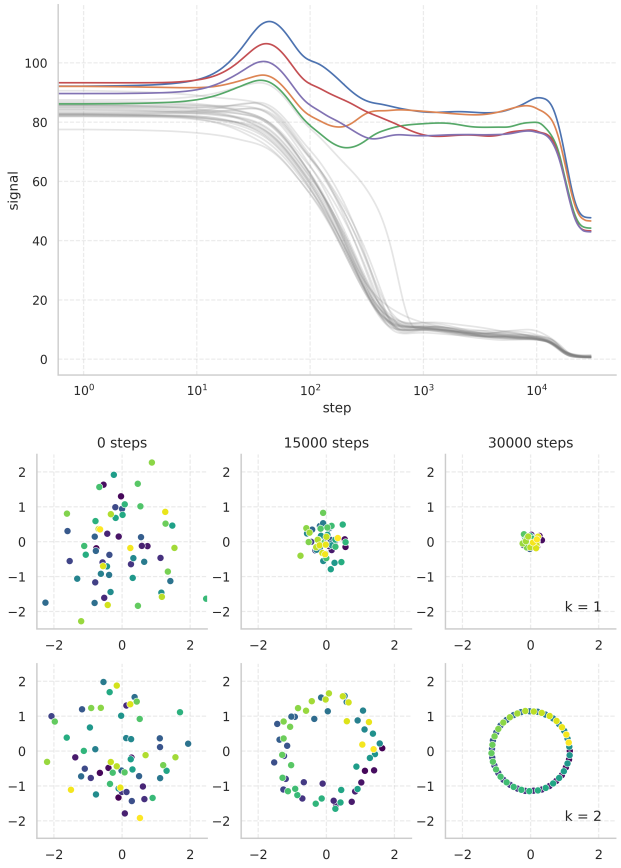

Figure 2. (*Top*) The signals, the magnitude of the Fourier coefficients, of each embedding frequencies over time shown on a logarithmic scale. The surviving frequencies clearly separate themselves from the rest of the frequencies that quickly go to 0. (*Bottom*) Snapshots of the embedding projected onto a dead frequency (top) and a survived frequency (bottom) at different timesteps during training.

tions into a linear combination of circles of different frequencies. Denoting $E_n \in \mathbb{R}^d$ to be the embedding vector for token n, we define the Fourier coefficients of frequency k to be

$$F_k = \sum_{n=0}^{p-1} e^{-i2\pi \frac{k}{p} n} E_n. \tag{1}$$

The circle of frequency k is located on the plane spanned by two vectors $\mathrm{Real}(E_k), \mathrm{Imag}(E_k) \in \mathbb{R}^d$. We define the Fourier signal of frequency k as $\|F_k\|^2 = \sum_{j=0}^{d-1} (F_k^j)^2$ and observe how the signal of each frequency evolves over training steps, shown in Figure 2(Top). We observe that only a few circles have significant signals (hence *survived*) in the end, while the signals of all other circles decay to almost zero (hence *dead*). Shown in Figure 2(Bottom), projections of a surviving frequency stabilize into a clear circle (bottom), while a dead frequency collapses towards its center (top).

While prior works have found circles either with Fourier transformation (Nanda et al., 2023) or with principal component analysis (PCA) (Liu et al., 2022), we use the Fourier transformation since it can reveal circular representations better (see Figure 3). Because each frequency corresponds to a circle, we use the phrase "circle", "circular representations" or "frequencies" interchangeably throughout the whole paper.

Research question: which circles survive? In Figure 2(Top), we observe that the surviving circles in the end tend to have higher initial signal. Can we use this information to predict survived circles? Unfortunately, we find that initially large frequencies do not *always* lead to surviving circles, but large initial signals do lead to *higher probability* of surviving. Therefore, we resort to statistical analysis rather than deterministic analysis by aggregating training results with a fixed embedding and random initializations of MLPs and datasets and report the mean and 95% confidence interval.

3. Survival of the Fittest

In an ecological system with constrained resources, only the fittest will survive. We hypothesize that this theory also holds true in the case of modular addition. Motivated by the ecological analogy, we wish to analyze the competitive dynamics of how the model selects certain circles as its representation by answering **Q1**: how many circles survive? **Q2**: what are the properties of the circles that survive?

3.1. Q1: How many circles survive? A Resource Perspective.

In ecological systems, with more resources available, it is intuitive to think that more species could have a chance to survive. Similarly, in neural networks, the embedding dimension can be made analogous to the total resources available, as a larger model has stronger approximation power and can lead to better performance on the desired task (Sharma & Kaplan, 2020; Michaud et al., 2024; Song et al., 2024). We expect that higher-dimensional embeddings, even just randomly initialized, provide more "resources" for model training. Specifically, in the setting of modular addition, when $d \geq p$, with probability one, perfect circles can be obtained by linearly projecting d-dimensional random representations into suitable subspaces. [2]

[2]Finding a subspace where a perfect circle of frequency k lives on is equivalent to find two linear projections p_1 and p_2 such that $E_a \cdot p_1 = \sin(2\pi ak/p)$ and $E_a \cdot p_2 = \cos(2\pi ak/p)$ for all $a = 0, 1, \cdots, p-1$. Since these equations are linear and linearly independent (due to random initializations), when the number of unknown variables $2d$ is larger (smaller) than the number of equations $2p$, the system is underdetermined (overdetermined), leading to the existence (nonexistence) of solutions.

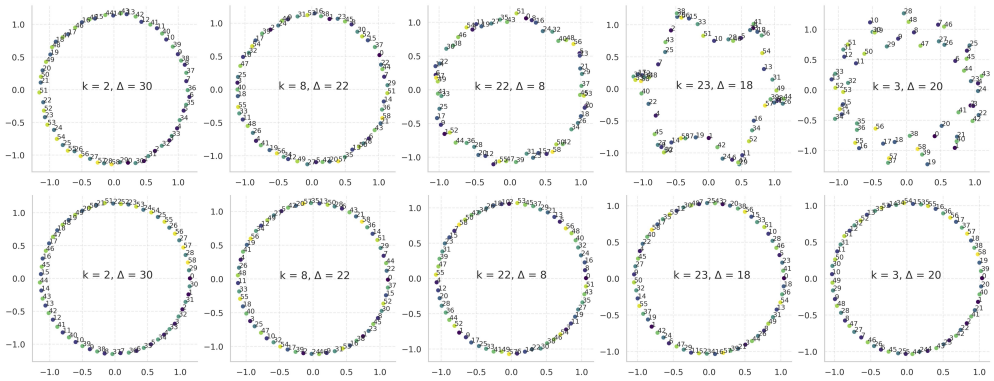

Figure 3. (*Top*) The model embedding projected onto the first 10 principal components in pairs, the only components with significant singular values. (*Bottom*) The model embedding projected onto the Fourier basis of frequencies in descending order of signal magnitude. The Δ between adjacent tokens shows a correspondence between PCA and FFT. Note that Δ can be calculated as the inverse of frequency k modulo p. This indicates that PCA is a loose approximation of circles associated with Fourier frequencies.

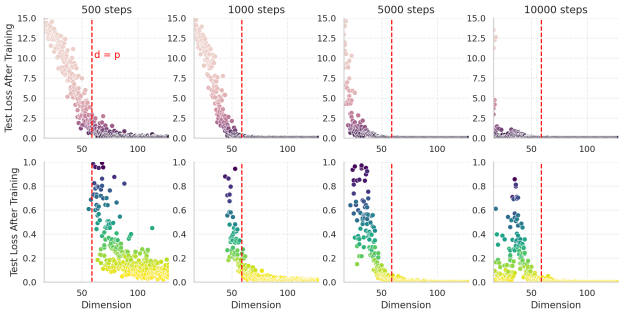

Figure 4. Freezing the initial embedding and training only the MLP, test loss (zoomed in on the *bottom* to < 1.0) in relation to embedding dimension d. One can notice a dip in loss at $d = p$.

