On the Limits of Linear Representation Hypotheses in Large Language Models: A Dynamical Systems Analysis

Abhinav Muraleedharan
University of Toronto
abhi@cs.toronto.edu

Abstract

Linear representation hypotheses and steering vector control methods are increasingly popular in mechanistic interpretability, suggesting that perturbations in latent space yield predictable changes in model behavior. We provide a rigorous theoretical critique of this perspective by analyzing the chaotic dynamics inherent in deep residual networks through the lens of dynamical systems theory. We prove that two latent vectors which are initially ϵ-close can diverge exponentially within $O(\log(1/\epsilon))$ layers under positive Lyapunov exponents, fundamentally undermining the assumption that linear operations in latent space reliably control model outputs. Our analysis reveals that the exponential sensitivity to initial conditions characteristic of chaotic systems makes linear approximations inherently unreliable in deep networks, providing a theoretical foundation for understanding the limitations of current interpretability methods.

1 Introduction

Mechanistic interpretability seeks to unravel the internal computations of neural networks by providing structured explanations for their behavior [7]. A prominent paradigm in this field is the *linear representation hypothesis*, which posits that semantic features and controllable behaviors correspond to linear directions in the network's latent space. This assumption underlies numerous interpretability techniques, including steering vectors [6], concept activation vectors [5], and linear probing methods [1]. While these linear methods have shown empirical success in controlled settings, their theoretical foundations remain poorly understood, particularly for the deep, highly nonlinear architectures that characterize modern large language models (LLMs). The assumption that local linearity persists across dozens of layers warrants rigorous theoretical scrutiny from the perspective of dynamical systems theory.

The key insight driving our analysis is that deep residual networks can be viewed as discrete-time dynamical systems, and such systems often exhibit chaotic behavior characterized by sensitive dependence on initial conditions. When the dynamics are chaotic, as evidenced by positive Lyapunov exponents, small perturbations grow exponentially, making long-term prediction and control fundamentally difficult. Furthermore, in such systems it becomes difficult to map directions in the latent space to concepts. For instance, suppose one define a direction $\vec{e_g}$ in the latent space as a 'good direction'. Another latent vector which is ϵ close to this 'good vector' would produce totally different logit distribution, since nonlinear dynamics inherent in these networks would be able to amplify even minute deviations. A control technique trying to detect the model's "goodness" along this direction would therefore be unstable: infinitesimal deviations would be magnified into qualitatively distinct downstream behaviors. This instability challenges the core assumption that latent directions remain semantically meaningful as they propagate through depth. Our results formalize this phenomenon

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Mechanistic Interpretability.

and show that, beyond shallow layers, the notion of a globally coherent linear semantic direction becomes mathematically untenable.

In this work, we provide a comprehensive dynamical systems analysis of linear representation hypotheses in residual networks. Our main contribution establishes that ϵ-close initial perturbations diverge exponentially, with separation occurring within $O(\log(1/\epsilon))$ layers under chaotic dynamics. This result fundamentally challenges the assumption that small perturbations remain predictably small throughout deep networks, revealing the inherent instability of linear approximations in chaotic dynamical systems. We then analyze the implications of this exponential divergence for interpretability methods and suggest directions for developing chaos-aware approaches. Our analysis reveals fundamental limitations of linear interpretability methods and provides a theoretical foundation for understanding why these techniques often produce inconsistent results in large-scale networks.

2 Mathematical Framework

2.1 Residual Networks as Dynamical Systems

Consider a residual network with L layers, where each layer applies a transformation of the form:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{F}_k(\mathbf{x}_k), \quad k = 0, 1, \dots, L-1 \tag{1}$$

This discrete dynamical system can be viewed as a forward Euler discretization of the continuous-time ODE:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{F}(t, \mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0 \tag{2}$$

where $\mathbf{F}(t, \mathbf{x})$ interpolates the discrete transformations $\mathbf{F}_k(\mathbf{x})$. This dynamical systems perspective enables us to apply tools from chaos theory, particularly Lyapunov exponent analysis, to understand the stability and predictability of trajectories.

