Why Can't Neural Networks Master Extrapolation ? Insights from Physical Laws

Ramzi Dakhmouche
Institute of Mathematics, EPFL, Switzerland
Computational Engineering Lab, Empa, Switzerland
ramzi.dakhmouche@epfl.ch

Hossein Gorji
Computational Engineering Lab, Empa, Switzerland
mohammadhossein.gorji@empa.ch

Abstract

Motivated by the remarkable success of Foundation Models (FMs) in language modeling, there has been growing interest in developing FMs for time series prediction, given the transformative power such models hold for science and engineering. This culminated in significant success of FMs in short-range forecasting settings. However, extrapolation or long-range forecasting remains elusive for FMs, which struggle to outperform even simple baselines. This contrasts with physical laws which have strong extrapolation properties, and raises the question of the fundamental difference between the structure of neural networks and physical laws. In this work, we identify and formalize a fundamental property characterizing the ability of statistical learning models to predict more accurately outside of their training domain, hence explaining performance deterioration for deep learning models in extrapolation settings. In addition to a theoretical analysis, we present empirical results showcasing the implications of this property on current deep learning architectures. Our results not only clarify the root causes of the extrapolation gap but also suggest directions for designing next-generation forecasting models capable of mastering extrapolation.

1 Introduction

In physics and engineering, the most impactful models are those that remain reliable even beyond the training data or observation domain, where controlled experiments or simulations are expensive, sparse, or beyond reach. Yet, the ability to master extrapolation remains out of reach for state-of-the-art machine learning models, including the latest forecasting approaches [15, 11, 12, 10] such as Foundation Models (FMs). FMs [8, 17] are transformer-based neural networks with a considerably large number of parameters (hundreds of millions) pretrained on large datasets from diverse time series domains, but still have been shown to be outperformed by simple linear or seasonal models [15] in extrapolation or long-range forecasting settings. More generally, deep learning models contrast with physical models having a symbolic structure, in the range of data regimes they can encode or generate. Indeed, in the context of fluid dynamics for instance, training a neural network solely on laminar flow data would lead to highly inaccurate predictions [6, 14, 2] for turbulent flows. Whereas a symbolic differential equation such as the Navier-Stokes equation- or its discretization in space, represents a very flexible data-generating process that covers a very wide range of regimes. Furthermore, symbolic learning has demonstrated stronger extrapolation properties in biological applications [5, 4]. This indicates a fundamental difference between symbolic models and neural networks and raises the question of characterizing this difference precisely to identify and design models that would extrapolate effectively. Although neural networks are typically over-parameterized and have a particular structure involving composing the same type of functions repeatedly, they are still approximators with explicit expressions, just like symbolic models, hence it is not clear what

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Machine Learning and the Physical Sciences.

qualitative properties make them less suitable for extrapolation. Consequently, we tackle this question by making the following contributions:

- We identify a precise characterization of structural variability as a key property that allows symbolic models to be better extrapolators than neural networks.
- We propose a theoretical analysis demonstrating the role of this property in ensuring improved extrapolation under the Occam's razor hypothesis, thereby providing a principled basis for model selection outside the training range.
- Building on this insight, we propose a minimal neural network architecture change as a first step toward better extrapolation and showcase its performance gains on synthetic and electricity time-series data.

2 Problem Setting

Given a dataset $(X_i, Y_i)_{i\leq n}$ of input and response variables $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$, we consider the regression task of predicting the response value Y for new samples X. Assuming $(X_i)_{i\leq n}$ are sampled from a given domain $\mathcal{D} \subseteq \mathbb{R}^d$ with $Y_i = f(X_i)$ for all $i \leq n$, the goal in extrapolation is to achieve low prediction error on samples outside of $\mathcal{D}$. Note that, this is different from standard generalization in statistical learning theory, in that the latter focuses on ensuring low prediction error for new samples from the training domain or distribution. Importantly, extrapolation is an ill-posed problem in general, since outside of $\mathcal{D}$ the ground truth data generating function f could have infinitely many different behaviors. Hence, to tackle this question in a meaningful way, we restrict the setting to cases where f has an explicit expression. Such a set includes both neural networks and elementary functions such as polynomials, exponentials of polynomials, logarithms, trigonometric functions ...etc. One unifying property of all these functions is that they satisfy polynomial differential equations [13, 9]. This allows us to define a selection principle based on Occam's razor [16], by assigning preference to functions that encode less information. For a concise presentation, we assume $d = 1$. A relevant measure of information in this case is the number of bits needed to represent the polynomial ordinary differantial equation (ODE) satisfied by each function, and which is of the form $P(x, y', y'', y^{(3)}, \dots) = 0$ in the scalar input setting. This contrasts with previous pathwise or continuous information measures which are typically asymptotic [3, 1]. As illustrated in figure 1, functions corresponding to simpler ODEs encode less qualitative variation and hence correspond to simpler hypotheses. In the following, we therefore consider a model to yield better extrapolation, if it does so on simpler functions in this sense.