Freezing embeddings We confirm with experiments that the initial random embeddings already encapsulate rich representations for MLPs to learn. To verify, we freeze the embeddings at their initialization and train only the MLP. We vary the dimension d of the embedding and observe evolution of test loss. In Figure 4, we show snapshots of the test loss at different timesteps as a function of embedding dimension, where we indeed notice a phase transition around $d = p$—the test loss is sharply better when $d > p$ than when $d < p$. Interestingly, we observe that at 10,000 steps, models with the small ds ($d \leq 20$) have lower loss than models with "medium" ds ($20 \leq d \leq 59$), an intriguing observation similar to the phenomenon of double descent (Nakkiran et al., 2020). Our speculation is: networks with small ds can benefit from feature learning, while networks with large ds ($d \geq p = 59$) have more sheer approximation resources, as we argue above, despite being in a lazy learning regime (Geiger et al., 2020). Networks of medium ds fail perhaps because they are both lazy and resource constrained. A full investigation of this phenomenon is left for future work.

Trainable embeddings Now, we go back to the standard setup where embeddings are trainable. Since dimensionality is analogous to resources in ecological systems, we want to understand if more circles can survive in embeddings of larger dimensions. The answer is yes. As both p and d determine the embedding dimension, we conduct experiments varying one while fixing the other. In Figure 5(Top), we fix d and study the effect of varying p on the expected number of surviving frequencies after sampling over 100 trials, from which we identify clear positive correlation between p and the number of circles the model learns. Similarly, we fix p and vary d to see d's effect on the total number of surviving frequencies in Figure 5(Bottom), which similarly shows an upward curve. As the embedding dimension increases, the number of circles (algorithm redundancy) increases. This redundant mechanism potentially makes neural networks more robust, but less parameter-efficient and interpretable. Understanding this mechanism would be an intriguing topic for future work.

3.2. Q2: Which Circles? Some Are "Fitter" at Initialization.

We want to predict the final surviving circles from the initialization. Following the *Survival of the Fittest hypothesis*, we wish to define a few "fitness" measures that make certain frequencies more likely to survive than others. We indeed observe some properties that make some frequencies more "fit," including large initial signals and large initial gradients. Intuitively, one can think of these "fitness" measures as being born either strong or fast in the ecosystem analogy.

3.2.1. Initial Signal—1st Fitness Measure

We define survival rate as the number of times a particular frequency survives and becomes part of the final represen-

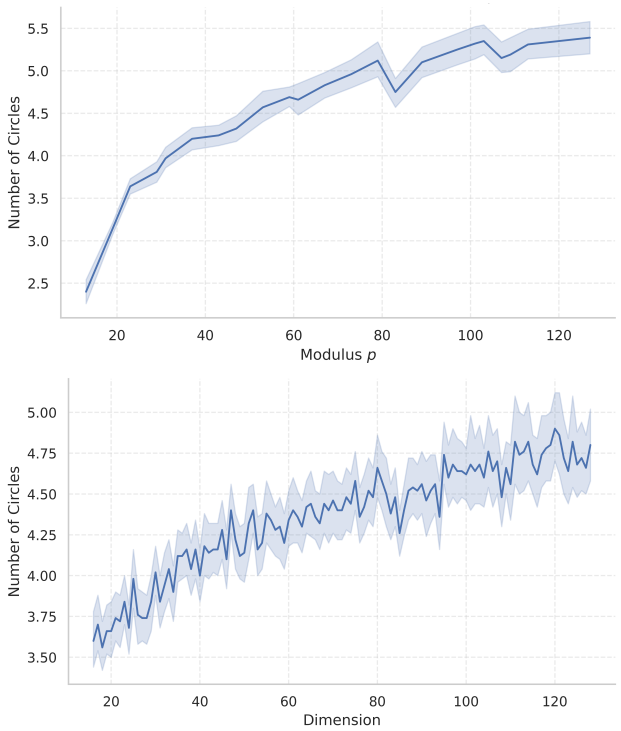

Figure 5. (*Top*): The number of circles the model chooses for its final representation in relation to the number of tokens, p, over 100 random trials. (*Bottom*): The number of circles as the embedding dimension d increases from 16 to 128.

tation over the total number of the trials where we fix the embedding and randomize over different MLPs and dataset initializations. In Figure 6(Left), we show that over 10 different embedding initializations and 50 trials each, establishing a linear correlation: the higher the initial signal, the more likely the model is to choose that frequency as its final representation. To confirm, we find the Pearson correlation between survival rate and initial signal is 0.85 with a p-value smaller than 10^{-3}.

To corroborate, we conduct a perturbation experiment on the initial embedding. Specifically, for a given embedding at initialization, we perform Fourier transform and find the initial coefficients of an arbitrary frequency. We manually enlarge or shrink its magnitude and perform an inverse FFT to restore the embedding, from which we perform model training. We demonstrate in Figure 6(Middle) the survival rate of the perturbed frequency in different initializations as it deviates from the mean of the rest of the signals. We observe that if the frequency is much higher than the rest, survival rate is near 100%, while the frequency rarely survives if it is much lower than the mean. This experiment indeed suggests that as we control the environment much more closely, the initial signal of the frequency plays a unique role in determining the final representation.

To further substantiate our findings, we manually construct an embedding to evaluate survival rate. We first randomly sample 2 out of p frequencies and set their signals to be of a varying ratio $r \in [0, 1]$. Concretely, the largest frequency will have signal magnitude s, while the second highest will have a signal $r \cdot s$. We set all other frequencies to have a signal of small $\epsilon = 10^{-6}$. In this setup, we show in Figure 6(Right) that the frequency with the highest signal will always survive, while the second highest frequency increases in survival rate as its signal increases and differentiates itself further from the rest of the signals. Interestingly, despite all other frequencies having a signal near 0 at initialization, the model sometimes chooses to revive them rather than always choosing the two clear frontrunners, a phenomenon that warrants more investigation in the future.

3.2.2. Initial Gradient—2nd Fitness Measure

In evolution theory, species best adapted to their environment would survive. In neural networks, we hypothesize that not only the representations with high initial signals, but also those can quickly adapts into circles are more likely to survive. This observation motivates us to examine initial expanding velocity (gradient). We simply calculate the gradient as the difference in the signals before and after a given timestep i, taking into account both the embedding gradients and weight decay.

In Figure 7(Left), we show the frequencies' initial gradient values alongside their survival status. Due to the weight decay mechanism, all gradient signals decrease over time, but those with higher initial gradients tend to shrink less and are more likely to survive. In Figure 7(Middle), we show that as the initial gradient increases, the survival rate of those frequencies increases. To analyze the possibly compounding effect, we show that frequencies with both high initial signal and gradient are more likely to survive in Figure 7(Right), as the top right corner is more lit with oranges, indicating more survived frequencies. To further verify the above observation, we train a linear support vector machine (SVM) that separates the dead and survived frequencies, achieving an 83.8% accuracy.

