Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks

Lukas Braun
lukas.braun@psy.ox.ac.uk
Department of Experimental Psychology,
University of Oxford, UK

Erin Grant
erin.grant@ucl.ac.uk
Gatsby Unit & SWC
University College London, UK

Andrew M. Saxe
a.saxe@ucl.ac.uk
Gatsby Unit & SWC,
University College London, UK

Abstract

A foundational principle of connectionism is that perception, action, and cognition emerge from parallel computations among simple, interconnected units that generate and rely on neural representations. Accordingly, researchers employ multivariate pattern analysis to decode and compare the neural codes of artificial and biological networks, aiming to uncover their functions. However, there is limited analytical understanding of how a network's representation and function relate, despite this being essential to any quantitative notion of underlying function or functional similarity. We address this question using fully analysable two-layer linear networks and numerical simulations in nonlinear networks. We find that function and representation are dissociated, allowing representational similarity without functional similarity and vice versa. Further, we show that neither robustness to input noise nor the level of generalization error constrain representations to the task. In contrast, networks robust to parameter noise have limited representational flexibility and must employ task-specific representations. Our findings suggest that representational alignment reflects computational advantages beyond functional alignment alone, with significant implications for interpreting and comparing the representations of connectionist systems. [1]

Introduction

The *parallel distributed processing* hypothesis posits that function in artificial and biological networks emerges from interactions among simple interconnected units that compute with distributed representations (Rumelhart et al., 1986). Accordingly, one might aim to identify function from network observables such as connectivity weights and neural activity patterns; however, this is often complicated by the inherent complexity and partial observability of these systems. In particular, the structure of artificial and biological networks is often *non-identifiable* in the sense that networks can be structurally distinct, yet implement the same input-output mapping.

Even deep *linear* networks are non-identifiable. Such networks effectively perform a linear transformation (Laurent & von Brecht, 2018), but do so through multistage computations that give rise to hidden-layer representations. As a result, the optimisation landscape of a deep linear network is non-convex and enjoys a high-dimensional manifold of minima whose shape is determined by the statistics of training data and the network architecture (Arora et al., 2019; Baldi & Hornik, 1989; Saxe et al., 2014), making it a useful surrogate for studying representation learning (Braun et al., 2022; Dominé et al., 2024; Saxe et al., 2019). Here, we leverage the analytical tractability of deep linear networks to study functionally equivalent parametrisations at global minimum error. Crucially, these solutions employ different internal representations, which has significant *computational* consequences, most notably in their affordances for linear decoding, representational similarity analysis and their sensitivity to noise.

[1] A longer version of this work appeared as Braun et al. (2025).

Solution manifold of two-layer linear networks

Deep linear networks are highly overparametrised, admitting many weight configurations that achieve the global optimum—collectively known as the solution manifold. Within this manifold (*e.g.*, , Figure 1A,B), we analytically identify four distinct sub-regions: (a) general linear (GLS), (b) least-squares (LSS), (c) minimum representation norm (MRNS), and (d) minimum weight norm solutions (MWNS). Crucially, all subregions attain the same minimal training error but differ in their hidden-layer representations. While GLS and LSS can achieve almost arbitrary hidden-layer representations (Figure 1D), MRNS and MWNS have hidden-layer representations that appear arbitrary and unstructured, however, their representational similarity matrix (RSM) is fixed and reveals the underlying structure of the task (Figure 1E,F).

Implications for analysis of representations

A common approach to comparing activity patterns across conditions, stimuli, models, or participants is to assess the similarity of their representational geometry (Haxby et al., 2014; Kriegeskorte et al., 2008). However, our analytical results show that in both general and least-squares regimes, geometric relationships may not reflect the underlying computation. Linear predictivity—how well one model's activations linearly predict another's (*e.g.*, Yamins & DiCarlo, 2016; Yamins et al., 2014)—is often taken as evidence of functional similarity. Yet, our analysis reveals that strong linear predictivity does not guarantee functional alignment. As shown in Figure 2A. High R^2 values can arise from predictions within task-agnostic or task-specific solutions, while low values occur between different solution types—highlighting the risk of misinterpreting functional equivalence when solution types are not clearly distinguished.