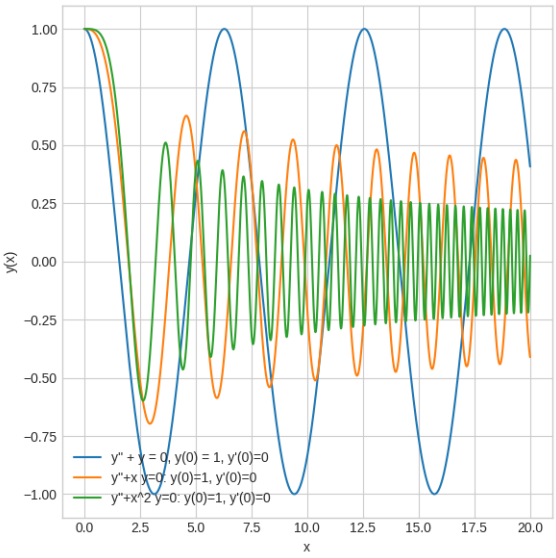

Figure 1: Information levels illustration. The green curve oscillates twice as much as the orange one, while they both have shrinking modes requiring more bits to be encoded.

3 Measuring Structural Variability

In order to ensure improved extrapolation, the key idea is to design a statistical learning model with structurally diverse building blocks, as we will analyze below. But first, how to measure structural variability ? Classical geometric notions such as dot product are too strict, while analytical measures such as norms are agnostic to structure. For instance, in $L^2(0, 2\pi)$, trigonometric functions have dot product equal to 0 while having the same structure modulo a small shift (see figure

2). To overcome that, we propose to measure variability via algebraic objects derived from the ODEs satisfied by the model building blocks. Specifically, the measure has two components:

1. The order of the ODE: the highest order derivative of the ODE represents the most elementary discriminating quantity to consider, and encodes coarse qualitative change. For instance, oscillatory behavior cannot be observed in solutions of first order ODEs and require a second order term, as illustrated in figure 2.
2. Algebraic reduction classification: We focus here on polynomials of at most second degree, since the Navier-Stokes equation discretized in space is a quadratic ODE, yet captures regimes which are arguably as far apart as possible. More precisely, the linear case can be analyzed via the companion univariate annihilating polynomial defined for an ODE $y^{(n)} + c_n y^{(n-1)} + \cdots + c_1 y = 0$ as $p(D) = D^n + c_n D^{n-1} + \cdots + c_1$, where variability of the roots induces variability of the corresponding solutions. As for the quadratic case, it can be mapped to Sylvester's Law of Inertia, which reduces second degree multivariate polynomials to sums of second degree monomials with coefficients in $\{-1, 0, 1\}$. As illustrated in figure 5, varying the choice of these coefficients for each term leads to a different qualitative behavior for the ODE solutions.

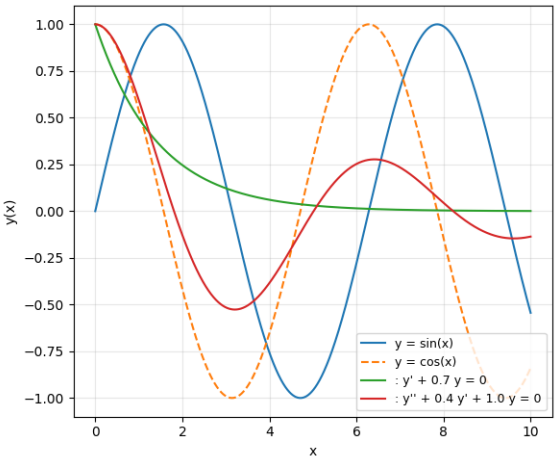

Figure 2: Structural variation illustration. The cosine function is just a shifted version of sine.