Relation to Lottery Ticket Hypothesis Our analysis can be related to the Lottery Tickets Hypothesis (LTH) (Frankle & Carbin, 2019), where some subnetworks are "winning tickets" that achieve comparable test accuracy to the original network when trained in isolation. Our results suggest that with large embedding dimensions, good circles exist even at initialization, similar to "winning tickets." Our analysis is technically different from LTH in two ways: (1) LTH requires training and pruning to identify "winning" tickets, while circles are mathematically defined without training (but specific to modular addition); (2) LTH only states the existence of "winning" tickets at initialization, while we

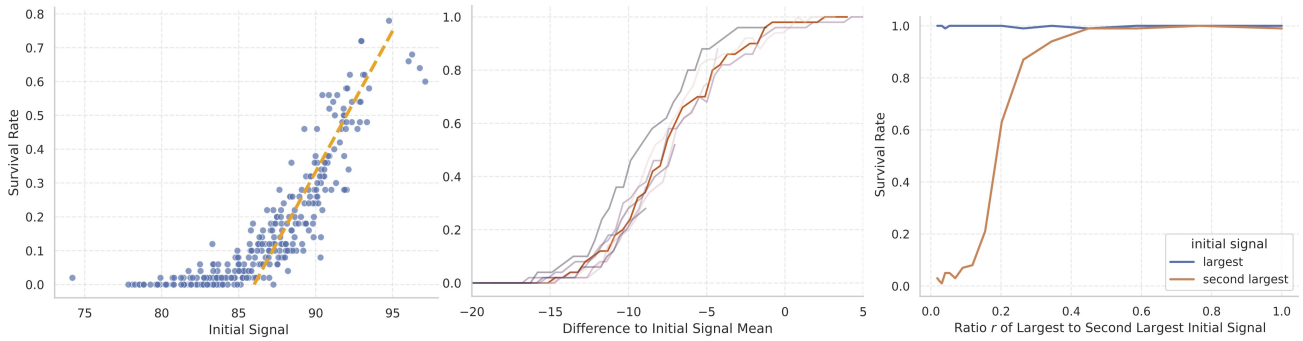

Figure 6. (*Left*) Empirical observation: survival rates of frequencies given their initial Fourier signals over many randomized trials. (*Middle*) Perturbation: survival rate of an arbitrary perturbed frequency versus deviation from the mean initial signal, with different lines showing the same frequency in different embeddings. (*Right*) Manual construction: survival rates of the largest frequency and the 2 largest as they differ by r.

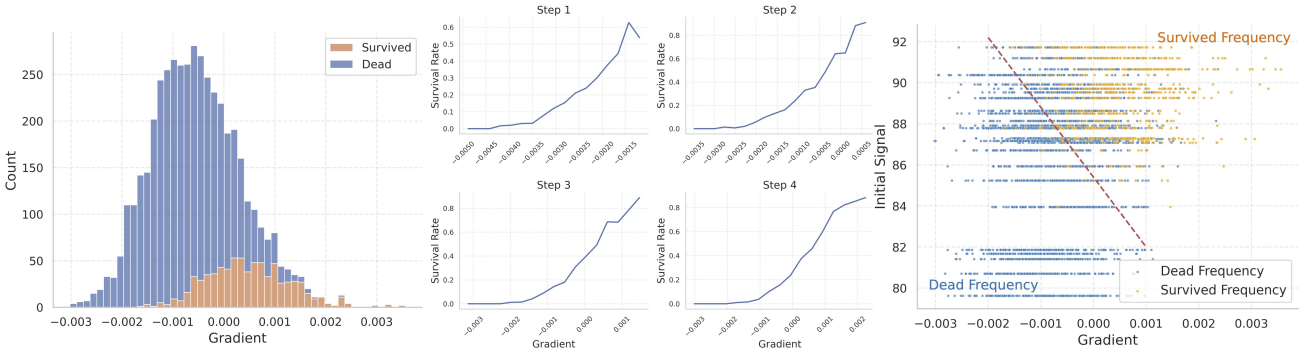

Figure 7. Distribution of the frequency gradients at step 4 with the frequency's survival status. (*Middle*) Survival rate in relation to frequency gradients for step 1 to 4. (*Right*) Survived (orange) and dead (blue) frequencies characterized by their initial signals and gradients at step 4.

manage to characterize the properties of "winning" circles, that, when trained in isolation, have comparable accuracy (see Appendix C). Our work provides representation-level insights to the studies of LTH.

4. Circles Can Collaborate or Compete

In the last section, we demonstrated how *Survival of the Fittest* explains the evolution of circles. However, what is lacking from this explanation is the interaction between circles, considering "fitness" is defined on individual circles. This section seeks to understand circle interaction from two aspects: (1) understand how a group of n circles "collaborate" to reduce losses on the task at hand; (2) understand the (effective) differential equations that govern the evolution of circle signals.

4.1. Circles Have to Collaborate to Reduce Loss

We observed that models naturally form multiple circles in training. Why is this the case? Can't a single circle perform modular addition effectively? Here, we show that a strong cooperative pattern exists between the circles of different frequencies. We investigate this relationship by conducting ablation studies, by manually isolating 1, 2, and 3 different frequencies (removing all other frequencies) and see if these isolated frequencies can solve the task of modular addition. We find that the model is not able to perform the modular addition with only one circle, still has considerably high loss using two circles, and reaches near zero loss with 3 circles. The loss achieved over time in $30,000$ training steps, with 1, 2 and 3 circles respectively, is shown in Figure 8.

4.2. Modeling Circle Dynamics with Differential Equations

In ecology, the Lotka-Volterra equations are famous for using first-order nonlinear differential equations to model the relationship between prey x and predators y (Alon, 2019). They have the following form $\mathrm{d}x/\mathrm{d}t = \alpha x - \beta xy, \mathrm{d}y/\mathrm{d}t = \delta xy - \gamma y$, where x and y represent the population density of prey and predators, respectively, and $\mathrm{d}x/\mathrm{d}t$ and $\mathrm{d}y/\mathrm{d}t$ are the instantaneous growth rates of the two populations. This can be easily generalized to more than two species by

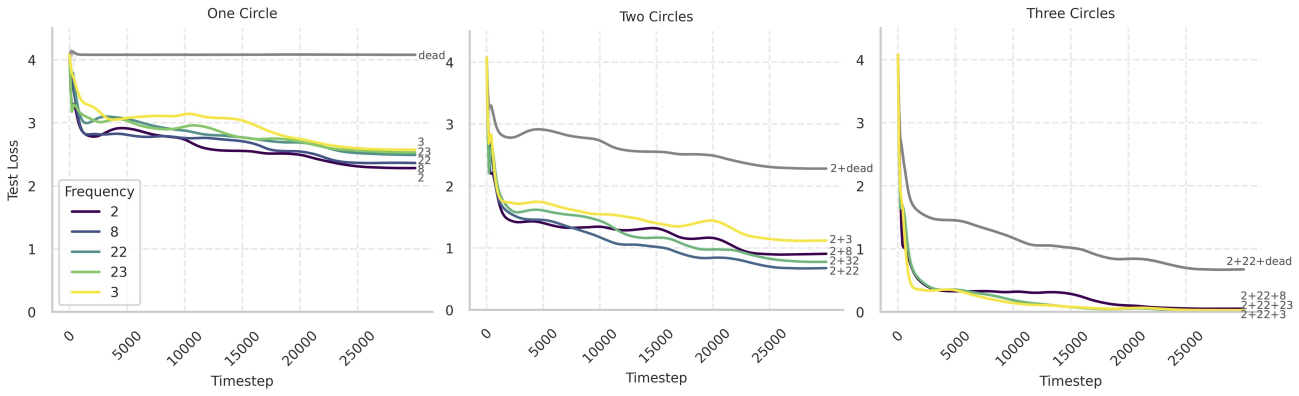

Figure 8. Test loss over training steps in ablation study, with only one (left), two (middle), three (right) circles remaining and other circles removed. Three circles are needed to reduce the loss to zero.

involving linear single-body terms and quadratic two-body terms. Interpreting species population as frequency signals, we have:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{i=1}^{N_c} \alpha_i x_i + \sum_{i=1}^{N_c}\sum_{j=i}^{N_c} \beta_{i,j} x_i x_j, \quad i = 1, 2, \cdots N_c, \tag{2}$$

where x_i represents signals of each frequency.

