Gradient descent dynamics in multilayer neural networks

Problem formulation
We analyse a fully linear three layer network $y = \langle \langle x, h \rangle, x \rangle$ trained on patterns $(x, y) = (r, f)$ via gradient descent on $E$, the squared error.

Decomposing input-output correlations
The learning dynamics can be expressed using the SVD of $S$: $S = U \Sigma V^T$. Mode $i$ limits a set of coherently varying properties $y^i$ of a set of coherently varying items $x^i$ with strength $\sigma_i^2$.

Analytical learning trajectory
The network's input-output map is exactly
$$x^i \rightarrow \langle \langle x, h \rangle, x \rangle \rightarrow \langle \langle \langle x, h \rangle, h \rangle, x \rangle \rightarrow \langle \langle \langle \langle x, h \rangle, h \rangle, h \rangle, x \rangle$$
for a special class of initial conditions and $S_1\| = 1$.

• Each mode evolves independently.
• Each mode is learned in time $t \approx \sqrt{\frac{1}{\sigma_i^2}}$.
• Eps. give good approximations 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.

Singular values
The normalized inner product between items, $\langle x, y \rangle = \sum_i \sigma_i x^i y^i$, depends only on the level $l$ of their nearest common ancestor.

Singular vectors
The singular vectors mirror the tree structure:
- Vectors at level $l = 0$ detect fine-scale distinctions.
- They come in $M_0$ families, one for each node $v$ at level $l = 0$.
- Each vector is zero except on the $R_0$ descendants of $v$.
- The nonzero values are induced by functions that sum to 1.

For a binary tree structure, this yields hourglasses.

Spatially-structured data
Approach can be extended to other sorts of structure.

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

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 studied
- Recency and familiarity effects.
- Induction property judgments.
- "Distractive" feature effects.
- "Category" coherence effects.
- Perceptual correlations.
- Practice effects.

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

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 aperiodical learning trajectories.
- The ratio of the transition period to the time in hell mastery can be arbitrarily small.

Models of semantic development

Many neural network simulations have captured aspects of broad empirical patterns in semantic development (Bower & Tuck, 1994; Regier & McClelland, 2006)

Main idea: Semantic knowledge arises from incrementally learning about the properties of items, e.g., a dog.
- "Canary, salmon, oak, and rose." (Simplex).
- "Has petals" (property).
- "Salmon can swim." (relation).

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 (GT: our analytical results; white)

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?