Given the previous framework, we show in the following result that lacking a level of structural variability prevents accurate extrapolation.

Proposition 1. (Informal)
Consider a parameterized set of regression functions $\{f_\theta : \mathbb{R} \longrightarrow \mathbb{R},\ \theta \in \mathbb{R}^p\}$ such that for $x \in \mathbb{R}$

$$f_\theta(x) = f_{1,\theta}(x) + \cdots + f_{k,\theta}(x)$$

Assume $(f_{j,\theta})_{j \leq k}$ do not span the variability classes, then for all $M > 0$, there exists a smooth function g_M such that

$$\inf_{\theta \in \mathbb{R}^p} \|f_\theta - g\|_\infty > M$$

Proof. We restrict the proof to the linear case and postpone it to Appendix B.
Next, we analyze the corresponding behavior for neural networks.

4 Differential Annihilators of Neural Networks

A natural question that arises given the proposed approach to measure structural variability is: what about the structure of the ODE satisfied by a neural network ? We provide an answer to this question in the following proposition, and will refer to such an equation as the differential annihilator.

Proposition 2. (Informal)
Consider a neural network f_θ with Tanh or Sigmoid activation functions. Then,

- The minimal polynomial ODE satisfied by f_θ is of degree $\sum_\ell^L m_\ell$, where L is its length and m_ℓ the width of each layer, and admits constant solutions which are highly dependent on the training data.
- f_θ converges exponentially to a constant as it approaches the border of the training domain.

Hence, the set of functions that a neural network can approach globally, that is, those whose differential annihilators can be well approximated by the annihilator of a neural network is considerably reduced, due to the constraint of constant solutions. We report the proof in Appendix C.

5 Numerical Results

We evaluate our proposed framework by making a non-standard yet simple change in Multi-layer Perceptron (MLP) architecture to illustrate the gain obtained by structural variability. For that matter, we train a MLP that is a linear combination of subnetworks of varying length. This makes the order of the differential annihilators of each sub-network different. We use sigmoid activation functions and 3-layers with width of 16. We train with early stopping to prevent overfitting.

5.1 Synthetic Dataset

We first evaluate the architectural change on elemantry function fitting, namely sine, complex periodic, quadratic and hyperbolic tangent. We report the extrapolation errors in table 1 below. We report estimated trajectories for sin functions in figures 3 and 4

Function	StandardNet (MSE)	Proposed Net (MSE)
Sin	1.172	0.082
Complex Periodic	1.592	0.416
Quadratic	0.218	0.149
Tanh	0.0018	0.0076

Table 1: Average extrapolation MSE for each function across window sizes.

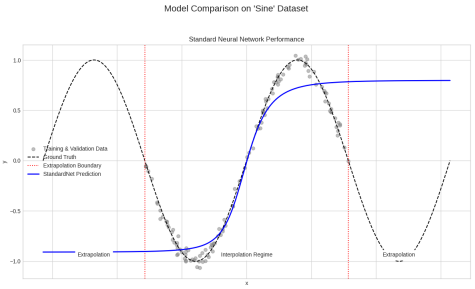

Figure 3: Predicted trajectory- Standard MLP

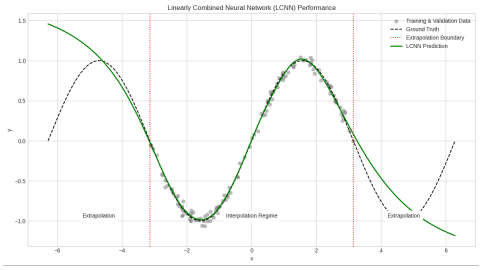

Figure 4: Predicted trajectory - Proposed MLP

5.2 Real-World Dataset

We further evaluate on electric time series data from the ETTH dataset [18, 7]. We train on a window of 2500 samples and evaluate. We report the extrapolation results in table 2 below. Once again, we see an improvement in extrapolation accuracy via the proposed minimal architectural change.

