# A Mathematical Theory of Context Controlled Semantic Development

**Devon Jarvis (devon.jarvis@wits.ac.za)**
School of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Ave, Braamfontein, Johannesburg, 2000, South Africa

**Richard Klein (richard.klein@wits.ac.za)**
School of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Ave, Braamfontein, Johannesburg, 2000, South Africa

**Benjamin Rosman (Benjamin.Rosman1@wits.ac.za)**
School of Computer Science and Applied Mathematics, University of the Witwatersrand, 1 Jan Smuts Ave, Braamfontein, Johannesburg, 2000, South Africa

**Andrew Saxe (a.saxe@ucl.ac.uk)**
Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, University College London, 25 Howland Street, London W1T 4JG, United Kingdom

## Abstract

**A number of phenomena which underlie human semantic cognition, and emerge during childhood, have been established. Recent work has provided a mathematical theory for the context unaware phenomena, using deep linear neural networks. Here we extend this theory to encompass aspects of cognitive control in semantic learning. This follows later in a child's development and requires the ability to "gate" aspects of the cognitive computation. Gating is a nonlinear process where portions of the computation are inhibited in some contexts. We use a neural network with ReLU activation to perform the gating and model three more behaviours in semantic development, namely domain specific attribute weighting, new attribute induction and conceptual reorganisation. We use a Gated Deep Linear Network to model the ReLU network, providing the full training dynamics and interpretability in its implementation of cognitive control. We find that the ReLU network uses an intricately structured latent representation which is mixed selective. Thus, we demonstrate how reusable, generalizable and mixed-selective latent representations may emerge, three properties which have previously seemed incongruent.**

**Keywords:** Controlled Cognition, Semantic Learning, Gated Linear Networks, Mixed Selectivity

## Introduction

The development of semantic knowledge throughout childhood has been studied extensively, yielding a number of identified empirical phenomena which underlie human cognition (McClelland & Rogers, 2003; Luna, 2009; Crone & Steinbeis, 2017). For example, 1) learning happens rapidly following a slow initial period; 2) knowledge is acquired in order of generality (progressive differentiation); and 3) the absence of fine-grained distinctions early in learning results in general concepts being extended to all objects within too broad a class (illusory correlations). Recent work provided a mathematical theory for semantic development based on deep neural networks and modelled these three phenomena, which result from the interplay of dataset statistics and the inductive biases of neural architectures (A. M. Saxe et al., 2019). Yet, this theory used the linear activation function to maintain analytical tractability of the neural network learning dynamics. As a consequence, that theory could not capture the many context-dependent effects observed in semantic development (Monchi et al., 2001; Fuster, 2001; Friedman & Robbins, 2022). In this work we aim to extend this mathematical theory to encompass nonlinear network activations which are necessary to incorporate contextual control into semantic cognition (Rumelhart, 1990; Rumelhart & Todd, 1993). We focus on modelling three additional semantic phenomena: 4) domain specific attribute weighting – attending to different information in distinct contexts (domains); 5) new attribute induction – limiting the acquisition of new knowledge to the context in which it is introduced; 6) conceptual reorganisation – flexibly remapping the relationship between objects based on context (McClelland & Rogers, 2003).

## Method

Similar to prior work we create a task where the network is provided an item as input and required to produce a set of corresponding features (A. M. Saxe et al., 2019; Braun et al., 2022). Figure 1 summarizes our entire setup. Each item is queried with a one-hot representation and one-hot context feature, such that all items are present in each context. The features then impart structure into the dataset based on the similarity of items. For example, the output labels from "Alive" to "Petals" form a hierarchy structure. All items are alive, only animals can move, only birds have feathers and so on. This block of features should be activated regardless of the queried context. In contrast the three other blocks of output labels need to be activated only in one of the contexts which requires cognitive control and a nonlinear network mapping. We train a network with full-batch gradient descent and quadratic loss to perform this task. Importantly, the one hidden layer of the