Although the fit yields an R^2 value close to 1, the model suffers from overfitting due to too many free parameters of $x_i \cdot x_j$ terms. When we attempt to compute a trajectory using our estimates, errors accumulate, leading to a rapid loss of numerical stability. This prompts a natural question: are the quadratic nonlinear terms really necessary?

To our satisfaction, the answer is no. Deviating from the Lokta-Volterra equations, we found an even simpler, linear differential equation, that can estimate the training trajectory well. Removing the second-order terms from Equation 2, we get

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_{j=1}^{N_c} \alpha_{i,j} x_j + b_i, \quad i = 1, 2, \cdots, N_c, \tag{3}$$

or in matrix form $\frac{\mathrm{d}x}{dt} = Ax + b$. With this new set of equations, we can model the trajectory well. We report the R^2 of our fit for both linear regression and Lasso in Figure 9 over many trials with embeddings of varying size, using 80% of the data for training and evaluate the model on the other 20%. For two frequencies in a given trained model, one survived and one dead, we compare the original trajectory, the estimated trajectory from Linear Regression and from Lasso in Figure 9. While linear regression gives us an almost perfect fit at the risk of overfitting, Lasso provides comparable estimations with reasonable errors bounds and extremely sparse coefficients.

As the differential equations are linear, we can find an analytical solution to the ODE system. Assume A as the coefficient matrix of the regression model and b as the intercept, for a given x_0, we have the following solution

$$x(t) = e^{At}x_0 + (e^{At} - I)A^{-1}b. \tag{4}$$

Note that for the coefficient matrix of the Lasso fit, the matrix is extremely sparse and is not naturally invertable, so we have added a small $\epsilon I (\epsilon = 10^{-8})$ to A. The analytical solution similarly provides us with a comparable estimation of the trajectory in Figure 9.

Relation to Neural Tangent Kernel Despite the fact that the Lotka-Volterra equations work well to model ecological dynamics, it is not too surprising that a linear ODE is sufficient to model the training dynamics in neural networks, when the neural networks are wide enough to be characterized by the neural tangent kernel (Jacot et al., 2018). However, our analysis is still novel in the sense that we can disentangle the entire embedding space into individual circles and study their interactions linearly. The idea of decomposition allows the analysis of a complicated system to be broken down to analysis of many simple subsystems and their interactions.

5. Related Work

Mechanistic Interpretability on Algorithmic Tasks A lot of work has been done to reverse engineer how neural networks implement algorithmic tasks (Nanda et al., 2023; Zhong et al., 2023; Liao et al., 2023; Chughtai et al., 2023; Stander et al., 2023; Quirke & Barez, 2024; Quirke et al., 2024) because they are mathematically well-defined and simple. However, even on these toy tasks, neural networks already display some intriguing phenomena, including phase changes during training (Nanda et al., 2023), different algorithms (Zhong et al., 2023; Liao et al., 2023), or show the existence of multiple copies of algorithms (Nanda

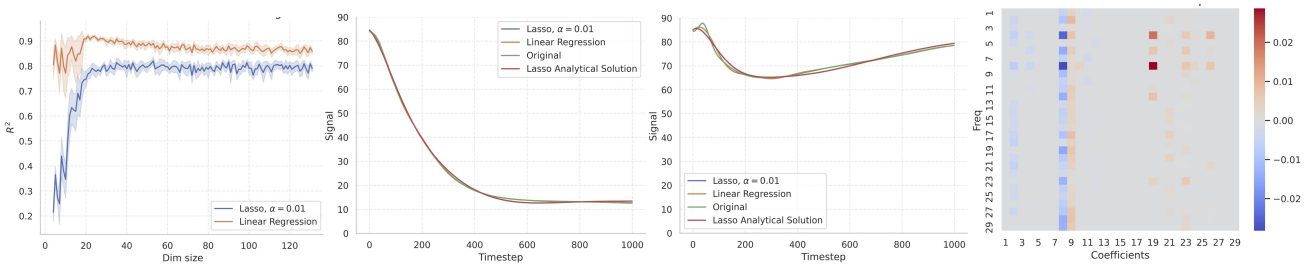

Figure 9. Using linear ODE to model circles' training dynamics. (1) R^2 for fit on embeddings of varying d. (2) Estimated Trajectory for Dead Frequency. (3) Estimated Trajectory for Surviving Frequency. (4) Coefficient Matrix for a representative Lasso fit.

et al., 2023; Zhong et al., 2023).

Training Dynamics Training dynamics strives to understand what happens internally within a model during training. Two most studied phenomena in this area are "grokking" (Power et al., 2022; Thilak et al., 2022; Liu et al., 2022; 2023; Barak et al., 2023) and "double descent" (Nakkiran et al., 2020; Yilmaz & Heckel, 2022; Davies et al., 2023; Schaeffer et al., 2024).Other works have studied training dynamics at various abstraction levels, such as on emerging capabilities level (McCoy et al., 2019; Hoogland et al., 2024), on the circuit level (Olsson et al., 2022; Chen et al., 2023; Singh et al., 2024), and on the neuron level (Quirke et al., 2023). Similar to our analysis of circle interactions using ODEs, previous work attempted to model representation dynamics during training using simple effective dynamics (Liu et al., 2022; Hu et al., 2023; Baek et al., 2024; van Rossem & Saxe, 2024).

Representation Learning Representation learning is key for networks to generalize (Bengio et al., 2013; Le-Khac et al., 2020; Zou et al., 2023; Huh et al., 2024). Many learning paradigms aim to encourage better representations, including weak supervised learning (Zhou, 2018), contrastive learning (Jaiswal et al., 2020; Le-Khac et al., 2020), and Siamese learning (Grill et al., 2020; Chen & He, 2021). Similar to our paper revealing circles on different frequencies, prior works have shown redundant representations and algorithms in models (Doimo et al., 2023; Song et al., 2024) and similarly circular representations in general language models (Engels et al., 2024).

Lottery Ticket Hypothesis The Lottery Ticket Hypothesis (Frankle & Carbin, 2019) posits that some subnetworks–"winning tickets"—identified at initialization and trained in isolation can match the test accuracy of the original, dense network. It has inspired extensions, such as a stronger conjecture on finding such subnetworks without training (Zhou et al., 2019; Ramanujan et al., 2020; Malach et al., 2020; Orseau et al., 2020; Pensia et al., 2020; Diffenderfer & Kailkhura, 2021; da Cunha et al., 2022), transferring winning tickets across setups (Morcos et al., 2019; Chen et al., 2021), and improving methods of pruning to find the subnetworks (Lee et al., 2019; Frankle et al., 2020; Wang et al., 2020; Tanaka et al., 2020; Frankle et al., 2021). Our work relates to LTH that circles can be treated as subnetworks with distinct signals at initialization. In our setting, "winning tickets" exist at initialization and have nice properties like high initial signals, which allow them to eventually become the learned representations.