Extrapolation Window size	**StandardNet MSE**	**Proposed Net MSE**
0.79	15.005008	11.537091
1.57	14.166123	11.067607
2.36	13.238490	11.514800
3.14	12.184800	12.040763

Table 2: Extrapolation MSE averaged over 3 runs.

6 Conclusion

In this work, we investigated the fundamental limitations of current neural networks for extrapolating beyond the training regime. By identifying and formalizing a key property that governs extrapolation performance, we provided both a theoretical explanation for the observed gap and empirical evidence demonstrating its impact. Beyond clarifying the root causes of this phenomenon, our analysis points to concrete avenues for bridging the extrapolation gap, such as merging symbolic and over-parameterized models to increase structural variability.

Acknowledgements

This work was supported by the Swiss National Science Foundation under grant No. 212876. We acknowledge computational resources from the Swiss National Supercomputing Centre CSCS. R.D. acknowledges Dr. Ivan Lunati for providing laboratory infrastructure and computational resources and for valuable discussions.

References

[1] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural computation*, 13(11):2409–2463, 2001.

[2] C.-Y. Chuang, A. Torralba, and S. Jegelka. Estimating generalization under distribution shifts via domain-invariant representations. *Proceedings of Machine Learning Research*, 119, 2020.

[3] T. M. Cover and J. A. Thomas. Entropy rates of a stochastic process. *Elements of Information Theory*, pages 63–65, 1991.

[4] R. Dakhmouche, I. Lunati, and H. Gorji. Robust symbolic regression for dynamical system identification. *Transactions on Machine Learning Research*, 2025.

[5] R. Dakhmouche, I. Lunati, and H. Gorji. Robust symbolic regression for network trajectory inference. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2025.

[6] K. Fukami, K. Hasegawa, T. Nakamura, M. Morimoto, and K. Fukagata. Model order reduction with neural networks: Application to laminar and turbulent flows. *SN Computer Science*, 2(6):467, 2021.

[7] R. Ilbert, A. Odonnat, V. Feofanov, A. Virmaux, G. Paolo, T. Palpanas, and I. Redko. Samformer: Unlocking the potential of transformers in time series forecasting with sharpness-aware minimization and channel-wise attention. *arXiv preprint arXiv:2402.10198*, 2024.

[8] M. Jin, S. Wang, L. Ma, Z. Chu, J. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-f. Li, S. Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

[9] E. R. Kolchin. *Differential algebra & algebraic groups*, volume 54. Academic press, 1973.

[10] Z. Li, X. Qiu, P. Chen, Y. Wang, H. Cheng, Y. Shu, J. Hu, C. Guo, A. Zhou, C. S. Jensen, et al. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 5595–5606, 2025.

[11] X. Liu, J. Liu, G. Woo, T. Aksu, Y. Liang, R. Zimmermann, C. Liu, S. Savarese, C. Xiong, and D. Sahoo. Moirai-moe: Empowering time series foundation models with sparse mixture of experts. *arXiv preprint arXiv:2410.10469*, 2024.

[12] G. Pucher, A. Dada, F. Nensa, M. Schuler, C. Reinhardt, J. Kleesiek, and C. M. Sauer. Evaluating zero-shot foundation models for time series forecasting in clinical settings: A simulation study with electronic health records. *Studies in health technology and informatics*, 329:820–824, 2025.

[13] J. F. Ritt. *Differential algebra*, volume 33. American Mathematical Soc., 1950.

[14] T. Sutter, A. Krause, and D. Kuhn. Robust generalization despite distribution shift via minimum discriminating information. *Advances in Neural Information Processing Systems*, 34:29754–29767, 2021.

[15] W. Toner, T. L. Lee, A. Joosen, R. Singh, and M. Asenov. Performance of zero-shot time series foundation models on cloud data. *arXiv preprint arXiv:2502.12944*, 2025.

[16] S.-M. Udrescu and M. Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science advances*, 6(16):eaay2631, 2020.