Definition 1 (Lyapunov Exponents)**.** *For the linearized dynamics around a trajectory $\{\mathbf{x}_k\}$, the* Lyapunov exponents $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ *characterize the exponential growth rates of infinitesimal perturbations. Formally, they are defined as:*

$$\lambda_i = \lim_{k \to \infty} \frac{1}{k} \log \sigma_i \left(\prod_{j=0}^{k-1} \mathbf{J}_j \right) \tag{3}$$

where $\mathbf{J}_j = \mathbf{I} + \nabla \mathbf{F}_j(\mathbf{x}_j)$ is the Jacobian matrix at step j, and σ_i denotes the i-th singular value of the matrix product.

Definition 2 (Chaotic Dynamics)**.** *A dynamical system exhibits* chaotic behavior *if it has sensitive dependence on initial conditions, characterized by at least one positive Lyapunov exponent $\lambda_{\max} > 0$. This means that initially nearby trajectories separate exponentially at rate $\lambda_{\max}$.*

3 Main Theoretical Results

3.1 Exponential Divergence of Close Trajectories

Our main result rigorously establishes the exponential divergence of initially close perturbations in chaotic residual networks.

Theorem 1 (Exponential Divergence in Chaotic Residual Networks)**.** *Let $\{\mathbf{x}_k\}_{k \geq 0}$ and $\{\mathbf{y}_k\}_{k \geq 0}$ be trajectories of the residual network evolution:*

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{F}_k(\mathbf{x}_k) \tag{4}$$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \mathbf{F}_k(\mathbf{y}_k) \tag{5}$$

Assume that $\mathbf{F}_k$ is continuously differentiable and that the system exhibits chaotic dynamics with maximal Lyapunov exponent $\lambda_{\max} > 0$. Then for sufficiently small initial separation $\epsilon = \|\mathbf{x}_0 - \mathbf{y}_0\|$, there exist constants $C_1, C_2 > 0$ such that:

$$C_1 \epsilon e^{\lambda_{\max} k} \leq \|\mathbf{x}_k - \mathbf{y}_k\| \leq C_2 \epsilon e^{\lambda_{\max} k} \tag{6}$$

for sufficiently large k. In particular, $\|\mathbf{x}_k - \mathbf{y}_k\| = \Omega(1)$ once

$$k \geq \frac{\log(1/\epsilon)}{\lambda_{\max}} + O(1) \tag{7}$$

Proof. Let $\boldsymbol{\delta}_k = \mathbf{y}_k - \mathbf{x}_k$ denote the perturbation vector at step k. The evolution of this perturbation is governed by:

$$\boldsymbol{\delta}_{k+1} = \boldsymbol{\delta}_k + \mathbf{F}_k(\mathbf{y}_k) - \mathbf{F}_k(\mathbf{x}_k) \tag{8}$$

Since $\mathbf{F}_k$ is continuously differentiable, we can write the difference, $\mathbf{F}_k(\mathbf{y}_k) - \mathbf{F}_k(\mathbf{x}_k)$ as :

$$\begin{aligned}
\mathbf{F}_k(\mathbf{y}_k) - \mathbf{F}_k(\mathbf{x}_k) &= \mathbf{F}_k(\mathbf{x}_k + \boldsymbol{\delta}_k) - \mathbf{F}_k(\mathbf{x}_k) && \text{(define perturbation } \boldsymbol{\delta}_k = \mathbf{y}_k - \mathbf{x}_k\text{)} && (9)\\
&= \mathbf{F}_k(\mathbf{z}(1)) - \mathbf{F}_k(\mathbf{z}(0)) && \text{(define path } \mathbf{z}(t) = \mathbf{x}_k + t\boldsymbol{\delta}_k\text{)} && (10)\\
&= \int_0^1 \frac{d}{dt}\mathbf{F}_k(\mathbf{z}(t))\,dt && \text{(fundamental theorem of calculus in 1D for } t\text{)} && (11)\\
&= \int_0^1 \nabla\mathbf{F}_k(\mathbf{z}(t))\frac{d\mathbf{z}}{dt}\,dt && \text{(chain rule)} && (12)\\
&= \int_0^1 \nabla\mathbf{F}_k(\mathbf{x}_k + t\boldsymbol{\delta}_k)\,\boldsymbol{\delta}_k\,dt && \text{(since } \frac{d\mathbf{z}}{dt} = \boldsymbol{\delta}_k\text{).} && (13)
\end{aligned}$$

$$\mathbf{F}_k(\mathbf{y}_k) - \mathbf{F}_k(\mathbf{x}_k) = \int_0^1 \nabla\mathbf{F}_k\big(\mathbf{x}_k + t\boldsymbol{\delta}_k\big)\,\boldsymbol{\delta}_k\,dt.$$