6. Conclusion

In this paper, we show that the *Survival of the Fittest* theory can explain the training dynamics of the toy modular addition task. Qualitatively, embeddings can be decomposed into circles of different frequencies, deemed as species interacting with one another. Under the resource constraint of model sizes, circles with large signals and gradients are more likely to survive. Quantitatively, the dynamics of circle interaction can be described by a simple linear differential equation. Our results highlight simple laws underlying seemingly complicated representation dynamics and open the door for more fine-grained analysis of representation dynamics for mechanistic interpretability.

Limitations We have focused on a single learning problem: modular addition. Our work studying dynamics between different representations is made possible because representations in modular addition models are well-defined and well-understood. Significant additional work is needed to scale these analyses to even more complex, general models, which remains a challenge.

Broader Impact

Understanding training dynamics contributes broadly to the field of mechanistic interpretability, which allows us to better understand and control neural networks in ways we desire, such as making them more accurate and safe. Neural networks, just like any other dual-use technologies, have accompanying risks, so one should exercise caution when deploying the models that this work contributes to building.

Acknowledgements

We thank Wes Gurnee, Julian Yocum, and Ziqian Zhong for helpful conversations and suggestions. We acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center (Reuther et al., 2018) for providing HPC resources that have contributed to the research results. This work is supported by the Rothberg Family Fund for Cognitive Science, the NSF Graduate Research Fellowship (Grant No.2141064), and IAIFI through NSF grant PHY-2019786.

References

Alon, U. *An introduction to systems biology: design principles of biological circuits*. Chapman and Hall/CRC, 2019.

Baek, D. D., Liu, Z., and Tegmark, M. Geneft: Understanding statics and dynamics of model generalization via effective theory. *arXiv preprint arXiv:2402.05916*, 2024.

Barak, B., Edelman, B. L., Goel, S., Kakade, S., Malach, E., and Zhang, C. Hidden progress in deep learning: Sgd learns parities near the computational limit, 2023.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., Lasenby, R., Wu, Y., Kravec, S., Schiefer, N., Maxwell, T., Joseph, N., Hatfield-Dodds, Z., Tamkin, A., Nguyen, K., McLean, B., Burke, J. E., Hume, T., Carter, S., Henighan, T., and Olah, C. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Chen, A., Schwartz-Ziv, R., Cho, K., Leavitt, M. L., and Saphra, N. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. *arXiv preprint arXiv:2309.07311*, 2023.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Chen, X., Cheng, Y., Wang, S., Gan, Z., Liu, J., and Wang, Z. The elastic lottery ticket hypothesis. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 26609–26621. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/dfccdb8b1cc7e4dab6d33db0fef12b88-Paper.pdf.

Chughtai, B., Chan, L., and Nanda, N. A toy model of universality: reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

da Cunha, A., Natale, E., and Viennot, L. Proving the lottery ticket hypothesis for convolutional neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Vjki79-619-.

Dalvi, F., Sajjad, H., Durrani, N., and Belinkov, Y. Analyzing redundancy in pretrained transformer models. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4908–4926, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.398. URL https://aclanthology.org/2020.emnlp-main.398.

Davies, X., Langosco, L., and Krueger, D. Unifying grokking and double descent. *arXiv preprint arXiv:2303.06173*, 2023.

Diffenderfer, J. and Kailkhura, B. Multi-prize lottery ticket hypothesis: Finding accurate binary neural networks by pruning a randomly weighted network. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=U_mat0b9iv.

Doimo, D., Glielmo, A., Goldt, S., and Laio, A. Redundant representations help generalization in wide neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114011, 2023.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Engels, J., Liao, I., Michaud, E. J., Gurnee, W., and Tegmark, M. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024.

Frankle, J. and Carbin, M. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJl-b3RcF7.

Frankle, J., Dziugaite, G. K., Roy, D. M., and Carbin, M. Stabilizing the lottery ticket hypothesis, 2020.

Frankle, J., Dziugaite, G. K., Roy, D., and Carbin, M. Pruning neural networks at initialization: Why are we missing the mark? In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Ig-VyQc-MLK.

Geiger, M., Spigler, S., Jacot, A., and Wyart, M. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, nov 2020. doi: 10.1088/1742-5468/abc4de. URL https://dx.doi.org/10.1088/1742-5468/abc4de.

Gould, R., Ong, E., Ogden, G., and Conmy, A. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kvcbV8KQsi.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Gurnee, W. and Tegmark, M. Language models represent space and time. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jE8xbmvFin.

Gurnee, W., Horsley, T., Guo, Z. C., Kheirkhah, T. R., Sun, Q., Hathaway, W., Nanda, N., and Bertsimas, D. Universal neurons in gpt2 language models. *arXiv preprint arXiv:2401.12181*, 2024.

Hoogland, J., Wang, G., Farrugia-Roberts, M., Carroll, L., Wei, S., and Murfet, D. The developmental landscape of in-context learning. *arXiv preprint arXiv:2402.02364*, 2024.

Hu, M. Y., Chen, A., Saphra, N., and Cho, K. Latent state models of training dynamics. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=NE2xXWo0LF.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning. *Technologies*, 9(1):2, 2020.

Lampinen, A. K., Chan, S. C., and Hermann, K. Learned feature representations are biased by complexity, learning order, position, and more. *arXiv preprint arXiv:2405.05847*, 2024.

Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive representation learning: A framework and review. *Ieee Access*, 8:193907–193934, 2020.

Lee, N., Ajanthan, T., and Torr, P. SNIP: SINGLE-SHOT NETWORK PRUNING BASED ON CONNECTION SENSITIVITY. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1VZqjAcYX.

Liao, I., Liu, Z., and Tegmark, M. Generating interpretable networks using hypernetworks, 2023.

Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., and Williams, M. Towards understanding grokking: An effective theory of representation learning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=6at6rB3IZm.

Liu, Z., Michaud, E. J., and Tegmark, M. Omnigrok: Grokking beyond algorithmic data. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=zDiHoIWa0q1.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Malach, E., Yehudai, G., Shalev-Schwartz, S., and Shamir, O. Proving the lottery ticket hypothesis: Pruning is all you need. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6682–6691. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/malach20a.html.

Marks, S. and Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024.

McCoy, R. T., Min, J., and Linzen, T. Berts of a feather do not generalize together: Large variability in generalization across models with similar test set performance. *arXiv preprint arXiv:1911.02969*, 2019.

Michaud, E., Liu, Z., Girit, U., and Tegmark, M. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36, 2024.

Morcos, A. S., Yu, H., Paganini, M., and Tian, Y. *One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9XFSbDPmdW.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., Amodei, D., Brown, T., Clark, J., Kaplan, J., McCandlish, S., and Olah, C. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Orseau, L., Hutter, M., and Rivasplata, O. Logarithmic pruning is all you need. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Pensia, A., Rajput, S., Nagle, A., Vishwakarma, H., and Papailiopoulos, D. Optimal lottery tickets via subset-sum: logarithmic over-parameterization is sufficient. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Power, A., Burda, Y., Edwards, H., Babuschkin, I., and Misra, V. Grokking: Generalization beyond overfitting on small algorithmic datasets, 2022.

Quirke, L., Heindrich, L., Gurnee, W., and Nanda, N. Training dynamics of contextual n-grams in language models. *arXiv preprint arXiv:2311.00863*, 2023.

Quirke, P. and Barez, F. Understanding addition in transformers, 2024.

Quirke, P., Neo, C., and Barez, F. Increasing trust in language models through the reuse of verified circuits, 2024.

Ramanujan, V., Wortsman, M., Kembhavi, A., Farhadi, A., and Rastegari, M. What's hidden in a randomly weighted neural network? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11890–11899, 2020. doi: 10.1109/CVPR42600.2020.01191.