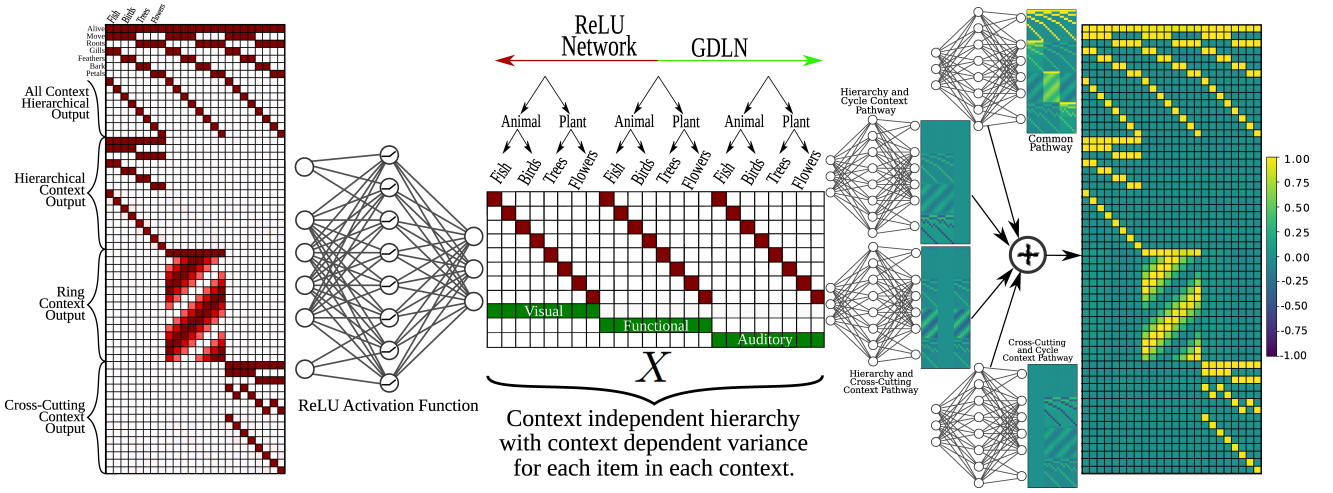


Figure 1: Dataset used to train the ReLU network (left) and GDLN architecture used to imitate it (right). Inputs (middle matrix) are created by appending a one hot vector encoding object identity to a one hot vector encoding context such that each item appears in each context. Target outputs (left and right matrices) contain some context-independent (top block) and some context specific properties (bottom three blocks). These datasets broadly follow a hierarchical structure across items (hierarchical tree depicted in middle over input columns), but with some variation in each context-specific block. All structures are taken from A. M. Saxe et al. (2019). The analysis in this work shows that the ReLU network dynamics arise from four implicit modules, made explicit by the GDLN pathways towards the right, which receive different subsets of inputs and generate different subsets of outputs. Together these graded mixed-selective pathways couple together to produce the correct output labels for each object. While each context-specific pathway is only on in two contexts (blocks of columns) they still produce labels for all three context-specific parts of the output space (blocks of rows). This creates errors which other pathways learn to remove. If this fine balance of excitation and inhibition is broken then errors will be incurred.

network uses the ReLU activation function which imparts the ability to gate on portions of the computation. We use a linear output layer and do not regularize or bias the network towards context specificity in its hidden neurons. Finally, we also implement a Gated Deep Linear Network (GDLN) (A. Saxe et al., 2022) using the same setup with the aim of replicating the ReLU network's gating pattern and learning trajectory. The only difference between the two models is that the gating of the GDLN's hidden neurons is explicit and present from the start, while the ReLU network learns the gating pattern using the nonlinearity. By imitating the ReLU network the GDLN provides interpretability into how the ReLU network implements controlled semantic cognition.

## Results

We find that the ReLU network is able to model all six old and new phenomena of controlled semantic cognition. Figure 2(a) depicts the loss trajectory for the ReLU network on the three context task and we note that it achieves zero error by the end of training. This demonstrates Domain Specific Attribute Weighting as an appropriate gating pattern is learned. By analysing the loss trajectory we note sudden drops in loss, progressive differentiation (each drop gets smaller) and illusory correlations prior to convergence, demonstrated in Figure 2(b). Figure 2(c) depicts new attribute induction as in each context one new attribute is introduced for only the first item, generalised within the context and the activation strength is reflective of how similar each item is to the first. Finally, Figure 2(d) depicts the multi-dimensional scaling (MDS) plot of the network's hidden layer over time. Three distinct clusterings occur which have different relative positions of the eight items based on which context is queried, demonstrating conceptual reorganisation.

