A deep learning theory of perceptual learning dynamics

Andrew M. Saxe

Stanford University

Overview

With practice, humans and other organisms dramatically improve their accuracy in simple perceptual discriminations. Experiments have reported learning-induced changes in neural tuning at many levels of the cortical hierarchy, and the magnitude of changes within an area has been found to depend strongly on its position in the hierarchy. A fundamental challenge for theory is to understand this distribution of changes across brain areas.

Here we propose that depth—the brain's layered structure—is a key factor controlling the size and timing of neural changes across the cortical hierarchy.

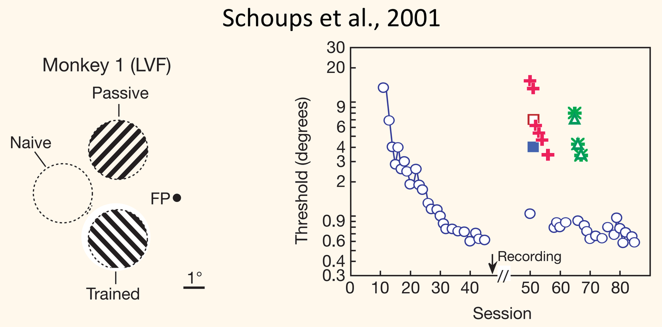
We construct a quantitative, analytic theory of perceptual learning by analyzing gradient descent dynamics in deep linear neural networks. Deep networks exhibit several learning pathologies, including nonconvexity, nonlinear coupling, and scaling symmetries, which strongly impact learning dynamics.

Our results uncover a fundamental dichotomy between learning in 'shallow' parallel structure and 'deep' serial structure: learning in parallel structures targets the 'most informative neurons,' while learning in serial structures targets the 'least informative layers.'

The model's predictions accord with a diverse set of experimental findings, including the pattern of changes within layers; the size and timing of changes across layers; the effects of high precision vs low precision tasks; and the transfer of performance to untrained locations.

Simple perceptual discriminations

Practicing orientation discrimination leads to improved behavioral performance



Ahissar & Hochstein, 1997
Dosher & Lu, 1998
Schoups et al., 2001
Ghose et al., 2002
Yang & Maunsell, 2004
Raiguel et al., 2006
Adab et al., 2011

A wealth of experiments have documented

- Changed neural representations in multiple cortical areas (IT, V4, V1)
- Much larger changes in higher areas than lower areas

Remarkably, V1 changes are modest even after extensive training on high precision orientation discrimination tasks

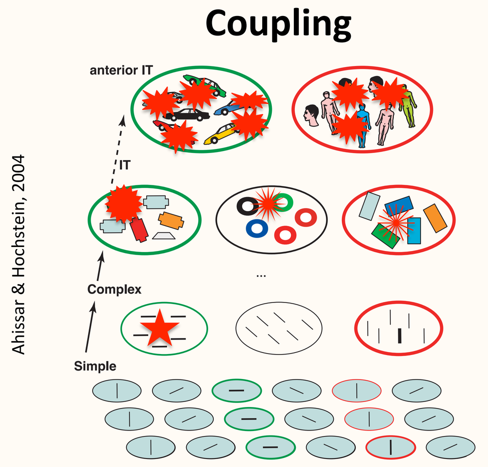
What determines the distribution of changes across cortical areas?

Pathologies of depth

We propose that depth—the brain's layered structure—is a key factor controlling the size and timing of neural changes across the cortical hierarchy.

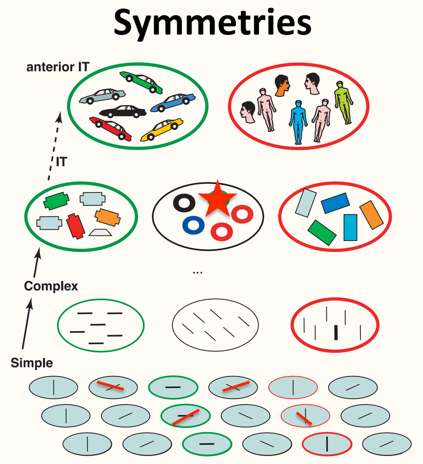
Depth substantially complicates the learning process (Hochreiter, 1991; Bengio et al., 1994) by introducing nonconvexity, vanishing gradients, nonlinear coupling, and scaling symmetries

Coupling



Changing V1 disrupts subsequent layers

Symmetries



Many equivalent changes

Learning in a deep network must overcome these difficulties, yielding dramatically different learning dynamics in comparison to shallow networks (Saxe et al., 2014).

