A Mathematical Theory of Semantic Development

Andrew M. Saxe, James L. McClelland and Surya Ganguli

Stanford University

Overview

A wide array of psychology experiments have revealed remarkable regularities in the developmental time course of human cognition. For example, infants generally acquire broad categorical distinctions (i.e., plant/animal) before finer-scale distinctions (i.e., dog/cat), often exhibiting rapid, or stage-like transitions during learning. What are the theoretical principles underlying the ability of neuronal networks to discover categorical structure from experience?

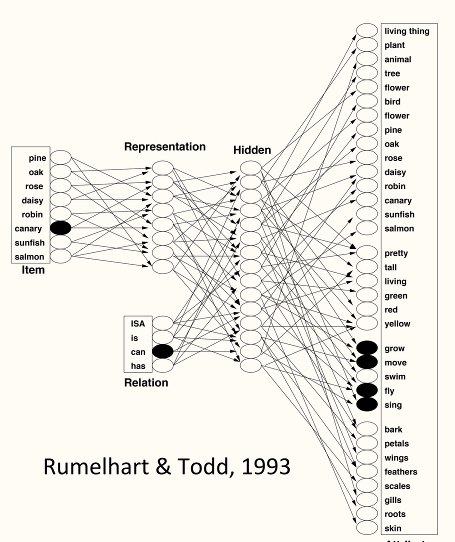
We develop a mathematical theory of hierarchical category learning through an analysis of the learning dynamics of multilayer networks exposed to hierarchically structured data. Our theory yields new exact solutions to the nonlinear dynamics of error correcting learning in deep, three layer networks. These solutions reveal that networks learn input-output covariation structure on a time scale that is inversely proportional to its statistical strength.

We further analyze the covariance structure of data sampled from hierarchical probabilistic generative models, and show how such models yield a hierarchy of input-output modes of differing statistical strength, leading to a hierarchy of time-scales over which such modes are learned.

Our results reveal that even the second order statistics of hierarchically structured data contain powerful statistical signals sufficient to drive complex experimentally observed phenomena in semantic development, including progressive, coarse-to-fine differentiation of concepts and sudden, stage-like transitions in performance punctuating longer dormant periods.

Models of semantic development

Many neural network simulations have captured aspects of broad empirical patterns in semantic development (Rumelhart & Todd, 1993; Rogers & McClelland, 2004)



Rumelhart & Todd, 1993

Main idea: Semantic knowledge arises from incrementally learning about the properties of items, e.g., that a

- *'Canary* can *grow'*
- *'Rose* has *petals'*
- *'Salmon* can *swim'*

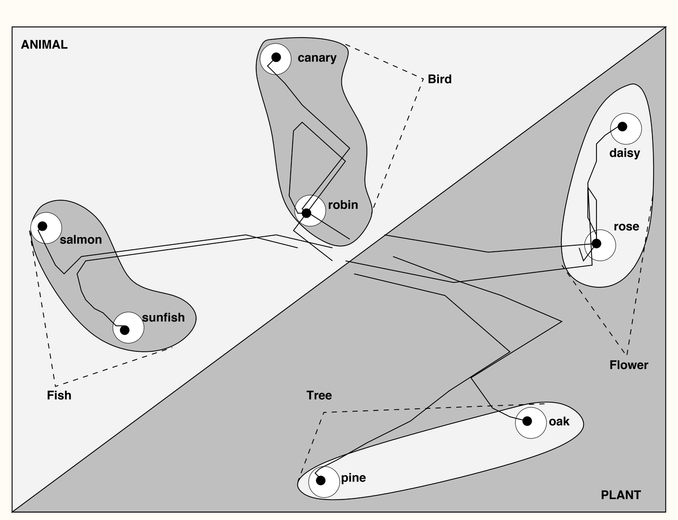
Instantiated in, e.g., the Rumelhart network

- *Item* layer codes input item
- *Hidden* layers learn internal representations
- *Attribute* layer codes output properties

The internal representations of such networks exhibit both **progressive differentiation** and **stage-like transitions**.

Trajectory of internal representations during learning obtained through simulation

(Cf. our analytical results, right)



However the theoretical basis for the ability of neuronal networks to exhibit such strikingly rich nonlinear behavior remains elusive. What are the essential principles that underlie such behavior?