Reuther, A., Kepner, J., Byun, C., Samsi, S., Arcand, W., Bestor, D., Bergeron, B., Gadepally, V., Houle, M., Hubbell, M., Jones, M., Klein, A., Milechin, L., Mullen, J., Prout, A., Rosa, A., Yee, C., and Michaleas, P. Interactive supercomputing on 40,000 cores for machine learning and data analysis. In *2018 IEEE High Performance extreme Computing Conference (HPEC)*, pp. 1–6. IEEE, 2018.

Schaeffer, R., Robertson, Z., Boopathy, A., Khona, M., Pistunova, K., Rocks, J. W., Fiete, I. R., Gromov, A., and Koyejo, S. Double descent demystified: Identifying, interpreting & ablating the sources of a deep learning puzzle. In *The Third Blogpost Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=muC7uLvGHr.

Sharma, U. and Kaplan, J. A neural scaling law from the dimension of the data manifold. *arXiv preprint arXiv:2004.10802*, 2020.

Singh, A. K., Moskovitz, T., Hill, F., Chan, S. C. Y., and Saxe, A. M. What needs to go right for an induction head? a mechanistic study of in-context learning circuits and their formation, 2024.

Song, J., Liu, Z., Tegmark, M., and Gore, J. A resource model for neural scaling law. *arXiv preprint arXiv:2402.05164*, 2024.

Stander, D., Yu, Q., Fan, H., and Biderman, S. Grokking group multiplication with cosets, 2023.

Tanaka, H., Kunin, D., Yamins, D. L., and Ganguli, S. Pruning neural networks without any data by iteratively conserving synaptic flow. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6377–6389. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/46a4378f835dc8040c8057beb6a2da52-Paper.pdf.

Thilak, V., Littwin, E., Zhai, S., Saremi, O., Paiss, R., and Susskind, J. The slingshot mechanism: An empirical study of adaptive optimizers and the grokking phenomenon, 2022.

van Rossem, L. and Saxe, A. M. When representations align: Universality in representation learning dynamics. *arXiv preprint arXiv:2402.09142*, 2024.

Wang, C., Zhang, G., and Grosse, R. Picking winning tickets before training by preserving gradient flow. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgsACVKPH.

Yilmaz, F. F. and Heckel, R. Regularization-wise double descent: Why it occurs and how to eliminate it. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pp. 426–431. IEEE Press, 2022. doi: 10.1109/ISIT50566.2022.9834569. URL https://doi.org/10.1109/ISIT50566.2022.9834569.

Zhong, Z., Liu, Z., Tegmark, M., and Andreas, J. The clock and the pizza: Two stories in mechanistic explanation of neural networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S5wmbQc1We.

Zhou, H., Lan, J., Liu, R., and Yosinski, J. Deconstructing lottery tickets: Zeros, signs, and the supermask. In *Advances in Neural Information Processing Systems*, 2019.

Zhou, Z.-H. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M. J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J. Z., and Hendrycks, D. Representation engineering: A top-down approach to ai transparency, 2023.

A. Weight Decay

In the main paper, we have estimated how dimensionality is analogous to total resources in an ecosystem, while weight decay represents the resource constraint in the environment that decay at each time step.

Considering the extreme case where weight decay is zero: there is no resource limit and it is not surprising that all the frequencies survive while the neural network fails to generalize, as it has trouble "grokking" (Liu et al., 2022; 2023). As weight decay slightly increases, the different frequencies pose a competitive dynamic against each other. The number of final circular representations gradually drops, as illustrated in Figure 10.

B. Circularity

In addition to the signal magnitude metric used throughout the paper, we compute another metric, circularity, to analyze the initial embedding. We modify the metric introduced in Zhong et al. (2023) by calculating through each frequency in the Fourier Basis instead of through the principal components. As established in the paper, if $X_0, ..., X_{p-1} \in \mathbb{R}^d$ are embeddings projected onto the Fourier basis, the circularity for a specific frequency f is defined as

$$c_k = \frac{2}{p\sum_{j=0}^{p-1} X_{k,j}^2} \left| \sum_{j=0}^{p-1} X_{k,j} e^{2\pi \mathbf{i} \cdot jk/p} \right|^2 \tag{5}$$

Using this metric, we conduct two experiments: one to measure the circularity of the embedding as its dimensionality varies and another to see if initial circularity plays a role in informing eventual representations.

To investigate the impact d exerts on the system, we randomly sample 50 initial embeddings of different dimensions and compute the mean and maximum circularity among all frequencies, as shown in Figure 11(Left). Indeed, circularity increases as dimensionality increases, confirming our observation that higher-dimensional embedding encodes more complex information at initialization.

We also aim to verify the hypothesis that if the embedding is initialized closer to a circle on a given frequency, that frequency is more likely to survive. However, the evidence shown in Figure 11(Right) is inconclusive as to whether larger initial circularity implies a better chance of survival.

C. Forcing Model to Learn < 3 Circles

In our training with an embedding of reasonably large size, such as $(p, d) = (59, 128)$, the model almost always chooses three or more circles as its final representations. As illustrated in Figure 8, these circles cooperate to solve the modular addition task instead of acting on their own. However, one would notice that if the model were using the Clock algorithm (Nanda et al., 2023), a single circle would be sufficient to solve the task perfectly. The observation naturally evokes the following questions: why does the model choose to learn multiple circles, and can we 'force' it to learn fewer than three circles under some contrived conditions?

To restrict the representations the model can learn, we manually ablate the initial embedding so that only one or two frequencies have non-zero signals on the Fourier basis.

Specifically, for a given embedding, we first train without any ablation, from which we identify the original circles the model chooses to learn. We then use k_1 to denote the frequency with largest signal after training and k_2 to denote the second largest frequency.

Using this information, we conduct four ablation experiments:

Experiment A At initialization, project the embedding onto the Fourier basis, set all frequencies except for k_1 to 0, and use inverse FFT to reconstruct the embedding. Train the model using this initial embedding.

Experiment B At initialization, use the same ablation procedure as above, but suppress all other frequencies except for k_1 and k_2 to 0.

Experiment C At initialization, randomly select a frequency $k_{r,1}$ and suppress all other frequencies except for $k_{r,1}$.

Experiment D At initialization, randomly select two frequencies $k_{r,1}$ and $k_{r,2}$ and suppress all other frequencies except for those two.

In Experiment A, although the embedding is initialized with only one frequency with significant signal, the model revives some frequencies with an originally 0 signal to form circular structures with strong signals, and eventually ends up with four learned circles. Figure 12(Left) shows the test loss for Experiment A in *orange*. We can infer from the loss curve that despite trying to learn with only one circle, the model struggles to achieve lower loss with this simple representation and has to make its representation more complex over time, leading to the periodic spikes in test loss. In none of our experiments were we able to construct a model that naturally forms a single-circle representation.