Deep linear network model

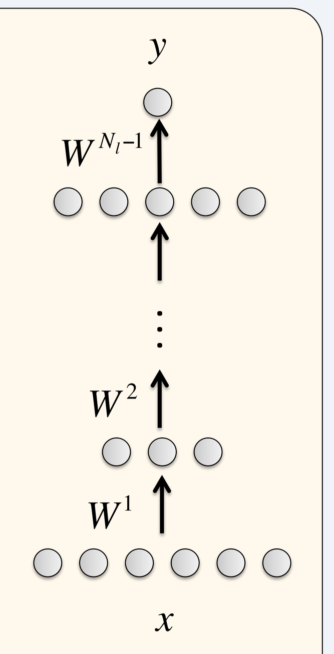
Seek quantitative theory:

- Which layer changes most?
- Which layer changes first?
- What changes in each layer?

Develop theory using simple model class: deep linear neural networks

- Minimal, tractable model
- Rich nonlinear learning dynamics
- Isolates specific impact of depth

Clarifies conceptual picture underlying nonlinear networks (Saxe et al., 2014; Dauphin et al., 2014; Goodfellow et al., 2015)



Gradient descent dynamics

Input-output map: Linear

$$y = \left(\prod_{i=1}^{D} W^i\right) x \equiv W^{tot} x$$

Objective function: Nonlinear; Nonconvex

$$\sum_{\mu} \left\| y^{\mu} - \left(\prod_{i=1}^{D} W^i\right) x^{\mu} \right\|^2$$

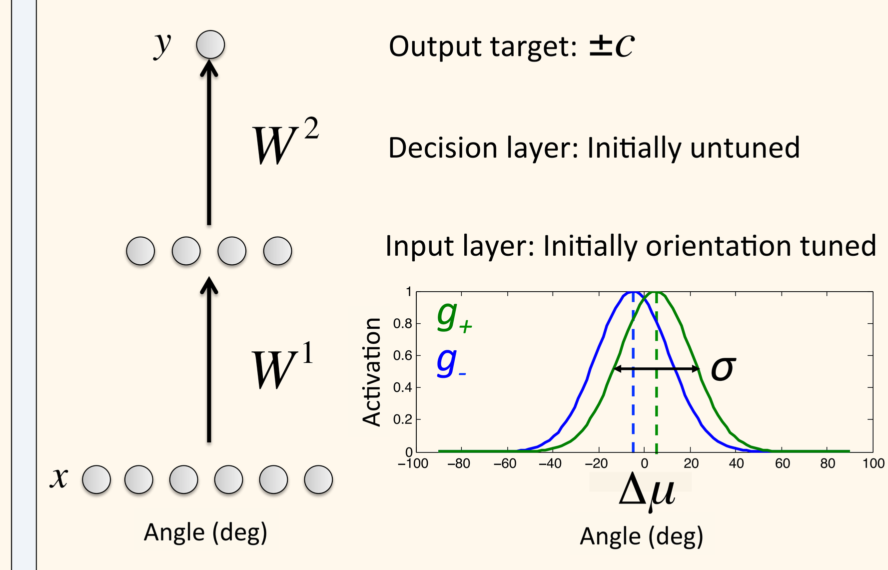
Gradient descent dynamics: Nonlinear; Coupled

$$\Delta W^l = \lambda \sum_{\mu=1}^{P} \left(\prod_{i=l+1}^{D} W^i\right)^T \left[y^{\mu} x^{\mu T} - \left(\prod_{i=1}^{D} W^i\right) x^{\mu} x^{\mu T} \right] \left(\prod_{i=1}^{l-1} W^i\right)^T$$

$$l = 1, \cdots, D$$

Useful for studying ***learning dynamics***, not increased representational power.