Gradient descent dynamics in multilayer neural networks

Problem formulation

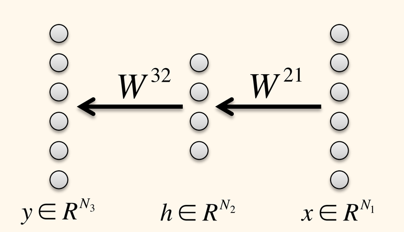
We analyze a fully linear three layer network $y = W^{32}W^{21}x$

trained on patterns

$$\{x^\mu, y^\mu\}, \mu = 1, \ldots, P.$$

via gradient descent on

$$\sum_\mu \left\| y^\mu - W^{32}W^{21}x^\mu \right\|^2.$$



This yields weight dynamics

$$\tau \frac{d}{dt} W^{21} = W^{32^T}\left(\Sigma^{31} - W^{32}W^{21}\Sigma^{11}\right)$$
$$\tau \frac{d}{dt} W^{32} = \left(\Sigma^{31} - W^{32}W^{21}\Sigma^{11}\right) W^{21^T}$$

Input correlations: $\Sigma^{11} \equiv E[xx^T]$

Input-output correlations: $\Sigma^{31} \equiv E[yx^T]$

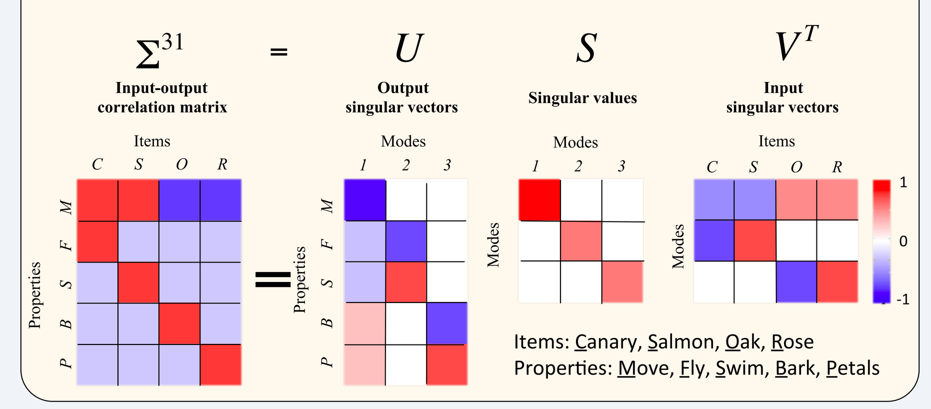
- Depends only on second order statistics of training data
- Coupled and nonlinear (despite linear input-output map)

Decomposing input-output correlations

The learning dynamics can be expressed using the SVD of Σ^{31}

$$\Sigma^{31} = U^{33} S^{31} V^{11^T} = \sum_{\alpha=1}^{N_1} s_\alpha u^\alpha v^{\alpha T}$$

Mode α links a set of coherently covarying properties u^α to a set of coherently covarying items $v^{\alpha T}$ with strength s_α



Items: Canary, Salmon, Oak, Rose
Properties: Move, Fly, Swim, Bark, Petals

Analytical learning trajectory

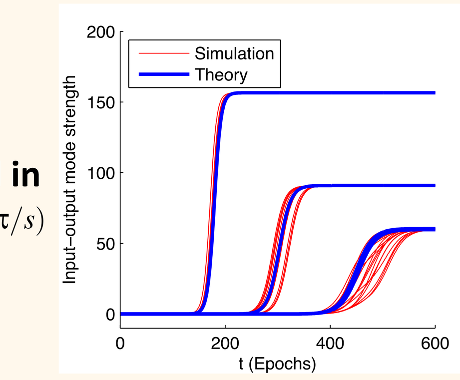
The network's input-output map is exactly

$$W^{32}(t)W^{21}(t) = \sum_{\alpha=1}^{N_2} a(t, s_\alpha, a_\alpha^0)\, u^\alpha v^{\alpha T}$$

where $a(t, s, a_0) = \dfrac{s e^{2st/\tau}}{e^{2st/\tau} - 1 + s/a_0}$

for a special class of initial conditions and $\Sigma^{11} = I$.