[17] G. Woo, C. Liu, A. Kumar, C. Xiong, S. Savarese, and D. Sahoo. Unified training of universal time series forecasting transformers. In *International Conference on Machine Learning*, pages 53140–53164. PMLR, 2024.

[18] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 11106–11115, 2021.

A Additional Numerical Results

We report graphical representations of solutions of quadratic ODEs corresponding to various classes, featuring structural variability.

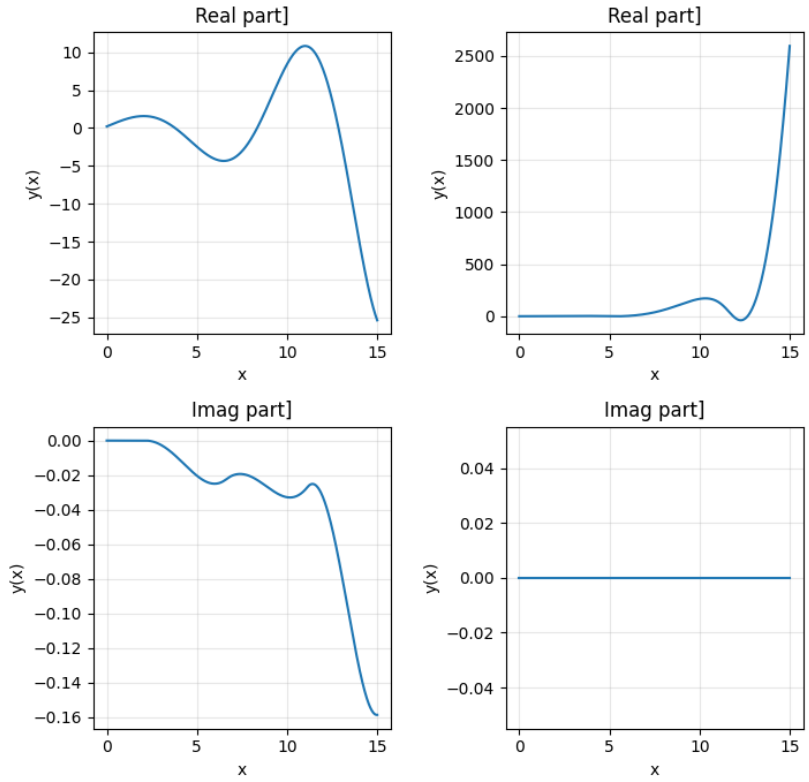

Figure 5: Structural variability illustration

B Proof of proposition 1

Notation. Fix a compact interval $K = [a, b] \subset \mathbb{R}$. We denote by $C(K)$ the Banach space of real–valued continuous functions on K with the sup–norm

$$\|h\|_\infty := \sup_{x \in K} |h(x)|.$$

For $r \in \mathbb{N}$, let $C^r(K) \subset C(K)$ be the subspace of functions with r continuous derivatives on K.

Model. Let $\{f_\theta : \mathbb{R} \to \mathbb{R} : \theta \in \mathbb{R}^p\}$ be a parameterized family with a decomposition

$$f_\theta(x) = \sum_{j=1}^{k} f_{j,\theta}(x), \qquad x \in \mathbb{R}.$$

Variability deficit (precise). We say that the building blocks $\{f_{j,\theta}\}$ *do not span the variability classes* if there exists a nonzero constant–coefficient linear differential operator

$$T = p(D) = D^r + c_r D^{r-1} + \cdots + c_1, \qquad r \geq 1,$$

such that

$$T f_{j,\theta} \equiv 0 \text{ on } K \quad \text{for all } \theta \in \mathbb{R}^p \text{ and } j = 1, \ldots, k.$$

Equivalently, every f_θ lies in the closed linear subspace

$$\mathcal{V} := \ker(T) \cap C(K) = \{y \in C^r(K) : Ty \equiv 0 \text{ on } K\} \subsetneq C(K).$$

Proposition 1 (sup–norm separation under variability deficit). Assume the variability deficit above holds. Then for every $M > 0$ there exists a function $g_M \in C^\infty(K)$ such that

$$\inf_{\theta \in \mathbb{R}^p} \|f_\theta - g_M\|_\infty \ > \ M.$$

Proof. Since $T \not\equiv 0$, the subspace $\mathcal{V} = \ker(T) \cap C(K)$ is a proper closed linear subspace of $C(K)$. Choose $g_0 \in C^\infty(K) \setminus \mathcal{V}$ (e.g., any smooth function with $Tg_0 \not\equiv 0$ on K).

