Bias-driven Alignment of Linear and ReLU Networks

Jirko Rubruck *
Experimental Psychology
University of Oxford

Kai J Sandbrink
Experimental Psychology
University of Oxford

Christopher Summerfield
Experimental Psychology
University of Oxford

Andrew M Saxe
Gatsby Unit & SWC
University College London

Devon Jarvis
School of Computer Science and Applied Mathematics
University of the Witwatersrand

Abstract

ReLU networks and their variants are a key building block of modern deep learning architectures. Despite their ubiquity, our understanding of learning dynamics in these models is still limited. Previous work has relied on a strong set of simplifying assumptions such as the removal of bias terms or predefined gating structures. Here, we explore empirically how the inclusion of bias terms influences learning dynamics in ReLU networks in the rich learning regime. Surprisingly, we find that the inclusion of bias terms simplifies learning dynamics, i.e. ReLU networks with bias terms have learning dynamics that are strongly aligned to those of well-understood linear models. Further, ReLU and linear networks with bias terms trained on nonlinear problems display a transient correspondence early in learning that is also reflected in highly structured, linear-like representations. We also highlight additional downstream effects of early linearity and find that the inclusion of bias terms boosts simplicity biases and the over-representation of features associated with simple tasks. We demonstrate the practical relevance of our results beyond simplified settings and show that bias terms can also induce early linearity on image classification tasks. Our results illustrate how seemingly minor and common architectural choices can change learning dynamics, biases towards simplicity, and representational alignment between systems.

1 Introduction

Neural networks, whether artificial or biological, learn representational structures that support a broad set of cognitive functions ranging from perception to complex reasoning. A large body of recent work has found remarkable similarities between representations of natural and artificial learning systems [Yamins et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014, Schrimpf et al., 2020]. In machine learning, the representational alignment between different artificial systems has also been extensively explored in the hope of understanding the role of representations in model behavior and to enhance model interpretability [Kornblith et al., 2019, Klabunde et al., 2025, Sucholutsky et al., 2024].

Representations in neural networks emerge through the complex interplay of architecture, dataset, and the dynamics of learning. Much theoretical work has explored how learning dynamics shape

*correspondence to jirko.rubruck@stx.ox.ac.uk

Preprint.

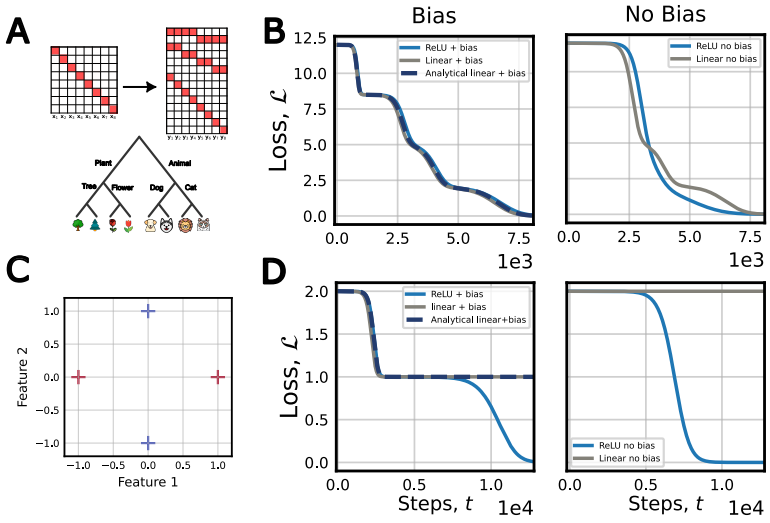

Figure 1: Functional similarity between linear and ReLU networks. A. The hierarchical learning task. This problem can be be solved by a linear network. **B.** Loss of ReLU and linear networks. Both network types are functionally similar when equipped with bias terms and display stage-like learning. Also note the good agreement of ReLU networks with exact solutions by [Saxe et al., 2014] devised for linear networks. **C.** A non-linear learning problem. **D.** ReLU networks with bias are also transiently aligned to linear networks on non-linear problems.

the neural network's internal representations [Saxe et al., 2014, Dominé et al., 2024, Braun et al., 2022] and noted that qualitatively and quantitatively similar representational structure can emerge in linear and non-linear connectionist models [Saxe et al., 2019, Zhang et al., 2025]. Despite this observed similarity there is still debate about how much similar patterns of representations relate to computation and behavior of models [Lampinen et al., 2024, Prince et al., Lampinen et al., 2025, Braun et al., 2025]. Furthermore, the factors that lead to the emergence of similarity have also been debated [Huh et al., 2024]. In particular, Lampinen et al. [2024] demonstrated that neural networks represent features relevant to simpler tasks more strongly. This difference is driven in part by the fact that simpler features are learned more quickly. This early learning ties the representational structure of neural networks to the simplicity bias of deep learning [Huh et al., 2023, Shah et al., 2020, Kalimeris et al., 2019].