- Each mode evolves independently
- Each mode is **learned in time** $t(s, \varepsilon) = \dfrac{\tau}{s} \ln \dfrac{s}{\varepsilon} = O(\tau/s)$
- Eqns. give good approximation for small random initial conditions



Singular values and vectors of hierarchically generated data

A hierarchical diffusion process

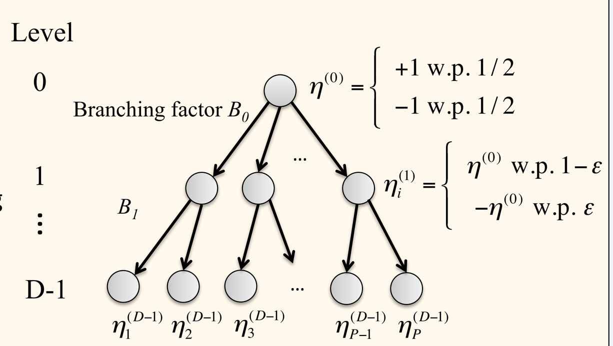
We consider training a neural network with data generated by a hierarchical generative model

We compute the statistical quantities $(s_\alpha, u^\alpha, v^\alpha)$ that drive learning, thereby explicitly linking hierarchical structure to learning dynamics

Generative model defined by *D*-level tree

Feature values diffuse down tree with small probability ε of changing along each link

Sampled *independently* *N* times to produce *N* features



Items assigned features based on values at leaves

Singular values

The *normalized inner product* between items, $q_k = \frac{1}{N}\sum_{i=1}^{N} y_i^{\mu_1} y_i^{\mu_2}$

depends only on the level *k* of their nearest common ancestor.

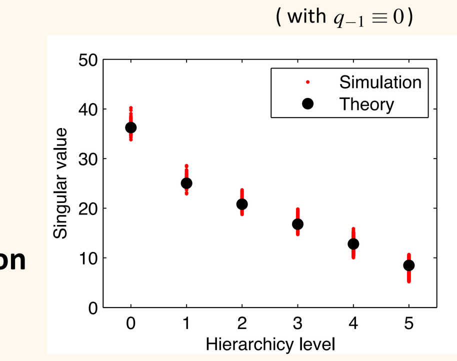
In particular, $q_k = (1 - 4\varepsilon(1-\varepsilon))^{D-1-k}$ which decreases with *k*.

Let M_l be the number of nodes at level *l*, and $\Delta_l \equiv q_l - q_{l-1}$ (with $q_{-1} \equiv 0$)

Then the **singular values** are

$$s_\alpha = \sqrt{NP \sum_{\ell=\alpha}^{D-1} \frac{\Delta_\ell}{M_\ell}}$$

which is **a *decreasing* function of the hierarchy level**.

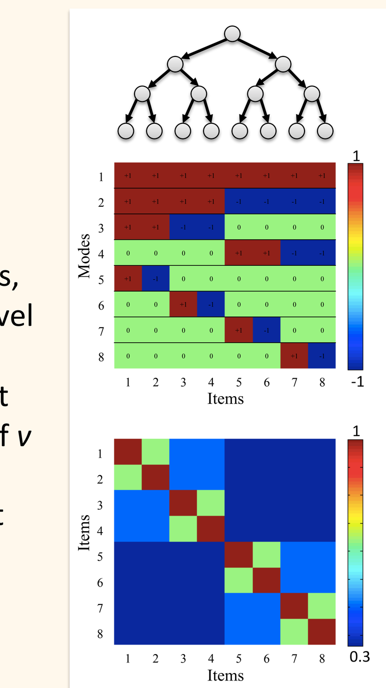


Singular vectors

The **singular vectors mirror the tree structure**

- Vectors at level *l* cannot detect finer-scale distinctions
- They come in M_{l-1} families, one for each node *v* at level *l-1*
- Each vector is zero except on the B_{l-1} descendants of *v*
- The nonzero values are induced by functions that sum to 0.