From Figure 2(a-d) we see that the GDLN loss trajectory exactly match that of the ReLU network and demonstrates all phenomena of controlled semantic cognition. We find that four partitions of the GDLN hidden layer are required to imitate the ReLU network. One portion of the hidden layer is gated on for all contexts, while the remaining three are gated on for two contexts. The output of each linear pathway through the network is depicted in Figure 1 (right). The predicted and simulated Singular Value Decomposition (SVD) for each GDLN pathway is depicted in Figure 2(e-f), giving rise to the predicted training dynamics in Figure 2(a). It can be proven that this is the unique GDLN which has this loss trajectory, and consequently this is the gating strategy implemented by the ReLU network. Thus, we have full training dynamics for the ReLU network in terms of the SVD of each of its effective modules. Thus, the inductive bias of gradient descent, paired with the dataset statistics and need for contextual control results in an intricate latent representation with structured mixed selectivity. This emerges since this is the architecture which minimizes the loss quickest by sharing computational units, as proposed by the neural race hypothesis (A. Saxe et al., 2022). Neural reuse has also been noted as a potential organisational principal for the human brain (Anderson, 2010), pointing further to the potential generality of our findings and the biological plausibility of this aspect of the ReLU network.

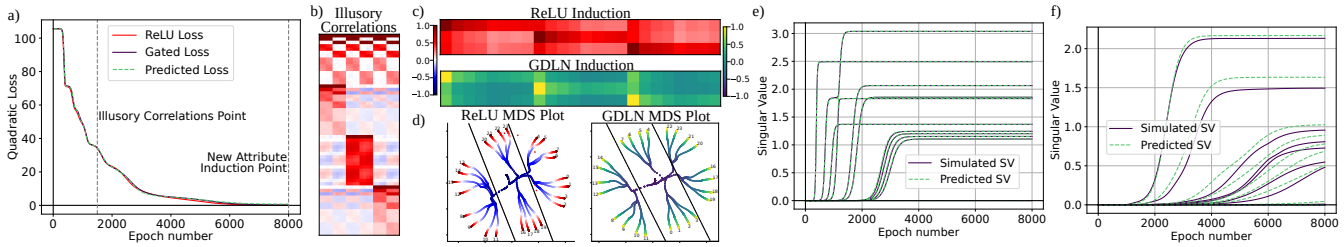


Figure 2: Summary of Results: a) Comparison between the predicted GDLN, actual GDLN and ReLU loss trajectory. Predicted loss dynamics are based on a GDLN which explicitly models the implicit pathways in a ReLU network due to the nonlinearity. b) Example of illusory correlations from the ReLU network output after $1500$ epochs of training. Features from the higher levels of the hierarchy are incorrectly attributed to all items in too broad of a category (eg: all birds can fly). c) Induction (context specific generalization) of one new feature per context (row) for the first object (column $0$, $8$, $16$) for the GDLN and the ReLU network. d) Conceptual Reorganisation: The context specific hidden layer rapidly changes the relative placement of the item in the latent space for each context based on the different feature commonalities between items. Gray lines indicate potential context boundaries in the projected space. Colour depicts progression in time. e and f) Singular Value dynamics comparing the predicted (dashed green) and empirical (solid purple) learning dynamics for the two pathway types of a GDLN model. e) depicts the dynamics of the common pathway and f) depicts the averaged dynamics of all context pathways together.

## Acknowledgments

This work was supported by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (216386/Z/19/Z) to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755). D.J. is a Google PhD Fellow and Commonwealth Scholar. A.S. and B.R. are CIFAR Azrieli Global Scholars in the Learning in Machines & Brains program.

## References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(4), 245–266.

Braun, L., Dominé, C., Fitzgerald, J., & Saxe, A. (2022). Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, *35*, 6615–6629.

Crone, E. A., & Steinbeis, N. (2017). Neural perspectives on cognitive control development during childhood and adolescence. *Trends in cognitive sciences*, *21*(3), 205–215.

Friedman, N. P., & Robbins, T. W. (2022). The role of prefrontal cortex in cognitive control and executive function. *Neuropsychopharmacology*, *47*(1), 72–89.

Fuster, J. M. (2001). The prefrontal cortex—an update: time is of the essence. *Neuron*, *30*(2), 319–333.

Luna, B. (2009). Developmental changes in cognitive control through adolescence. *Advances in child development and behavior*, *37*, 233–278.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience*, *4*(4), 310–322.

Monchi, O., Petrides, M., Petre, V., Worsley, K., & Dagher, A. (2001). Wisconsin card sorting revisited: distinct neural circuits participating in different stages of the task identified by event-related functional magnetic resonance imaging. *Journal of Neuroscience*, *21*(19), 7733–7741.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In *An introduction to neural and electronic networks* (pp. 405–420).

Rumelhart, D. E., & Todd, P. M. (1993). Ti learning and connectionist representations.

Saxe, A., Sodhani, S., & Lewallen, S. J. (2022). The neural race reduction: Dynamics of abstraction in gated networks. In *International conference on machine learning* (pp. 19287–19309).

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.