While some similarities have been observed between the dynamics and representations in linear and nonlinear models, their learning dynamics are typically distinct [Jarvis et al., 2025, Saxe et al., 2022] and have only been found to be equivalent in specific cases [Zhang et al., 2025]. Here we make the surprising observation that ReLU and linear networks have remarkably similar learning dynamics only when equipped with bias terms. We show how gradient-based learning in these different architectures arrives at similar solutions despite differences in architecture and expressivity. Our findings also underscore how bias terms can enhance simplicity biases and drive representational and functional alignment between distinct model classes.

Our contributions are as follows: (i) We show that ReLU networks and linear networks have equivalent learning dynamics when equipped with bias terms. This phenomenon also persists for non-linear task, where dynamics are aligned in early learning. (ii) We examine implications of this phenomenon for representations and find that bias-ReLU networks learn representations in a structured, linear-like fashion. (iii) We show that the alignment increases representational biases such that simple features are strongly overrepresented. (iv) We highlight that the bias-induced transient alignment extends beyond simple settings and can be observed in networks trained on image data.

2 Alignment of ReLU and linear networks

Setting. We are studying functional and representational alignment of ReLU and linear networks when these models are equipped with bias terms. We consider a learning task in which a network

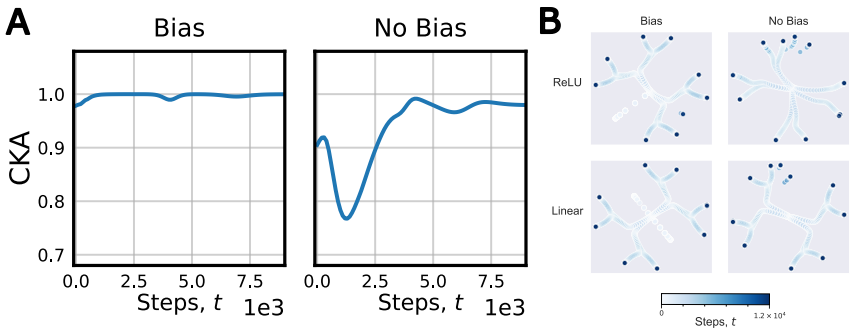

Figure 2: Representational alignment between linear and ReLU networks. A. Alignment is high when networks have bias terms. Centered Kernel Alignment (CKA) between linear and ReLU networks with (left) and without (right) bias terms throughout training. **B.** Multi-dimensional scaling of hidden representations. With bias terms representations in ReLU networks emerge in a structured, linear-like fashion. In non-bias ReLU networks representations evolve in a less structured manner and separate almost from the beginning of learning.

is presented with input vectors $\mathbf{x}_i \in \mathbb{R}^{N_{in}}$ that are associated to output vectors $\mathbf{y}_i \in \mathbb{R}^{N_{out}}$. The total dataset consists of $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$ with N samples. We consider two layer linear networks where the forward pass computes $\hat{\mathbf{y}}_i = \mathbf{W}^2(\mathbf{W}^1\mathbf{x}_i + \mathbf{b})$. We also train ReLU networks of the form $\hat{\mathbf{y}}_i = \mathbf{W}^2\sigma(\tilde{\mathbf{W}}^1\mathbf{x}_i + \mathbf{b})$ with $\sigma(x) = \max(x, 0)$. Here, the weight matrices are of dimensions $\mathbf{W}^1 \in \mathbb{R}^{N_{hid} \times N_{in}}$, $\mathbf{W}^2 \in \mathbb{R}^{N_{out} \times N_{hid}}$ and the bias vector is of dimension $\mathbf{b} \in \mathbb{R}^{N_{hid}}$. We train our networks to minimize a squared error loss of the form $\mathcal{L}(\hat{\mathbf{y}}) = \frac{1}{2}\sum_{i=1}^N \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|^2$. We optimize networks using full batch-gradient descent in the gradient flow regime starting from small initial conditions. For simulations in Fig. 1, Fig. 2, and Fig. 3 we use a hidden size of 64.