Thus,

$$\boldsymbol{\delta}_{k+1} = \left(\mathbf{I} + \int_0^1 \nabla\mathbf{F}_k(\mathbf{x}_k + t\boldsymbol{\delta}_k)\,dt\right)\boldsymbol{\delta}_k \equiv \mathbf{A}_k\boldsymbol{\delta}_k.$$

For sufficiently small $\|\boldsymbol{\delta}_k\|$, the integral is close to the Jacobian along the unperturbed trajectory. Let:

$$\mathbf{A}_k \approx \mathbf{I} + \nabla\mathbf{F}_k(\mathbf{x}_k).$$

And we have:

$$\boldsymbol{\delta}_{k+1} = (\mathbf{I} + \nabla\mathbf{F}_k(\boldsymbol{\xi}_k))\boldsymbol{\delta}_k = \mathbf{A}_k\boldsymbol{\delta}_k \tag{14}$$

where $\mathbf{A}_k = \mathbf{I} + \nabla\mathbf{F}_k(\boldsymbol{\xi}_k)$. By iteration:

$$\boldsymbol{\delta}_k = \left(\prod_{j=0}^{k-1}\mathbf{A}_j\right)\boldsymbol{\delta}_0 \tag{15}$$

The key insight is that the matrices $\mathbf{A}_j$ approximate the Jacobians along the unperturbed trajectory when $\|\boldsymbol{\delta}_j\|$ is small. By Oseledets' multiplicative ergodic theorem [8], for almost every initial direction $\boldsymbol{\delta}_0/\|\boldsymbol{\delta}_0\|$, we have:

$$\lim_{k\to\infty}\frac{1}{k}\log\left\|\prod_{j=0}^{k-1}\mathbf{A}_j\frac{\boldsymbol{\delta}_0}{\|\boldsymbol{\delta}_0\|}\right\| = \lambda_i \tag{16}$$

for some Lyapunov exponent λ_i. For the maximal Lyapunov exponent $\lambda_{\max}$, there exists a set of initial directions of positive measure such that:

$$\|\boldsymbol{\delta}_k\| \asymp \|\boldsymbol{\delta}_0\|e^{\lambda_{\max}k} = \epsilon e^{\lambda_{\max}k} \tag{17}$$

The separation becomes $O(1)$ when $\epsilon e^{\lambda_{\max}k} = O(1)$, which occurs for $k = O(\log(1/\epsilon)/\lambda_{\max})$. □

3.2 Implications for Linear Control

The exponential divergence established in Theorem 1 has implications for the reliability of linear control methods.

Corollary 1 (Breakdown of Linear Steering)**.** *Consider a steering intervention that applies a perturbation $\boldsymbol{\delta}_0$ with $\|\boldsymbol{\delta}_0\| = \epsilon$ at the input layer. Under chaotic dynamics, the steering effect becomes unpredictable after approximately*

$$L_{\textit{chaos}} = \frac{\log(1/\epsilon)}{\lambda_{\max}} \tag{18}$$

layers, where predictability is lost once the perturbation magnitude becomes $O(1)$.

Remark 1 (Practical Implications)**.** *For typical steering magnitudes $\epsilon \sim 10^{-3}$ to 10^{-1} and observed Lyapunov exponents $\lambda_{\max} \sim 0.1$ to 1.0 in deep networks, the chaos horizon $L_{\textit{chaos}}$ ranges from approximately 2 to 70 layers. This suggests that linear steering may be unreliable in networks deeper than a few dozen layers when operating in chaotic regimes.*

3.3 Lyapunov Spectrum and Perturbation Dynamics

To provide a more complete picture, we analyze how the full Lyapunov spectrum affects perturbation evolution.

Proposition 1 (Multi-Directional Divergence)**.** *Let $\lambda_1 > \lambda_2 > \cdots > \lambda_d$ be the Lyapunov exponents of the system, and let $\{\mathbf{v}_i\}$ be the corresponding Lyapunov directions. For an initial perturbation $\boldsymbol{\delta}_0 = \sum_{i=1}^{d} \alpha_i \mathbf{v}_i$, the evolution satisfies:*

$$\|\boldsymbol{\delta}_k\| \asymp \max_i |\alpha_i| e^{\lambda_i k} \tag{19}$$

In particular, if $\alpha_1 \neq 0$, then the perturbation grows at the maximal rate λ_1, regardless of the other components.