Further, representational similarity analysis (RSA) on the hidden-layer representations of two random walks on solution manifolds corresponding to different functions yields similarity scores that fluctuate randomly between statistical significance and insignificance (Figure 2B, left). In contrast, representational similarities *are* preserved among minimum weight-norm and among minimum representation-norm solutions, resulting in perfectly correlated representations when comparing ran-

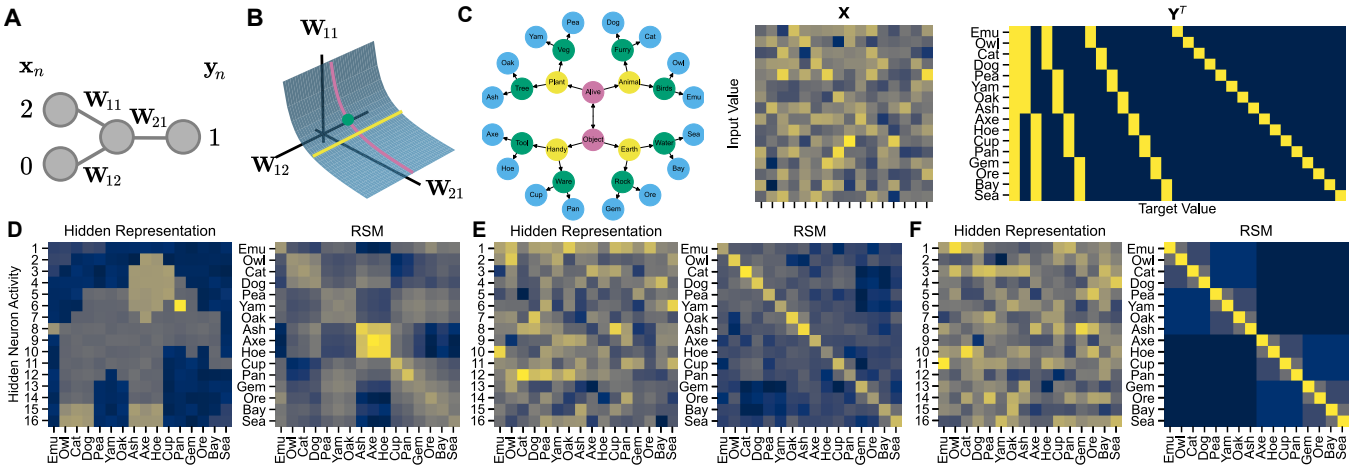

Figure 1: Solution manifold and hidden-layer representations. (A) Schematic of a two-layer linear network with a single training pair, $(\mathbf{x}_n, \mathbf{y}_n)$. The network has three weights which connect the two input, one hidden and one output neurons. (B) Solution manifold of network and task depicted in A. Weight $\mathbf{W}_{12}$ is in the null space of the input and thus can be set to an arbitrary value, whereas $\mathbf{W}_{11}$ and $\mathbf{W}_{21}$ are interdependent, if one of them increases the other one has to decrease accordingly. LSS (pink line), MRN (yellow line), and MWN solutions (green dot) are special subregions of the solution manifold (blue). (C) Schematic of a semantic-hierarchy task. Items (blue) in the semantic-hierarchy task are organised within a binary tree according to their properties (pink, yellow, green). Inputs are random vectors and corresponding target vectors encode for the position in the hierarchical tree. (D) Example least-squares solution, showing hidden-layer representations (left), (RSM, right). While the hidden-layer representations and RSM exhibit structure, they do not reflect the structure of the underlying semantic hierarchy. (E) Same as (C), but for a minimum-weight-norm solution. Here, neural representations are entirely determined by the training data and influenced by the unstructured encoding of input items. As a result, hierarchical structure is only weakly preserved. (F) Same as (C), but for a minimum-representation-norm solution. Neural representations align with the hierarchical structure, seen as structured representational similarity matrix (RSM) and multidimensional scaling plots, where items are grouped according to their similarity within the hierarchy.