Alignment of learning dynamics. To assess the effect of bias terms on ReLU network learning dynamics. We first train networks on a linearly solvable semantic learning problem similar to those considered by Saxe et al. (2019) [Saxe et al., 2019]. The problem is visualized in Fig. 1A. Linear networks in this setting display characteristic stage-like learning that is driven by the progressive acquisition of SVD modes of the dataset input-output correlation matrix (see Appendix A for a quick review of these dynamics). Surprisingly, we find that ReLU networks with bias terms closely track the dynamics of their linear counterparts, so much so that exact solutions developed for linear networks (see Appendix A) provide an excellent match. However, when removing bias terms the connection breaks down and dynamics diverge. We show the general pattern of our result in Fig. 1B.

We also assess the correspondence on a non-linear problem displayed in Fig. 1C. In Fig. 1D we show that even for non-linear problems bias-ReLU and bias-linear network dynamics are exactly matched in early training. As before, the correspondence breaks down when bias terms are ablated. We can see that ReLU networks in early training can be effectively characterized as linear networks when they are equipped with bias terms (Appendix B also shows similar results for biases in both layers). We will next examine the effect of bias terms on the representational structure in these networks.

Representational alignment of ReLU and linear models. We would like to know if the functional similarity between ReLU and linear networks also translates into a representational alignment of both models. In particular, we analyze network hidden representations in ReLU and deep linear networks during the training on the hierarchical learning task (Fig. 1A). To this end, we compare network representations using Centered Kernel Alignment (CKA, Kornblith et al. [2019]). In Fig. 2A we find that representations in ReLU and linear networks are strongly aligned when networks have bias terms. Further we visualize the evolution of hidden layer representations throughout training via Multi-dimensional scaling (Fig. 2B). We can see how hidden representations evolve in an orderly fashion in bias ReLU networks that show excellent agreement with linear networks. The representations for different classes with shared hierarchical features co-evolve, mirroring stage-like learning trajectories. In contrast, representations in bias-free ReLU networks evolve in a less structured manner, with the representations for the different classes separating at the very beginning of learning.

Simple before complex learning in bias-ReLU networks. Recent work demonstrated how learning order contributes to "representational biases" in which simple features and tasks explain large amount

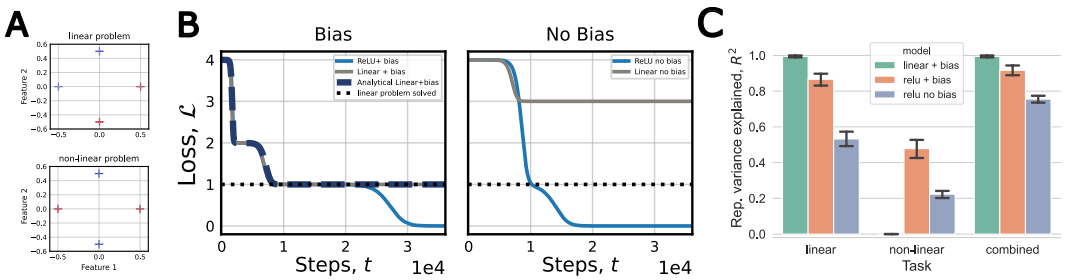

Figure 3: Linear to non-linear learning in ReLU networks. A. Models are trained on a linear problem (top) and a non-linear problem (bottom) problem in parallel. Problems are defined on the same four data-points. **B.** With bias (left) ReLU networks are aligned to linear models in early training and learn the linear task (dotted line) first. **C.** Representational variance explained via the linear and non-linear tasks. More variance in the representations of bias-ReLU networks is explained by the linear task than for non-bias ReLU networks. While ReLU networks without bias also represent features relevant to the linear task more strongly, the early linear alignment of ReLU networks with bias to linear models boosts this phenomenon. (Averages across 10 seeds, error bars 95%-CIs.)

of variance in representations [Lampinen et al., 2024]. To interrogate if such biases are enhanced by linear-like learning we design a task (see Fig. 3A) in which networks have to solve a linear and non-linear problem in parallel. Fig. 3B shows how bias-ReLU networks are aligned to linear networks in early training and initially only learn the linear problem. ReLU networks without bias also appear to learn slower after learning the linear task. However, the corresponding saddle point is less pronounced in these models.