Three layer model



Exact reduction to two variables

Solving these dynamics is hard in general; previous solutions (Fukumizu, 1998; Saxe et al., 2014) do not allow initially orientation-tuned neurons.

Here we give an exact reduction to two variables:

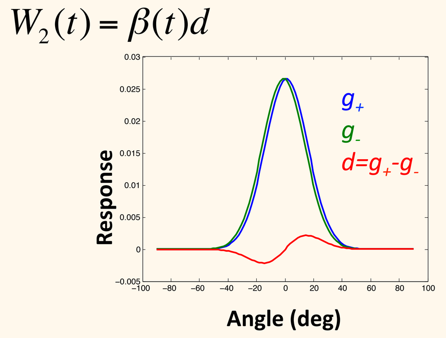
$\alpha(t)$ = size of change in input layer $\quad W_1(t) = \begin{bmatrix} g_+ & g_- \end{bmatrix} + \alpha(t) \begin{bmatrix} d & -d \end{bmatrix}$

$\beta(t)$ = size of change in decision layer $\quad W_2(t) = \beta(t) d$

With dynamics:

$$\tau \frac{d}{dt}\alpha = \beta(c - \nu\beta(1+2\alpha))$$

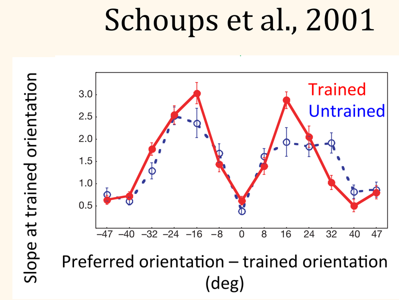
$$\tau \frac{d}{dt}\beta = (c - \nu\beta(1+2\alpha))(1+2\alpha)$$

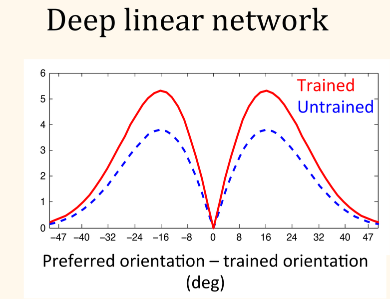


Where the constant $\nu = d^T g_+ = g_+^T g_+ - g_+^T g_-$ encodes task difficulty

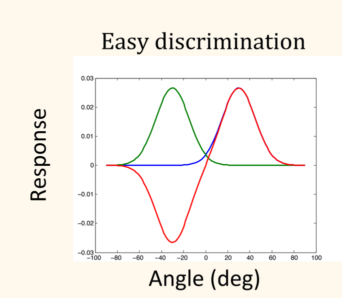
Dynamics of learning

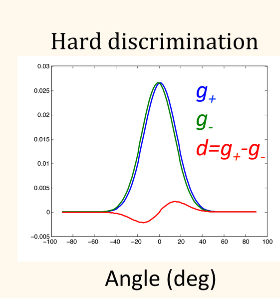
Within layer: most informative neuron changes most

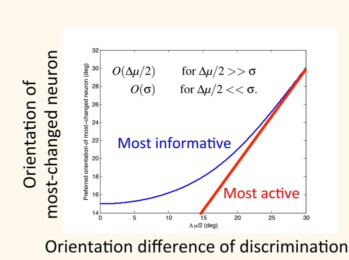




Task precision determines most informative neuron



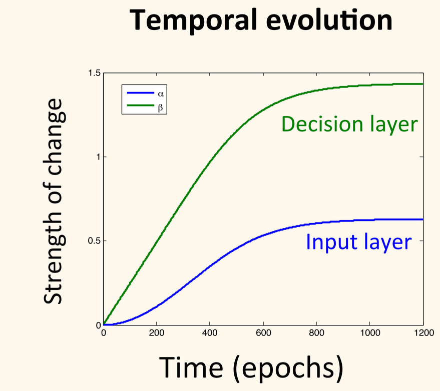




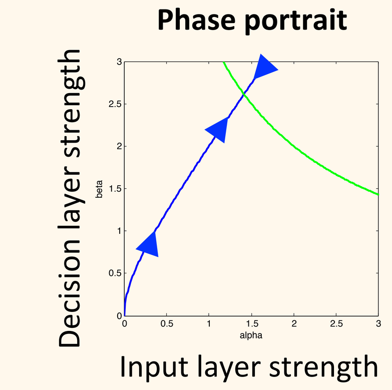
Coincides with prior shallow models (Jazayeri & Movshon, 2006)