However, in Experiment B, the model achieves good performance using only the two circles initialized with non-zero signals. Two circles seem sufficient for the model to achieve a loss as low as $1e^{-7}$, if the two circles are initalized well and we force the model to only use two. Therefore, we suspect that three circles are not necessary but rather a model choice to prefer redundant representations (Dalvi

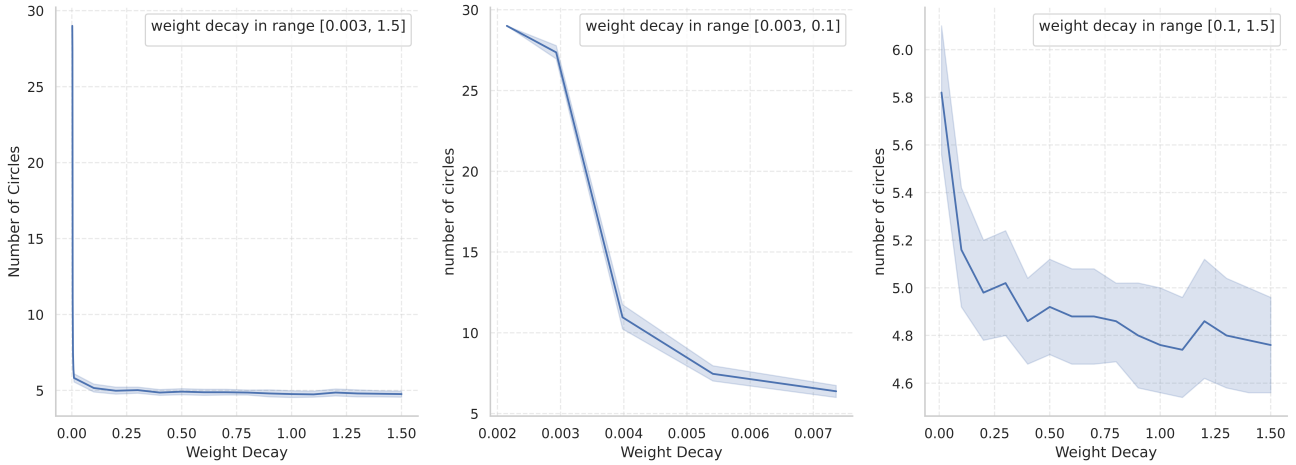

Figure 10. Number of circles survived as a function of weight decay. The *left* panel displays the complete range of weight decays tested in our experiments; the *middle* focuses on smaller weight decays, while the *right* illustrates the transition in the number of surviving circles at larger weight decays.

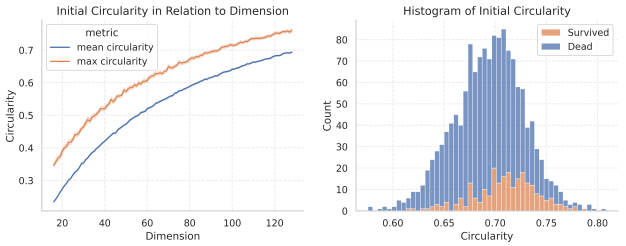

Figure 11. (*Left*) Circularity of projections of the embedding onto different Fourier frequencies at initialization as dimension d varies, calculated using Equation 5. (*Right*) Histogram of different frequencies by their gradients, along with their survival status.

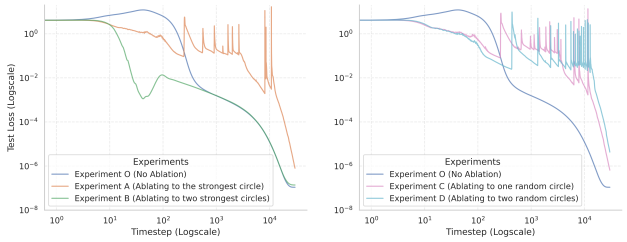

Figure 12. (*Left*) Test loss curve over training timesteps of the ablation experiment when keeping all frequencies (blue), the largest frequency (orange), and the largest two frequencies (green). (*Right*) Test loss curve of the ablation experiment when keeping all frequencies (blue), one random frequency (pink), and two random frequencies (cyan).

et al., 2020). Interestingly, Experiment B reaches lower test loss more quickly than the original, mainly because ablating to have only the strongest two frequencies already serves as a type of training and makes the model training process easier, similar to the findings of Zhou et al. (2019).

In comparison, Figure 12(Right) shows that the model struggles to learn with only 1 or 2 random frequencies; after training, at least three circles survive. Summarizing these results, we conclude that the choice of the circles are not arbitrary: the model can learn the task well with only the two most fitted frequencies but not two random frequencies. Here, we further corroborate that the embedding initialization plays a significant role in the model's preference for its circle representations.

Note that in experiments A, C, and D, the test loss curve exhibits several cycles of spiking and subsequent decay. This slingshot phenomenon is associated with the use of the Adam optimizer and often co-occurs with grokking (Thilak et al., 2022). We do not discuss this further, as it falls outside the scope of our research.

D. Embedding Gradients

In Section 3.2.2, we approximate the gradient as the difference of signals before and after a given timestep. We provide further justification here.

The actual gradient on the embeddings consists of two parts: the gradient $\frac{\partial \mathcal{L}}{\partial E_k}$ produced by MLP on the embedding through backpropagation and weight decay of the initial data. To understand the effect of the former on frequency signals, we use the same Fourier transform procedure to transform the gradients to the Fourier basis and compute their norm, as the Fourier transform is a linear transformation. In Figure 13, we visualize the norm of the gradient in Fourier basis over time. The Fourier gradient spikes around 100 steps and quickly diminishes to near zero after 1000 steps. After this point, the gradient becomes negligibly small, and weight

decay becomes the dominant factor affecting the evolution of signals.

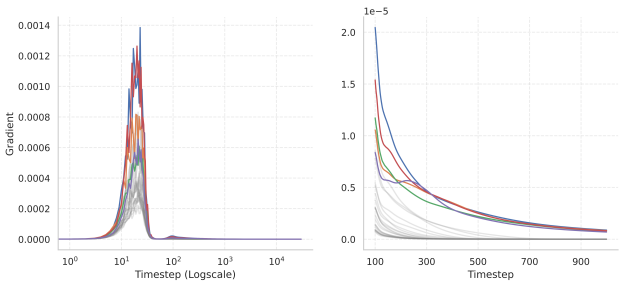

Figure 13. The embedding gradients through backpropagation projected onto the Fourier basis over time. The *right* zooms into timesteps 100 to 1000. The colored curves denote frequencies that eventually survived, while the grey ones represent gradients of dead frequencies.

These observations motivate two experimental decisions in our paper.

First, because it is difficult to reconstruct the effect of both backpropagation and weight decay compounded on top of each other on the Fourier basis, especially when weight decay dominates the gradient, we think the difference in signal is a simple and sufficient proxy to conduct experiments with in Section 3.2.2.

Second, since the embedding gradient becomes negligibly small after 1000 steps, we only use data from the first 1000 steps to fit the linear ODE system in Section 4. More datapoints after the first 1000 steps will only allow the regression model to capture the dynamics of weight decay, which distracts from our study of the dynamics between representations.

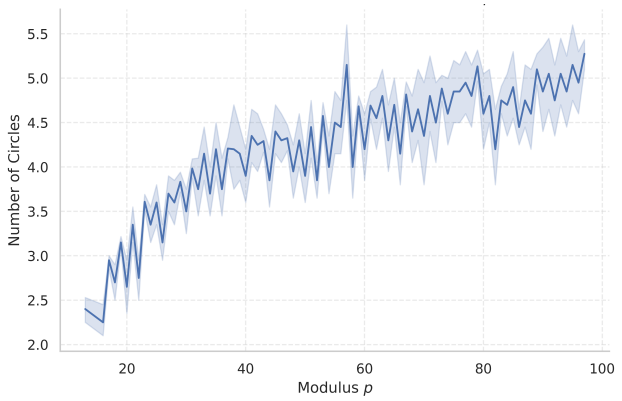

Figure 14. Number of survived circles over random trials as a function of p, where p ranges from 17 to 97.

E. Training Details

As discussed in Section 2, the model we trained for the modular addition task features a simple embedding-MLP architecture. The embedding has a size of (p, d), where p represents the modulus and d stands for the dimension. By default, we set $p = 59$ and $d = 128$. In our experiments, we vary p and d to study their effects on model's learned embeddings.