By the Hahn–Banach separation theorem in the Banach space $C(K)$, there exists a bounded linear functional $\Lambda \in C(K)^*$ such that

$$\Lambda|_{\mathcal{V}} = 0 \quad \text{and} \quad \Lambda(g_0) \neq 0.$$

By the Riesz representation theorem for $C(K)^*$, there is a finite signed Borel measure μ on K with

$$\Lambda(h) = \int_K h(x)\, d\mu(x), \qquad \|\Lambda\| = \|\mu\|_{\mathrm{TV}} < \infty.$$

For every θ we have $f_\theta \in \mathcal{V}$, hence $\Lambda(f_\theta) = 0$. Therefore, for all θ,

$$|\Lambda(g_0 - f_\theta)| = |\Lambda(g_0)| \leq \|\Lambda\|\, \|g_0 - f_\theta\|_\infty.$$

Taking the infimum over θ yields

$$\inf_\theta \|g_0 - f_\theta\|_\infty \ \geq \ \frac{|\Lambda(g_0)|}{\|\Lambda\|} \ =: \ \varepsilon_0 \ > \ 0.$$

Now fix $M > 0$ and define $g_M := \alpha g_0$ with $\alpha := (M+1)/\varepsilon_0$. Then, for all θ,

$$\|g_M - f_\theta\|_\infty \ \geq \ \frac{|\Lambda(g_M - f_\theta)|}{\|\Lambda\|} = \frac{|\Lambda(g_M)|}{\|\Lambda\|} = \frac{\alpha|\Lambda(g_0)|}{\|\Lambda\|} = \alpha\varepsilon_0 = M + 1 > \ M.$$

Hence $\inf_\theta \|f_\theta - g_M\|_\infty > M$, as claimed. □

Remarks. (1) The assumption "do not span the variability classes" is encoded here by the existence of a nontrivial annihilating operator T that all building blocks satisfy on K. In the linear (constant–coefficient) case, $\ker T$ is a finite–dimensional subspace spanned by exponentials and sinusoidals, and the proof above is entirely within the sup–norm on K.

(2) The same argument applies to any situation in which the realized model class $\{f_\theta|_K : \theta \in \mathbb{R}^p\}$ is contained in a proper *closed linear* subspace of $C(K)$. In particular, even if one arrives at $\mathcal{V}$ via quadratic jet constraints, it suffices that the resulting feasible set be contained in some proper closed linear subspace $V \subset C(K)$ (e.g., by linearization on a restricted regime). The Hahn–Banach/Riesz separation then yields the same conclusion.

C Proof of proposition 2

Proposition[Polynomial ODE and boundary behavior for 1D→1D MLPs with $\tanh$/sigmoid] Let $f_\theta : \mathbb{R} \to \mathbb{R}$ be a depth-L multilayer perceptron with widths $(m_1, \ldots, m_L)$, scalar input x, scalar output $f(x)$, and activation $\phi \in \{\tanh, \sigma\}$ (where $\sigma(u) = \frac{1}{1+e^{-u}}$). Set $M = \sum_{\ell=1}^{L} m_\ell$.

(i) There exists a nonzero polynomial $\mathcal{P} \in \mathbb{R}[T_0, \ldots, T_M]$ (with coefficients depending on θ) such that the scalar output satisfies the autonomous polynomial differential equation

$$\mathcal{P}\big(f(x), f'(x), \ldots, f^{(M)}(x)\big) \equiv 0.$$

In particular, the minimal order of a polynomial ODE satisfied by f_θ is at most M. Moreover, this ODE admits constant solutions: there exist constants c (depending on θ) such that $f(x) \equiv c$ solves the ODE, i.e. $\mathcal{P}(c, 0, \ldots, 0) = 0$.