Across layers: least informative layer changes most

Temporal evolution



Phase portrait



- Higher layers change **more**:

$\beta > \sqrt{2}\alpha \quad$ for all t

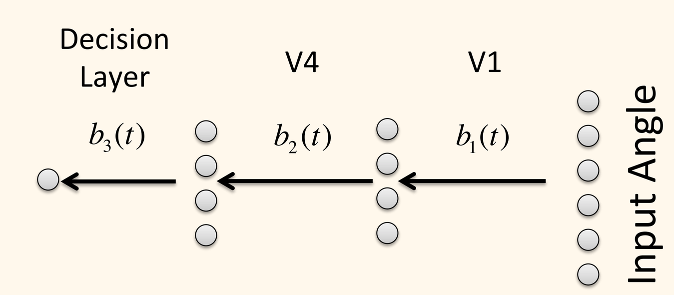
- Higher layers change **first**:

$\beta/\alpha \sim O(\sqrt{2}/\sqrt{\alpha}) \quad$ for small t

$\beta/\alpha \sim O(\sqrt{2}) \quad$ for large t

- Low precision tasks change only higher layer
- High precision tasks change both layers

Deeper networks: Reverse hierarchy of learning (Ahissar & Hochstein, 1997)



Poorly tuned higher layers:

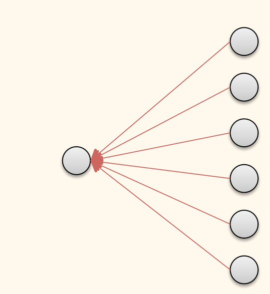
$$b_{N_l}(0) < \cdots < b_2(0) < b_1(0)$$

Implies:

$$\Delta_{N_l}(t) > \cdots > \Delta_2(t) > \Delta_1(t), \quad \forall t > 0$$

Two fundamental topologies

Parallel/Shallow

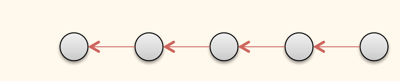


$$y = \sum w_j x_j$$

Weights are **independent**

Most informative neuron changes most

Serial/Deep



$$y = \left(\prod w_k\right) x$$

Weights are **coupled**

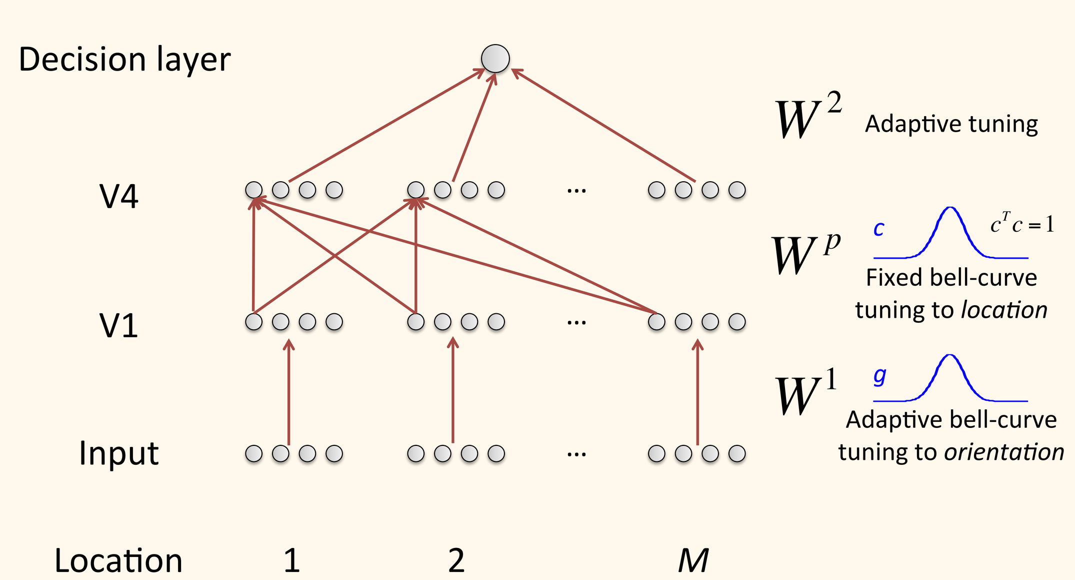
Least informative layer changes most

Transfer to untrained stimuli

Pooling-based invariance

Transfer of learning to novel contexts is a key focus of empirical work

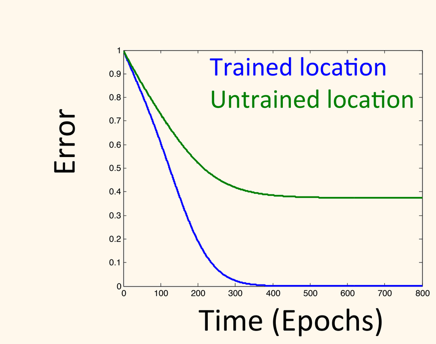
To address these findings, we consider spatial pooling



Dynamics of transfer after restricted-position training

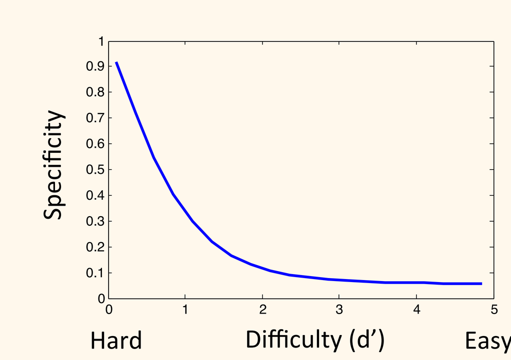
Restricted training at one location has identical reduced dynamics

Transfer to untrained location



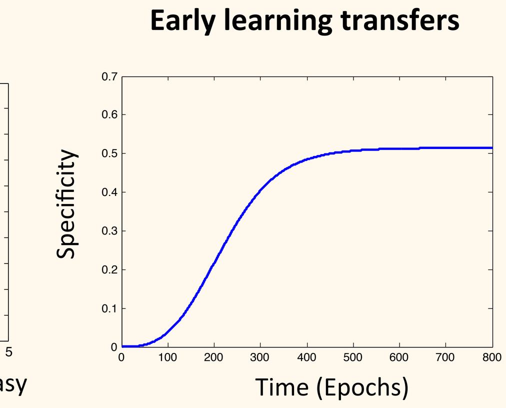
- Decision layer strength transfers, reduced by a constant factor $c_a^T c_b < 1$
- Input layer strength does not transfer

Low precision tasks transfer



- Ahissar & Hochstein, 1997
- Jeter et al., 2009

Early learning transfers



- Jeter et al., 2010
- Hung & Seitz, 2014

Conclusions

Depth may be a key factor determining the size and timing of changes across cortical layers

The pathologies of a deep system qualitatively change learning dynamics

First analytic model to address size and timing of changes across multiple layers

Effect of depth:

- Parallel structure: Most informative neuron changes most
- Serial structure: Least informative layer changes most

Empirical tests: Consistent with a variety of experimental findings

- Reverse hierarchy of learning
- Impact of task precision
- Transfer to novel contexts

Synthesis: Our model synthesizes and formalizes results from a number of previous theories: within-layer changes target most informative neurons (Schoups et al., 2001; Raiguel et al., 2006; Jazayeri & Movshon, 2006); changes follow reverse hierarchy (Ahissar & Hochstein, 1997); larger changes in decision readout (Dosher & Lu, 1998); coarse tasks transfer better than precision tasks (Ahissar & Hochstein, 1997).

Support: A.M.S. is supported by an MBC Traineeship.