For a binary tree structure, this yields Haar wavelets

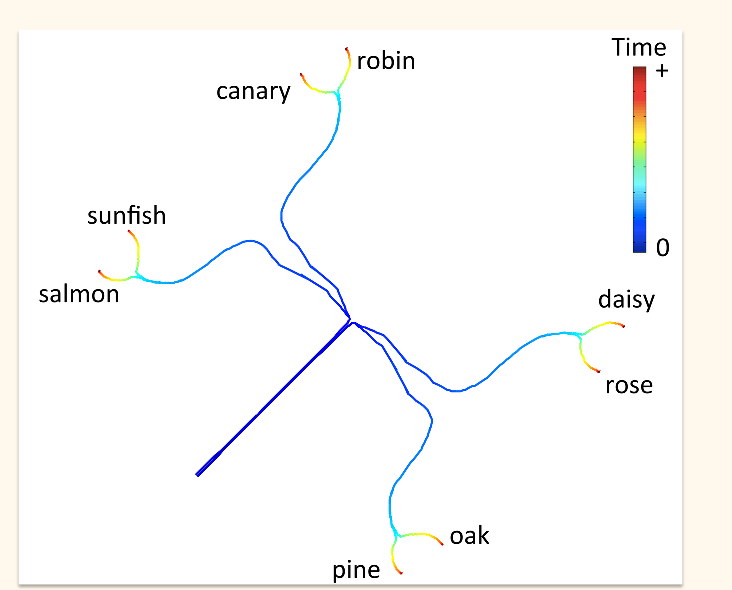


Progressive differentiation of hierarchical structure

Network **must** exhibit progressive differentiation on **any** dataset generated by this class of hierarchical diffusion processes:

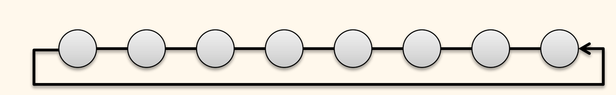
- Network learns input-output modes in time $O(\tau/s)$
- Singular values of broader hierarchical distinctions are larger than those of finer distinctions
- Input-output modes correspond exactly to the hierarchical distinctions in the underlying tree

MDS plot of internal representations during learning



Spatially-structured data

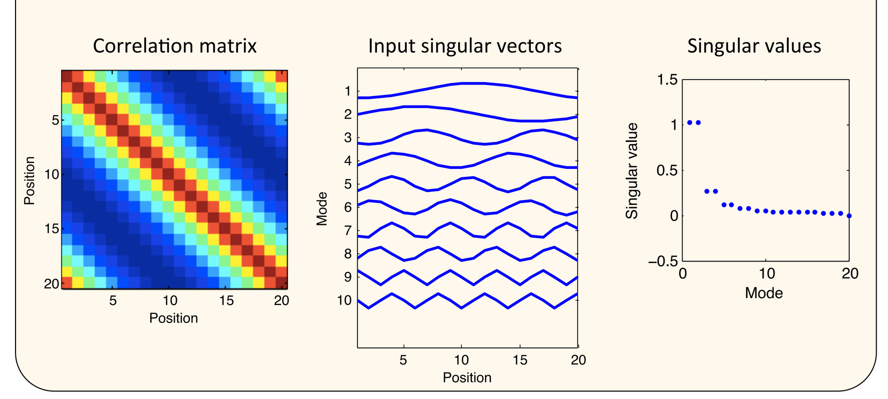
Approach can be extended to other sorts of structure



E.g., ring-structured Gaussian Markov random field

Covariance matrix is *circulant*

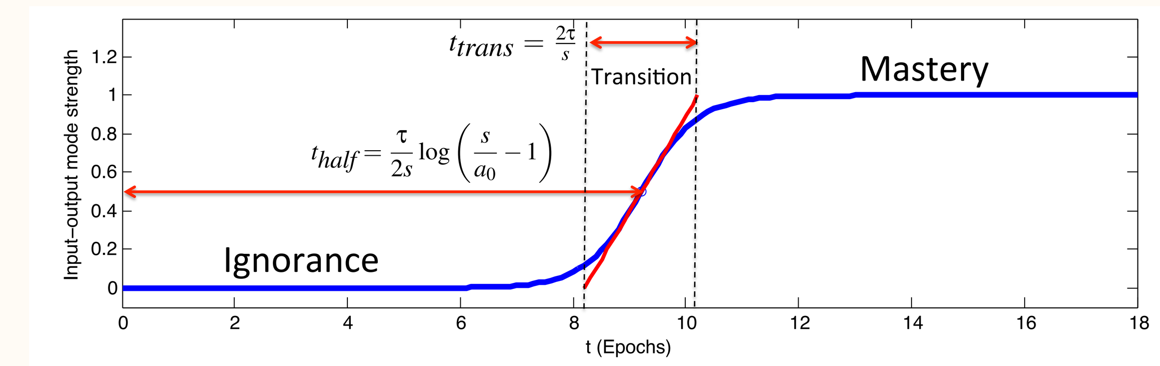
- Singular vectors are Fourier basis
- Singular values are related to Fourier coefficients
- Network learns about broad spatial information first



Stage-like transitions in learning

Empirical evidence suggests transitions during learning can be rapid and stage-like

- Our model exhibits such transitions
- Intuitively, arises from sigmoidal learning trajectories
- The ratio of the *transition period* to the *time to half-mastery* can be arbitrarily small