(ii) (Exponential convergence to constants at the boundary) Fix a bounded *training domain* $[a, b] \subset \mathbb{R}$. Then there exist constants $f_\infty^-, f_\infty^+ \in \mathbb{R}$ and positive constants $C_\pm, \kappa_\pm$ (depending on θ) such that

$$|f(x) - f_\infty^+| \leq C_+ e^{-\kappa_+(x-b)} \quad (x \to +\infty), \qquad |f(x) - f_\infty^-| \leq C_- e^{-\kappa_-(a-x)} \quad (x \to -\infty).$$

In words, as one moves beyond either border of the training interval, f_θ converges exponentially fast to a constant.

Proof. **Network and notation.** Write the usual forward equations

$$h^{(0)}(x) = x, \qquad z^{(\ell)}(x) = W^{(\ell)} h^{(\ell-1)}(x) + b^{(\ell)}, \qquad h^{(\ell)}(x) = \phi\big(z^{(\ell)}(x)\big) \in \mathbb{R}^{m_\ell},$$

and $f(x) = \alpha^\top h^{(L)}(x) + \beta$. Stack all hidden activations into

$$Y(x) \coloneqq \big(h^{(1)}(x), h^{(2)}(x), \dots, h^{(L)}(x)\big) \in \mathbb{R}^M,$$

and define the readout $G(Y) \coloneqq \alpha^\top h^{(L)} + \beta$, so $f(x) = G\big(Y(x)\big)$.

We shall use the elementary identities, valid for $y = \phi(u)$,

$$\phi = \tanh: \quad \phi'(u) = 1 - \phi(u)^2 = 1 - y^2, \qquad \phi = \sigma: \quad \phi'(u) = \phi(u)\big(1 - \phi(u)\big) = y(1-y). \tag{$*$}$$

(i) Existence of a polynomial ODE of order $\le M$ and constant solutions.

Step 1: An autonomous polynomial ODE for the hidden state. By the chain rule and $(*)$, each first-layer neuron satisfies

$$\frac{d}{dx} h_j^{(1)}(x) = w_j^{(1)}\, P\big(h_j^{(1)}(x)\big),$$

with $P(y) = 1 - y^2$ (tanh) or $P(y) = y(1-y)$ (sigmoid). Thus the first-layer derivatives are *polynomials in the first-layer activations*.

For layer 2,

$$\frac{d}{dx} h_k^{(2)}(x) = \phi'\big(z_k^{(2)}(x)\big) \sum_j a_{kj} \frac{d}{dx} h_j^{(1)}(x) = Q\big(h_k^{(2)}(x)\big) \sum_j a_{kj}\, w_j^{(1)}\, P\big(h_j^{(1)}(x)\big),$$

where $Q(y) = 1 - y^2$ (tanh) or $Q(y) = y(1-y)$ (sigmoid). Hence second-layer derivatives are *polynomials in $h^{(1)}$ and $h^{(2)}$*. Proceeding inductively over layers shows there exists a polynomial map

$$F : \mathbb{R}^M \to \mathbb{R}^M \quad \text{such that} \quad Y'(x) = F\big(Y(x)\big). \tag{1}$$

Step 2: $f^{(k)}$ are polynomial functions of Y. Define $H_0(Y) \coloneqq G(Y)$ and recursively

$$H_{k+1}(Y) \coloneqq \nabla H_k(Y) \cdot F(Y) \qquad (k \ge 0).$$

By construction and (1), H_k is a polynomial for each k, and along the network trajectory $Y(x)$ we have

$$f^{(k)}(x) = H_k\big(Y(x)\big), \qquad k = 0, 1, \dots. \tag{2}$$

Step 3: Algebraic elimination. Consider the polynomial map

$$\Psi : \mathbb{R}^M \to \mathbb{R}^{M+1}, \qquad \Psi(Y) = \big(H_0(Y), H_1(Y), \dots, H_M(Y)\big).$$

Its image is an algebraic set of dimension at most M. Hence there exists a nonzero polynomial $\mathcal{P} \in \mathbb{R}[T_0, \dots, T_M]$ that vanishes on $\operatorname{Im}\Psi$ (e.g., by elimination/Nullstellensatz). Evaluating along $Y(x)$ and using (2) yields the autonomous polynomial differential equation

$$\mathcal{P}\big(f(x), f'(x), \dots, f^{(M)}(x)\big) \equiv 0,$$

establishing the order bound $\le M$.