When interrogating model representations we find that early linearity is indeed a key contributor to representational biases. We follow the analysis by Lampinen et al. [2024] and fit linear regressions that predict activations of each hidden layer unit from binary inputs which represent input examples in terms of the linear and nonlinear task. In Fig. 3C we illustrate that ReLU networks with bias represent the linear task more strongly than their bias-free counterparts. However, both models represent linear features more prominently than non-linear features while solving both tasks with zero loss. Intriguingly, we also find that less overall representational variance is explained when fitting a regression that contains regressor for both linear and non-linear features in bias-free ReLU networks, hinting at less structured and task-attuned representations.

Alignment on naturalistic data. We next investigate if our insights also extend beyond simplified learning problems. We train ReLU networks on naturalistic image data, namely MINST and CIFAR-10 (grayscale) classification tasks. We maintain small initial weights, learning rate and squared error loss for these simulations. For MNIST we train models with a single hidden layer and increase hidden depth to two hidden layers for CIFAR-10. For naturalistic data full ablation of bias terms can be challenging as input correlations (e.g. constantly active pixels) can effectively act as bias terms [Rubruck et al., 2025]. To minimize this effect we normalize each input $\mathbf{x}_i$ by the pixel-wise mean over the full dataset, i.e. $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}} = \frac{1}{N}\sum_i^N \mathbf{x}_i$. As before, we find that loss curves of ReLU networks with bias terms are closely aligned to those of linear networks, while loss curves diverge faster when bias terms are ablated. Intriguingly, we observed the match between linear and ReLU networks to only be exact when hidden layer size of linear networks were reduced to be half the size of corresponding ReLU networks. The result indicates that dynamics between linear and ReLU networks are preserved albeit under different learning speeds similar to observations made by [Zhang et al., 2025]. Our results demonstrate that the phenomenon of early linearity of bias-ReLU networks can be observed for networks trained on naturalistic data.

3 Conclusion

We found strong functional and representational alignment between ReLU and linear networks when models are equipped with bias terms. We further demonstrated that this correspondence is not constrained to linear problems but that transient, early alignment can also be observed for non-linear problems. While our work does not yet provide an exact mathematical characterization learning

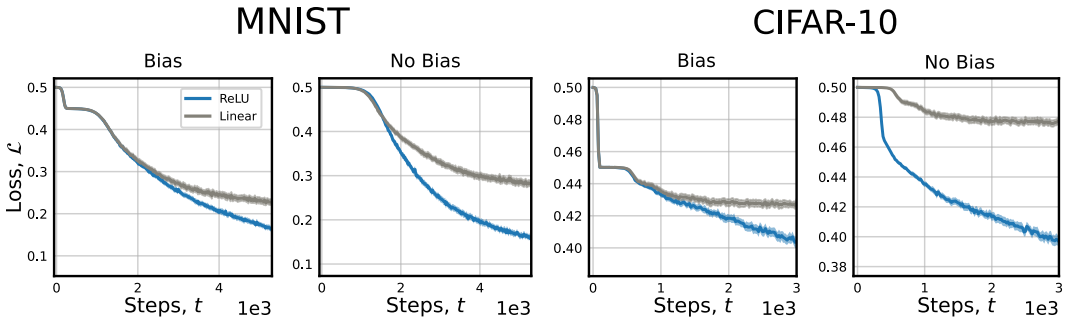

Figure 4: Transient alignment on naturalistic data. We show that bias terms drive alignment on naturalistic image data. Loss curves of linear and linear networks are more closely aligned in early training when models have bias terms (shaded region indicates SE).

dynamics in bias-ReLU networks the close match to exact solutions derived for linear networks indicates that a more exact understanding of ReLU learning dynamics in these cases should be attainable. This bias towards early linearity also induces strict linear before non-linear learning, while enhancing representational biases towards linear features in network hidden activations.