The MLP in our model consists of two layers: an input layer with dimension $2d$, two hidden layers, each with a width of 100, and an output layer with dimension p, which represents the logits for each token from 0 to $p - 1$.

Both the embedding and the MLP are initialized from a Gaussian distribution with $\mu = 0$ and $\sigma = 1$. We train the model using the AdamW optimizer (Loshchilov & Hutter, 2019), with a learning rate set to 0.01. The training loss is defined as the cross-entropy loss between the logits computed by the model and the ground truth. To encourage model generalization, we use a train-test split of 80-20 and apply a default weight decay of 0.5. Additionally, we experiment with other weight decay values to study their impact.

In each experiment, the model is trained for 3×10^4 steps.

F. Non-prime Modulus p

In Section 3.1, we only provide results for prime modulus p since non-prime p behaves differently in the modular addition task due to their non-trivial factors. For example, when $p = 12$, a circle with delta $\Delta = 2$ does not cover all the numbers in $[0, p - 1]$, and our previous analysis in Fourier basis no longer holds true. However, we present supplementary results on the effect of the number of surviving circles for all possible moduli in Figure 14.Compared with Figure 5(Left), one can observe more variation in Figure 14, but an upward trend can still be observed.

G. More ODE Approximated Trajectories

In Figure 9, we have shown the trajectories of two representative frequencies approximated by our linear ODE, one dead and one survived, and one representative coefficients heatmap. In Figure 15, we show the sparse coefficients heatmap for 3 more ODE fits with Lasso. In Figure 16, we report predicted trajectories for all 29 frequencies in one training run.

H. Experiments Compute Resource

All the model training is performed on NVIDIA V100 GPUs. We provide the GPU specs as follows:

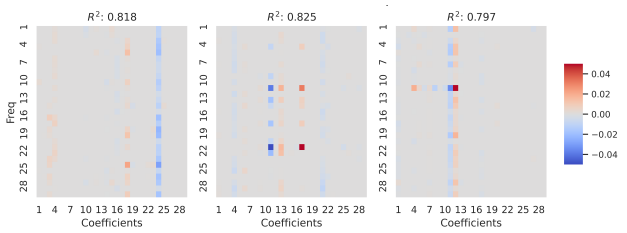

Figure 15. Coefficients heatmaps for 3 more Lasso fits of the embeddings' training dynamics.

- Processor: Intel Xeon Gold 6248
- Nodes: 224
- Clock Rate: 2.5GHz
- CPU cores: 40
- Node RAM: 384GB
- RAM per core: 9GB
- Accelerator type: Nvidia Volta V100
- Accelerators(per Node): 2
- Accelerator RAM: 32GB

The GPU days needed for each experiment are:

- Initial gradient/gradient experiment in 3.2: 1 GPU day
- Freeze embedding experiment in Section 3.1: 0.8 GPU days
- Varying modulus p experiment in Section 3.1: 15 GPU days
- Varying dimension experiment in Section 3.1: 9 GPU days
- Initial signal perturbation experiment in Section 3.2.1: 15 GPU days
- Manual construction experiment in Section 3.2.1: 0.5 GPU days
- Varying weight experiment in Appendix A: 2 GPU days
- Other small-scale experiments: 1 GPU day

Overall, the experiments compute resource adds up to about 45 GPU days. The full research project does not require more compute than the experiments reported in the paper.

I. Acknowledgement for Online Assets Used

We have utilized several online assets in creating Figure 1. The original illustration of the Fourier Transform can be accessed here. The icons are sourced from The Noun Project and Clker.

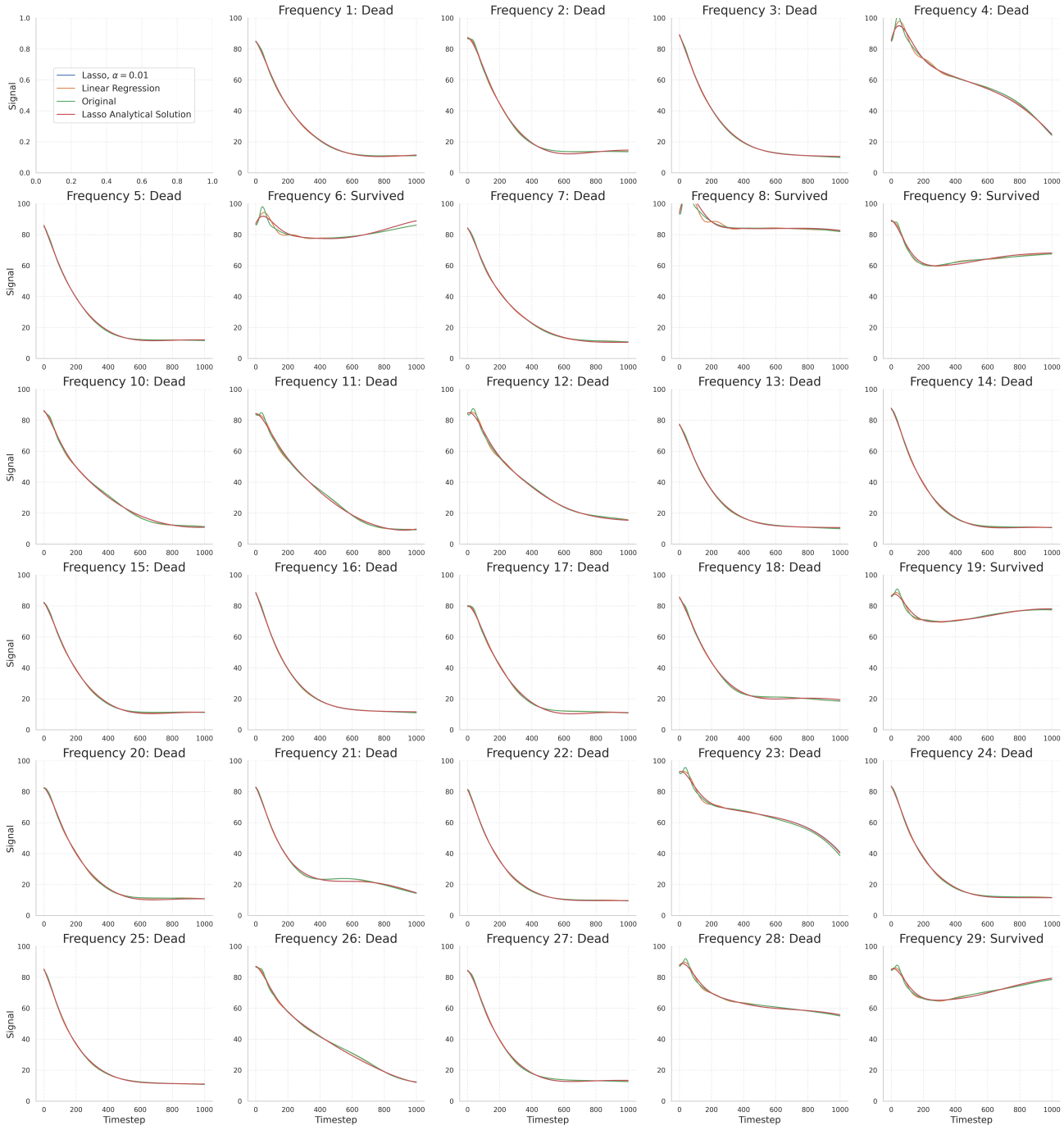

Figure 16. Original and estimated trajectories for signal evolution of all 29 frequencies during training.