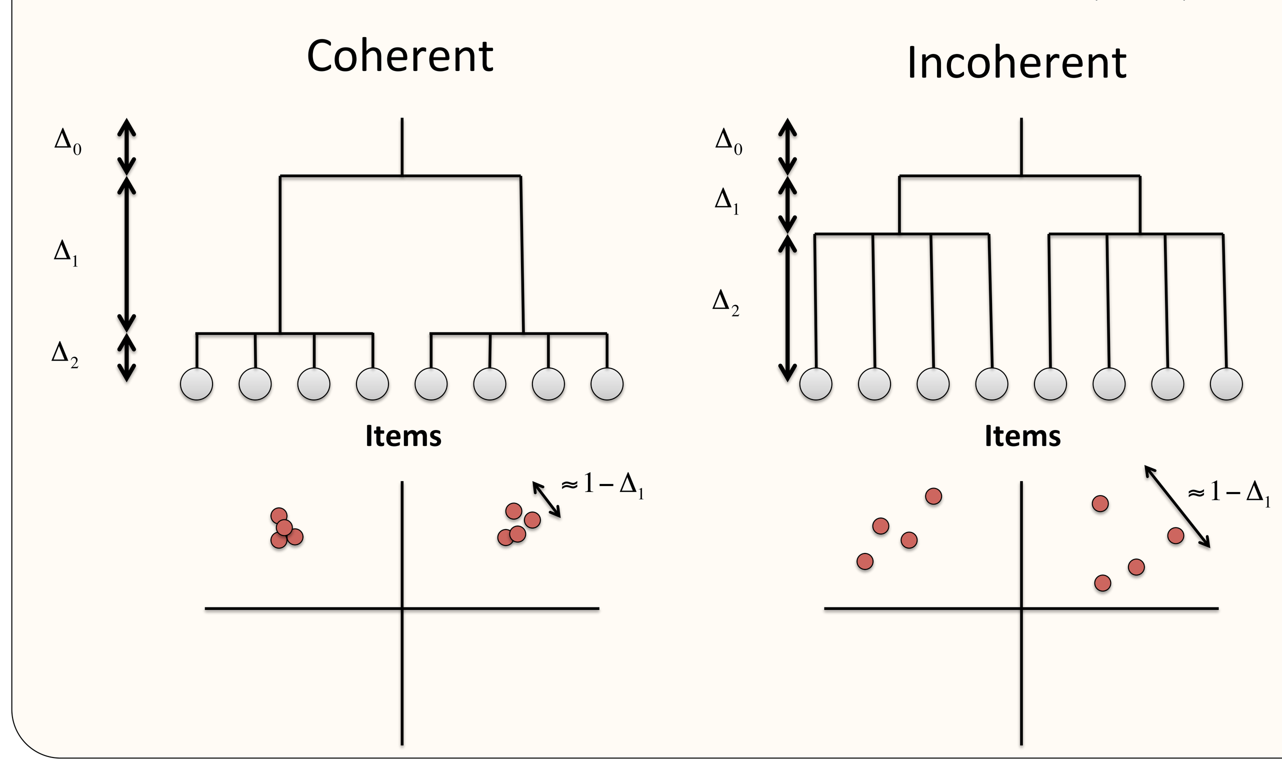


- Critically dependent on deep network structure; networks with no hidden layer do not exhibit stage-like transitions

Category coherence

Our analysis suggests $\Delta_l \equiv q_l - q_{l-1}$ as a quantitative measure of category coherence

- Measures overlap between siblings at level *l* vs level *l-1*.
- For large branching factors, directly relates to learning timescale: $\tau_l = O\left(\sqrt{\frac{M_l}{\Delta_l}}\right)$



Conclusion

Progressive differentiation of hierarchical structure is a general feature of learning in deep neural networks

Deep (but not shallow) networks exhibit **stage-like transitions** during learning

In a position to analytically understand many phenomena previously simulated

- Illusory correlations early in learning
- Familiarity and typicality effects
- Inductive property judgments
- 'Distinctive' feature effects
- Basic level effects
- Category coherence
- Perceptual correlations
- Practice effects

Our framework **connects probabilistic models** and **neural networks**, analytically linking structured environments to learning dynamics.

S.G. thanks DARPA, BioX, Burroughs-Wellcome, and Swartz foundations for support. J.L.M. was supported by AFOSR. A.S. was supported by a NDSEG Fellowship and MBC Traineeship. We thank Juan Gao and Jeremy Glick for useful discussions.