This result shows that even if most directions are stable (negative Lyapunov exponents), the presence of even a single unstable direction can cause exponential divergence.

4 Failure Analysis of Linear Methods

4.1 Fundamental Limitations

Our dynamical systems analysis reveals several fundamental reasons why linear interpretability methods fail in deep networks operating under chaotic dynamics.

The first and most critical limitation is exponential sensitivity to initial conditions. Theorem 1 demonstrates that steering vectors and other linear interventions become completely unpredictable after just $O(\log(1/\epsilon))$ layers, where ϵ represents the precision of the intervention. This means that even tiny implementation errors, numerical precision limitations, or uncertainty in the intervention magnitude can compound rapidly, causing the actual effect to diverge exponentially from the intended linear prediction. In practical terms, this makes precise control of model behavior difficult in deep networks operating in chaotic regimes.

The second fundamental limitation concerns the breakdown of linear superposition. In linear systems, the effects of multiple interventions simply add together, allowing for compositional control strategies. However, in chaotic systems, perturbations interact nonlinearly as they evolve, meaning that the effect of applying two steering vectors simultaneously is not the sum of their individual effects. This nonlinear interaction becomes more pronounced as perturbations grow, fundamentally undermining approaches that rely on decomposing complex behaviors into linear combinations of simpler interventions.

The third limitation involves the temporal instability of learned linear relationships. Even if linear relationships appear to hold at a given layer or for a specific set of inputs, the chaotic dynamics ensure that these relationships will not persist as inputs change or as analysis moves to different layers. This makes it impossible to develop stable, generalizable linear interpretability tools that work consistently across different contexts within the same network.

4.2 Empirical Predictions

Our theoretical framework generates several concrete and testable predictions about when and how linear interpretability methods will fail. The first prediction concerns the depth-dependence of steering effectiveness. Our analysis predicts that steering effectiveness should decrease exponentially with the depth at which interventions are applied, following the pattern effectiveness $\propto e^{-\lambda_{\max} \cdot \text{depth}}$. This can be tested by applying identical steering vectors at different layers and measuring the magnitude and consistency of the resulting output changes.

The second prediction relates to the relationship between network width and chaotic behavior. Wider networks typically have more degrees of freedom and may be more likely to exhibit chaotic dynamics with larger Lyapunov exponents. Our theory predicts that linear interpretability methods should become less reliable as network width increases, all else being equal. This can be tested by comparing the stability of steering interventions across networks of different widths but similar depth and training procedures.

The third prediction concerns the relationship between training dynamics and interpretability. Networks trained with techniques that promote smoother loss landscapes (such as weight decay, batch normalization, or specific initialization schemes) may have smaller Lyapunov exponents and thus be more amenable to linear analysis. Conversely, networks trained with techniques that increase expressivity (such as very deep architectures or specific activation functions) may exhibit more chaotic behavior and be less suitable for linear interpretability methods.

5 Implications for Interpretability Research

5.1 Fundamental Challenges

Our analysis reveals that the challenges facing linear interpretability methods are not merely technical limitations that can be overcome with better algorithms or more data. Instead, they represent fundamental mathematical constraints imposed by the chaotic nature of deep network dynamics. When neural networks operate in chaotic regimes—as evidenced by positive Lyapunov exponents—the exponential sensitivity to initial conditions makes long-range prediction and control inherently impossible, regardless of the sophistication of the interpretability method.

This has profound implications for the field of mechanistic interpretability. It suggests that the goal of achieving precise, predictable control over neural network behavior through linear interventions may be mathematically unattainable in deep networks. Rather than viewing this as a failure of current methods, we should recognize it as a fundamental constraint that must be incorporated into the design of interpretability tools.

The chaotic nature of deep networks also explains why many interpretability methods that work well on shallow networks or in controlled laboratory settings fail to scale to large, practical systems. The exponential amplification of small errors and the breakdown of linear superposition make it increasingly difficult to maintain reliable interpretability as system complexity grows.

