Précis of *Deep linear neural networks: A theory of learning in the brain and mind*

Andrew M. Saxe

Learning is a little-remarked miracle: all we–and indeed all many biological organisms–need do to improve at a given task is simply to continue to do it. The sort of task at which we can improve ranges from the exceptionally low level, such as visual Vernier discrimination in which subjects detect slight visual shifts smaller than the diameter of a single photoreceptor [9, 26], to the high level, such as semantic development, in which we progress from little knowledge of items and their properties as infants to a richly structured knowledge of entities in the world and their interrelations as adults [25, 4, 23, 30, 21]. Ultimately these behavioral improvements must be traceable to some change in the neuronal networks of the brain. Crossing these scales–linking behavior to its neural basis–is a critical challenge facing theories of brain function. What changes at the neural level enable these behavioral improvements at the psychological level? And what principles govern the dynamics of the learning process?

This thesis investigates the hypothesis that depth–the brain's chain-like, layered structure–is a critical factor shaping learning dynamics in the brain and mind. Depth refers to layered network structure (Fig. 1A) in which an intermediate layer cannot communicate directly with the input or output, but instead communicates only with the layers adjacent to it. This lack of direct access to inputs and outputs is simultaneously the strength and weakness of a deep architecture. To the good, depth permits a 'divide and conquer' strategy for complex tasks. Rather than solve an entire task in one go, each layer can solve a manageable subproblem, and contribute to the gradual implementation of a complex transformation. This is perhaps the key intuition behind deep learning, that complexity can be built up by the gradual composition of simple features. Yet depth incurs a cost: learning in a deep network is more complicated than in a shallow network. The same flexibility which makes deep networks highly expressive must be constrained during the learning process, so that ultimately one particular network configuration is chosen from amongst the vast array of options. As a result, learning in deep networks is often slow, and progress highly irregular (Fig. 1B).

Anatomically, the brain is a deep structure. Its many neocortical brain areas are arranged in a hierarchy [10], and within each area, the cortical sheet is further subdivided into several layers [8, 29], such that the overall "depth" of the visual cortex, for example, may be on the order of ten stages. To the extent that, anatomically, the brain is deep, and,

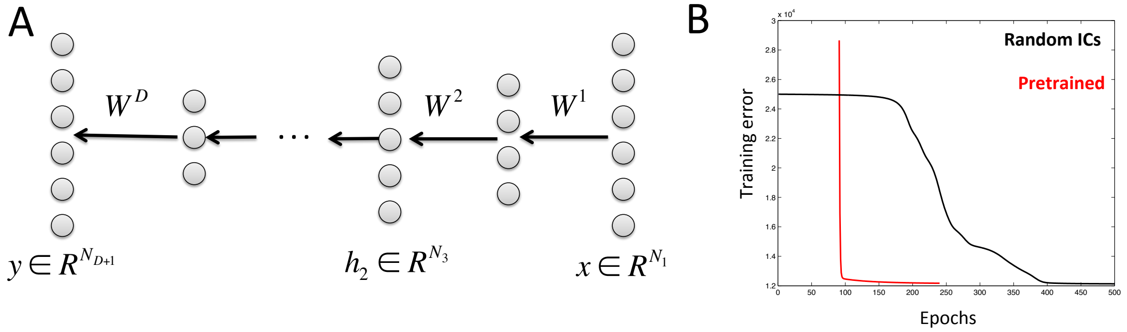

Figure 1: (A) A deep neural network with D layers. Signals may not pass directly from intermediate layers to inputs or outputs. (B) Learning difficulties in deep networks. Squared error of a five-layer deep network trained on a handwritten character recognition task over the course of training from small random weights (black). Initializing a deep network using unsupervised layerwise pretraining (red) speeds learning, even accounting for pretraining time (delay in initial start of red curve).

computationally, deep learning is hard, it is critical to understand how depth might impact learning dynamics in the brain. This thesis proposes that learning in the brain is powerfully sculpted by the special requirements of learning in a deep, chain-like structure. While a number of theoretical proposals have examined how plastic changes might be distributed *within* a single brain area, this work seeks to understand how the brain might apportion changes *across* many layers of the cortical hierarchy.

This thesis develops a quantitative theory of deep learning phenomena by studying learning dynamics in deep linear neural networks. Chapter 2 derives exact solutions to the entire time course of gradient descent learning in deep linear neural networks as a function of depth and the statistics of the task to be learned. These solutions analytically describe the trajectory of every weight in the network during learning. Depth fundamentally transforms the dynamics of learning even in these simple networks, yielding rich phenomena including long plateaus with little learning followed by stage-like transitions. In contrast to their shallow counterparts, deep networks show a sensitive dependence of learning dynamics on the initial weight configurations, and they exhibit behaviors such as illusory correlations that go beyond simple associationist accounts of learning. The theory provides quantitative answers to a variety of fundamental questions, such as why deep learning is often slow; and why a strategy called unsupervised layerwise pretraining can make it fast (Fig. 1B). As they possess little else other than their layered structure, deep linear networks isolate the specific impact of depth on learning dynamics, making them a vital tool for precisely specifying the ramifications of depth alone.