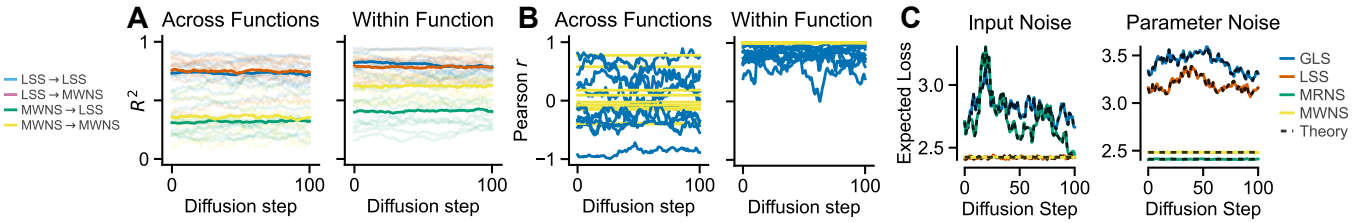

Figure 2: Consequences for neural analyses and the brain. (A) Linear decoding of random walks on the solution manifold of different functions yields high R^2 when predicting neural activity from task-agnostic representations (LSS) but low R^2 when predicting from task-specific representations (MWNS). Thus, it is valid to reject the hypothesis that two networks perform different functions only if both operate in the task-specific regime. Similarly, independent random walks on the same solution manifold maintain high R^2 when predicting task-specific from task-agnostic or task-specific representations, but predicting a task-agnostic network from a task-specific one results in low R^2, leading to invalid rejection of functional equivalence. (B) Same as (A) but using representational similarity analysis (RSA). Comparisons between LSS fluctuate randomly in significance, while task-specific MWNS yield stable and thus reliable measures of similarity. (C) Sensitivity to input and parameter noise. LSS and MRNS have optimal input noise robustness, while MWNS have optimal parameter noise robustness.

dom walks within the same function, however, not when comparing task-agnostic representations (Figure 2C, right).

Advantages of task-specific representations

A natural question arises from the observation that function and representation are dissociated: Why do we often observe representational alignment when comparing artificial and biological systems (Sucholutsky et al., 2023)? To address this, we demonstrate that solutions with task-specific neural representations offer significant computational advantages. Consequently, such solutions are likely to be the preferred functional implementation for both artificial and biological neural networks, resulting in representational alignment. We study the sensitivity of different points on the solution manifold to input and parameter noise (Figure 2C, left), and find that solutions with task-specific representations (MRNS, MWNS) are more robust to parameter noise.

Discussion

In this work, we give a complete analytical characterisation of the global minima manifold for deep linear networks, and demonstrate that sub-regions of this manifold provide differing affordances for computation and interpretation due to their representational structure. We conclude that the use of deep, overparametrised networks poses fundamental challenges for representational analysis, interpretation, and comparison, as the impact of variability in the parametrisation of functionally equivalent representations on these use cases is significant.

Acknowledgements

L.B. was supported by the Woodward Scholarship awarded by Wadham College, Oxford and the Medical Research Council (MR/N013468/1). E.G. and A.S. were supported by a Schmidt Science Polymath Award to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3850).

References

Arora, S., Cohen, N., Golowich, N., & Hu, W. (2019). A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks. *7th International Conference on Learning Representations*

Baldi, P., & Hornik, K. (1989). Neural Networks and Principal Component Analysis: Learning from Examples without Local Minima. *Neural Networks*, *2*(1), 53–58

Braun, L., Dominé, C., Fitzgerald, J., & Saxe, A. (2022). Exact learning dynamics of deep linear networks with prior knowledge. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems 35*

Braun, L., Grant, E., & Saxe, A. M. (2025, July). Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, V. Smith, F. Berkenkamp, & T. Maharaj (Eds.), *Proceedings of the 42nd international conference on machine learning*. PMLR

Dominé, C. C., Anguita, N., Proca, A. M., Braun, L., Kunin, D., Mediano, P. A., & Saxe, A. M. (2024). From lazy to rich: Exact learning dynamics in deep linear networks. *arXiv preprint arXiv:2409.14623*

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual review of neuroscience*, *37*(1), 435–456

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 249

Laurent, T., & von Brecht, J. (2018). Deep linear networks with arbitrary loss: All local minima are global. In J. G. Dy & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning* (pp. 2908–2913, Vol. 80). Pmlr

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, volume 1: Foundations*. MIT Press

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations*

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546

Sucholutsky, I., Muttenthaler, L., Weller, A., Peng, A., Bobu, A., Kim, B., Love, B. C., Grant, E., Groen, I., Achterberg, J., Tenenbaum, J. B., Collins, K. M., Hermann, K. L., Oktar, K., Greff, K., Hebart, M. N., Jacoby, N., Zhang, Q., Marjieh, R., . . . Griffiths, T. L. (2023). Getting aligned on representational alignment

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using Goal-Driven Deep Learning Models to Understand Sensory Cortex. *Nature Neuroscience*, *19*(3), 356–365

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624