Acknowledgments and Disclosure of Funding

We thank Satwik Bhattamishra and Yedi Zhang for useful feedback and discussions. This work was funded by a Wellcome Trust Discovery Award (227928/Z/23/Z) to CS, and a UKRI ESRC Grand Union Doctoral training partnership stipend awarded to JR. KS is funded by a Cusanuswerk Doctoral Fellowship. This work was also supported by a Schmidt Science Polymath Award to AMS, and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3850). AMS is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program.

References

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. doi: 10.1073/pnas.1403112111. URL https://www.pnas.org/doi/full/10.1073/pnas.1403112111. Publisher: Proceedings of the National Academy of Sciences.

Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11):e1003915, November 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003915. URL https://dx.plos.org/10.1371/journal.pcbi.1003915.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like?, January 2020. URL https://www.biorxiv.org/content/10.1101/407007v2. Pages: 407007 Section: New Results.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3519–3529. PMLR, May 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html. ISSN: 2640-3498.

Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. Similarity of Neural Network Models: A Survey of Functional and Representational Measures. *ACM Computing Surveys*, 57(9):1–52, September 2025. ISSN 0360-0300, 1557-7341. doi: 10.1145/3728458. URL http://arxiv.org/abs/2305.06329. arXiv:2305.06329 [cs].

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Christopher J. Cueva, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nathan Cloos, Nikolaus Kriegeskorte, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. Getting aligned on representational alignment, November 2024. URL http://arxiv.org/abs/2310.13018. arXiv:2310.13018 [q-bio].

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, February 2014. URL http://arxiv.org/abs/1312.6120. arXiv:1312.6120 [cond-mat, q-bio, stat].

Clémentine Carla Juliette Dominé, Nicolas Anguita, Alexandra Maria Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From Lazy to Rich: Exact Learning Dynamics in Deep Linear Networks. October 2024. URL https://openreview.net/forum?id=ZXaocmXc6d.

Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/2b3bb2c95195130977a51b3bb251c40a-Abstract-Conference.html.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, June 2019. doi: 10.1073/pnas.1820226116. Publisher: Proceedings of the National Academy of Sciences.

Yedi Zhang, Andrew M. Saxe, and Peter E. Latham. When Are Bias-Free ReLU Networks Effectively Linear Networks? *Transactions on Machine Learning Research*, January 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=Ucpfdn66k2.

Andrew Kyle Lampinen, Stephanie C. Y. Chan, and Katherine Hermann. Learned feature representations are biased by complexity, learning order, position, and more, September 2024. URL http://arxiv.org/abs/2405.05847. arXiv:2405.05847 [cs].

Jacob S Prince, George A Alvarez, and Talia Konkle. Representation with a capital 'R': measuring functional alignment with causal perturbation.

Andrew Kyle Lampinen, Stephanie C. Y. Chan, Yuxuan Li, and Katherine Hermann. Representation biases: will we achieve complete understanding by analyzing representations?, August 2025. URL http://arxiv.org/abs/2507.22216. arXiv:2507.22216 [q-bio].

Lukas Braun, Erin Grant, and Andrew M. Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. June 2025. URL https://openreview.net/forum?id=YucuAuXMpT.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic Representation Hypothesis, July 2024. URL http://arxiv.org/abs/2405.07987. arXiv:2405.07987.

Minyoung Huh, Hossein Mobahi, Richard Zhang, Brian Cheung, Pulkit Agrawal, and Phillip Isola. The Low-Rank Simplicity Bias in Deep Networks, March 2023. URL http://arxiv.org/abs/2103.10427. arXiv:2103.10427 [cs].

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The Pitfalls of Simplicity Bias in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9573–9585. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6cfe0e6127fa25df2a0ef2ae1067d915-Abstract.html.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on Neural Networks Learns Functions of Increasing Complexity. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Devon Jarvis, Richard Klein, Benjamin Rosman, and Andrew M. Saxe. Make Haste Slowly: A Theory of Emergent Structured Mixed Selectivity in Feature Learning ReLU Networks, March 2025. URL http://arxiv.org/abs/2503.06181. arXiv:2503.06181 [cs].

Andrew Saxe, Shagun Sodhani, and Sam Jay Lewallen. The Neural Race Reduction: Dynamics of Abstraction in Gated Networks. In *Proceedings of the 39th International Conference on Machine Learning*, pages 19287–19309. PMLR, June 2022. URL https://proceedings.mlr.press/v162/saxe22a.html. ISSN: 2640-3498.