5.2 Towards Chaos-Aware Interpretability

Given these fundamental limitations, we advocate for the development of chaos-aware interpretability methods that explicitly account for the nonlinear dynamics of deep networks. The first direction involves developing Lyapunov-aware analysis tools that compute local stability properties before attempting interpretability interventions. By measuring the local Lyapunov exponents, researchers can identify regions where linear methods might be temporarily reliable and avoid regions where chaotic dynamics make linear analysis futile.

The second direction focuses on developing short-horizon interpretability methods that operate within the predictability limits imposed by chaotic dynamics. Rather than attempting to control or predict network behavior across many layers, these methods would focus on understanding and influencing behavior within the chaos horizon $L_{\text{chaos}} = O(\log(1/\epsilon)/\lambda_{\max})$. This might involve developing layer-by-layer analysis techniques or focusing on understanding how information transforms across just a few consecutive layers.

The third direction involves developing ensemble-based interpretability approaches that account for the inherent uncertainty introduced by chaotic dynamics. Rather than seeking single, deterministic explanations, these methods would characterize the distribution of possible behaviors that could arise from small variations in inputs or interventions. This probabilistic approach would provide more realistic assessments of what can and cannot be reliably predicted or controlled in deep networks.

6 Related Work

Our work builds on the growing recognition that deep neural networks exhibit complex dynamical behavior that can be analyzed using tools from nonlinear dynamics [9, 11]. The connection between residual networks and continuous-time dynamical systems has been explored extensively [2, 4], but the implications for interpretability have received less attention.

Recent empirical work has documented various limitations of linear interpretability methods [10, 3], providing evidence consistent with our theoretical predictions. However, most of this work focuses on specific failure modes or methodological issues rather than the fundamental mathematical constraints we identify.

The application of chaos theory to neural networks has a long history [12], but most previous work focused on learning dynamics or computational capabilities rather than interpretability. Our contribution lies in connecting this dynamical systems perspective directly to the limitations of interpretability methods.

7 Limitations and Future Work

Our analysis relies on several assumptions that merit careful consideration. The most important assumption is that the networks under analysis exhibit chaotic dynamics with positive Lyapunov exponents. While there is empirical evidence that many deep networks operate in chaotic regimes [9], this may not be universal. Some networks, particularly those with specific architectural constraints or training procedures, may exhibit more regular dynamics that are more amenable to linear analysis.

Additionally, our analysis focuses on worst-case behavior and may overestimate the practical difficulties of linear interpretability. Real networks may have special structure, such as approximate low-rank dynamics or hierarchical organization that makes them more predictable than our general theory suggests. Understanding when and where such special structure exists represents an important direction for future work.

Future work should investigate the relationship between training procedures and dynamical properties, and develop practical tools for measuring and characterizing the chaotic properties of real networks.

8 Conclusion

We have provided a rigorous dynamical systems analysis of the fundamental limitations of linear representation hypotheses in deep residual networks. Our main contribution demonstrates that ϵ-close perturbations diverge exponentially within $O(\log(1/\epsilon))$ layers under chaotic dynamics, revealing that the exponential sensitivity to initial conditions characteristic of chaotic systems makes linear approximations inherently unreliable in deep networks.

This analysis provides the first rigorous theoretical explanation for the empirically observed instability and inconsistency of linear interpretability methods in large-scale networks. Rather than viewing these failures as technical limitations to be overcome, our results suggest they reflect fundamental mathematical constraints imposed by the chaotic nature of deep network dynamics.

Our work points toward the need for chaos-aware interpretability methods that explicitly account for the nonlinear dynamics of deep networks. By recognizing and working with the inherent limitations imposed by chaotic dynamics, the interpretability community can develop more robust and realistic approaches to understanding neural network behavior.

References

[1] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations Workshop Track*, 2017.

[2] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, pages 6571–6583, 2018.

[3] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

[5] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018.

[6] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2023.

[7] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024.001, 2020.

[8] Valery I Oseledets. A multiplicative ergodic theorem: Lyapunov characteristic numbers for dynamical systems. *Transactions of the Moscow Mathematical Society*, 19:197–231, 1968.

[9] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Advances in Neural Information Processing Systems*, 29:6561–6569, 2016.

[10] Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, 2021.

[11] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. The correspondence between random neural networks and dynamical systems. *arXiv preprint arXiv:1702.08360*, 2017.

[12] Haim Sompolinsky, Andrea Crisanti, and Hans-Jürgen Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262, 1988.