Constant solutions. An equilibrium of (1) produces a constant solution of the f–ODE. To see that such equilibria exist, choose any vector $s^{(1)} \in S_\phi^{m_1}$ with

$$S_\phi = \begin{cases} \{\pm 1\}, & \phi = \tanh, \\ \{0, 1\}, & \phi = \sigma, \end{cases} \qquad \text{so that } P(s_j^{(1)}) = 0 \ \ \forall j.$$

Then $h^{(1)}(x) \equiv s^{(1)}$ gives $h^{(1)\prime}(x) \equiv 0$, and by the layerwise formulae every higher-layer derivative also vanishes (they are linear combinations of lower-layer derivatives). Define recursively

$$s^{(\ell)} \coloneqq \phi\big(W^{(\ell)} s^{(\ell-1)} + b^{(\ell)}\big), \qquad \ell = 2, \dots, L,$$

and set $\bar{Y} \coloneqq (s^{(1)}, \ldots, s^{(L)})$. Then $F(\bar{Y}) = 0$, hence $H_0(\bar{Y}) = G(\bar{Y}) =: c$ and $H_k(\bar{Y}) = 0$ for $k \geq 1$. Plugging into the identity $\mathcal{P}(H_0, \ldots, H_M) \equiv 0$ gives

$$\mathcal{P}(c, 0, \ldots, 0) = 0,$$

so $f(x) \equiv c$ is a (constant) solution of the ODE. The set of such constants depends on θ through the affine maps and readout, hence is highly data/parameter dependent.

(ii) Exponential convergence to constants outside a bounded domain.

We prove the $x \to +\infty$ case; $x \to -\infty$ is analogous. For the first layer, each preactivation is $z_j^{(1)}(x) = w_j^{(1)} x + b_j^{(1)}$, so either $w_j^{(1)} = 0$ (then $h_j^{(1)}$ is constant) or $|z_j^{(1)}(x)| \geq |w_j^{(1)}|x - |b_j^{(1)}|$ grows linearly with x.

Use the standard tails, for all $z \in \mathbb{R}$,

$$\big|\tanh z - \operatorname{sgn}(z)\big| \leq 2e^{-2|z|}, \qquad \big|\sigma z - \mathbf{1}_{\{z>0\}}\big| \leq e^{-|z|}.$$

Therefore there exist constants $C_1, \kappa_1 > 0$ (depending on $w^{(1)}, b^{(1)}$) and a saturation vector $s^{(1)} \in \{\pm 1\}^{m_1}$ (tanh) or $s^{(1)} \in \{0, 1\}^{m_1}$ (sigmoid) such that

$$\|h^{(1)}(x) - s^{(1)}\| \leq C_1 e^{-\kappa_1 x} \qquad (x \to +\infty).$$

Inductively, write

$$z^{(\ell)}(x) = W^{(\ell)} s^{(\ell-1)} + b^{(\ell)} \; + \; W^{(\ell)}\big(h^{(\ell-1)}(x) - s^{(\ell-1)}\big).$$

The first two terms are constant; the last term is $O(e^{-\kappa_1 x})$. Since ϕ is globally Lipschitz with constant ≤ 1 for $\tanh$ and $\leq 1/4$ for σ, there exist $C_\ell, \kappa_\ell > 0$ and vectors $s^{(\ell)} = \phi\big(W^{(\ell)} s^{(\ell-1)} + b^{(\ell)}\big)$ such that

$$\|h^{(\ell)}(x) - s^{(\ell)}\| \leq C_\ell e^{-\kappa_\ell x} \qquad (x \to +\infty).$$

Propagating to the linear readout gives

$$|f(x) - f_\infty^+| \leq C e^{-\kappa x} \qquad (x \to +\infty),$$

with $f_\infty^+ = \alpha^\top s^{(L)} + \beta$ and suitable $C, \kappa > 0$ depending on θ. The $x \to -\infty$ case yields f_∞^- similarly. Translating from $x \to \pm\infty$ to "distance beyond the border" $x - b$ or $a - x$ proves the stated estimates. □