Jirko Rubruck, Jan Philipp Bauer, Andrew M. Saxe, and Christopher Summerfield. Early learning of the optimal constant solution in neural networks and humans. July 2025. URL https://openreview.net/forum?id=6Xyu486HRh.

A A Quick review of exact solutions in linear neural networks

We will quickly review the derivation of exact learning dynamics in deep linear networks by Saxe et al. [2014, 2019]. Consider the same setup as outlined in our section 2, **setting**. When training networks using full batch gradient descent in the gradient flow regime dynamics in linear networks are solely dependent on the dataset input-output and input-input correlation matrices. Using singular value decomposition (SVD), these matrices can be expressed as

$$\mathbf{\Sigma}^{yx} = \frac{1}{N}\mathbf{Y}\mathbf{X}^T = \mathbf{U}\mathbf{S}\mathbf{V}^T, \quad \mathbf{\Sigma}^{x} = \frac{1}{N}\mathbf{X}\mathbf{X}^T = \mathbf{V}\mathbf{D}\mathbf{V}^T. \tag{1}$$

Here $\mathbf{X} \in \mathbb{R}^{N_{in} \times N}$ and $\mathbf{Y} \in \mathbb{R}^{N_{out} \times N}$ contain the full set of input and output vectors. Crucially, if the right singular vectors $\mathbf{V}^T$ of $\mathbf{\Sigma}^{yx}$ diagonalize $\mathbf{\Sigma}^{x}$ the full evolution of network weights for deep and shallow networks can be described as

$$\mathbf{W}^2(t)\mathbf{W}^1(t) = \mathbf{U}\mathbf{A}(t)\mathbf{V}^T. \tag{2}$$

Here $\mathbf{A}(t)$ is a diagonal matrix. The evolution of these diagonal values $\mathbf{A}(t)_{\alpha\alpha} = a_\alpha(t)$ at each time-step t then follows a sigmoidal trajectory.

$$a_\alpha(t) = \frac{s_\alpha / d_\alpha}{1 - (1 - \frac{s_\alpha}{d_\alpha a_0})e^{-\frac{2s_\alpha}{\tau}t}} \tag{3}$$

In $s_\alpha = \mathbf{S}_{\alpha\alpha}$ and $d_\alpha = \mathbf{D}_{\alpha\alpha}$ denote the relevant singular values of $\mathbf{\Sigma}^{yx}$ and the eigenvalues of $\mathbf{\Sigma}^{x}$ respectively, a_0 are the singular values at initialization, and $\tau = \frac{1}{N\epsilon}$ is the time constant where ϵ is the learning rate.

B Alignment with bias terms in both layers.

For completeness, we also examine how dynamics are aligned in cases where networks contain bias terms in both layers. I.e. $\hat{\mathbf{y}}_i = \mathbf{W}^2\sigma(\tilde{\mathbf{W}}^1\mathbf{x}_i + \mathbf{b}_1) + \mathbf{b_2}$. We find that dynamics in these cases are also aligned and ReLU networks display characteristic, stage-like learning.

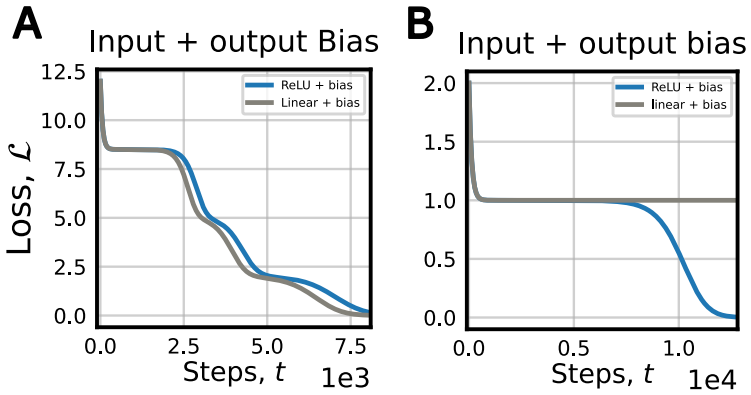

Figure 5: Alignment of ReLU and Linear networks with bias terms in both layers. A. Stage-like learning in ReLU networks that have bias terms on both layers on the hierarchical learning task in Fig. 1A. Dynamics appear preserved but under slightly different time constants. **B.** Early dynamics are also aligned on the non-linear problem from Fig. 1C.