After laying out the theory in Chapter 2, the thesis develops several applications to highlight the relevance of deep learning dynamics to a variety of neural and psychological/cognitive phenomena. In Chapter 3, I examine implications for the epoch of critical period plasticity during which neural receptive fields change in response to their input statistics early in an organism's life. I suggest that the difficulty of learning in a deep system may have led the brain to adopt a particularly successful strategy for mitigating the difficulties of depth, namely unsupervised layerwise pretraining. I further test the idea of domain general unsupervised learning, based on an extensive comparison of five unsupervised learning methods to electrophysiology data from nine different empirical studies across three sensory modalities. In Chapter 4, I show that deep linear networks predict the shape and distribution of changes in neural tuning across different layers of the cortical visual hierarchy in simple perceptual learning experiments. And finally, in Chapter 5, I examine learning about structured domains, and show that the theory captures aspects of the development of human semantic knowledge of items and their properties. These applications span levels of analysis from single neurons to cognitive psychology, demonstrating the potential of deep linear networks to connect detailed changes in neuronal networks to changes in high-level behavior and cognition.

Chapter 1: Introduction

Chapter 1 provides an integrative review of prior work on deep learning in engineering, neuroscience, and psychology. It begins with a historical review of efforts to train useful artificial deep neural networks within the computational sciences, where they have recently achieved state-of-the-art performance on a variety of pattern recognition problems [31, 24, 32]. From the computational literature, it distills six computational hypotheses of relevance to the brain sciences. These hypotheses express the computational rationale behind deep learning methods, and come more or less as a package: the main benefit of a deep architecture, namely the compact expression of complex functions (hypothesis H1), cannot be had without facing a correspondingly more difficult learning problem (H2). To overcome this difficulty, one can employ the unsupervised layerwise pretraining strategy; that this speeds learning is a key hypothesis (H3). Unsupervised layerwise pretraining is also held to improve generalization from limited data (H4). When it comes time to learn the eventual task of interest, this task-driven learning is thought to follow the direction of the gradient of task performance (H5). Overall, none of these hypotheses (H1-H5) relies on specifics of any one input modality, be it vision or audition or somatosensation, and hence the approach is meant to be a domain general way of learning good feature hierarchies and achieving high performance (H6).

The chapter then makes more precise arguments for the brain's depth. Cortical areas are arrayed hierarchically [10], and in addition to deep structure in the interconnection of brain areas, there is a layered anatomy within each brain area [8]. The cortical sheet may be anatomically divided into at least six layers, with inputs arriving in Layer 4 before being passed to Layers 2/3 and then Layers 5/6. And a sequential propagation of signals is also suggested by functional data, as neural responses show a progression of complexity from lower [17, 5, 6, 14] to higher levels [27, 18, 13]. Similar progressions of complexity occur in other sensory modalities [37, 29, 19], though they are less understood.

These two observations–that the brain is deep and deep learning is difficult–motivate the question of how depth impacts learning dynamics in the brain. I propose high-level mappings between the six computational hypotheses and neurocognitive phenomena, associating critical period plasticity with unsupervised layerwise pretraining; and task-driven learning with gradient descent fine-tuning. To make more detailed predictions for the consequences of depth in specific experimental paradigms requires a clear, quantitative theory of the impact of depth on learning dynamics. The chapter ends by introducing the main theoretical tool of the thesis, the deep linear neural network.

Chapter 2: Deep linear network learning dynamics

Chapter 2 derives the major features of the theory, including exact solutions to the full trajectory of learning in deep linear neural networks.[1] The central assumptions of the theory are that learning occurs within sequential, layered structure, represented by a deep linear neural network; and that connection strengths are adjusted according to error-driven gradient descent learning. From these two postulates, a variety of rich dynamical behavior emerges.

A deep linear neural network is a feed forward neural network model with no neural nonlinearity ($f(z) = z$). In particular, given a vector input $x \in R^{N_1}$, a D layer linear network computes the output

$$\hat{y} = W^{D-1}W^{D-2}\cdots W^2W^1x, \tag{1}$$

where the weight matrices $W^{D-1}, \cdots, W^1$ encode the strength of synaptic connections between neurons.

While historically depth and nonlinearity have gone hand in hand in neural network models, deep linear networks tease these two factors apart. Because the composition of linear functions is linear, the input-output map of the deep linear network can always be rewritten as a single shallow network with weight matrix W^{tot},

$$\hat{y} = W^{D-1}W^{D-2}\cdots W^2W^1x = W^{tot}x. \tag{2}$$

Hence deep linear networks fully control for representational power, as regardless of their depth, they only represent a linear function of the input. In this sense, they isolate the

[1] The work described in this chapter is published as A.M. Saxe, J.L. McClelland, and S. Ganguli. (2014). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Y. Bengio & Y. LeCun (Eds.), *International Conference on Learning Representations.* Banff, Canada.

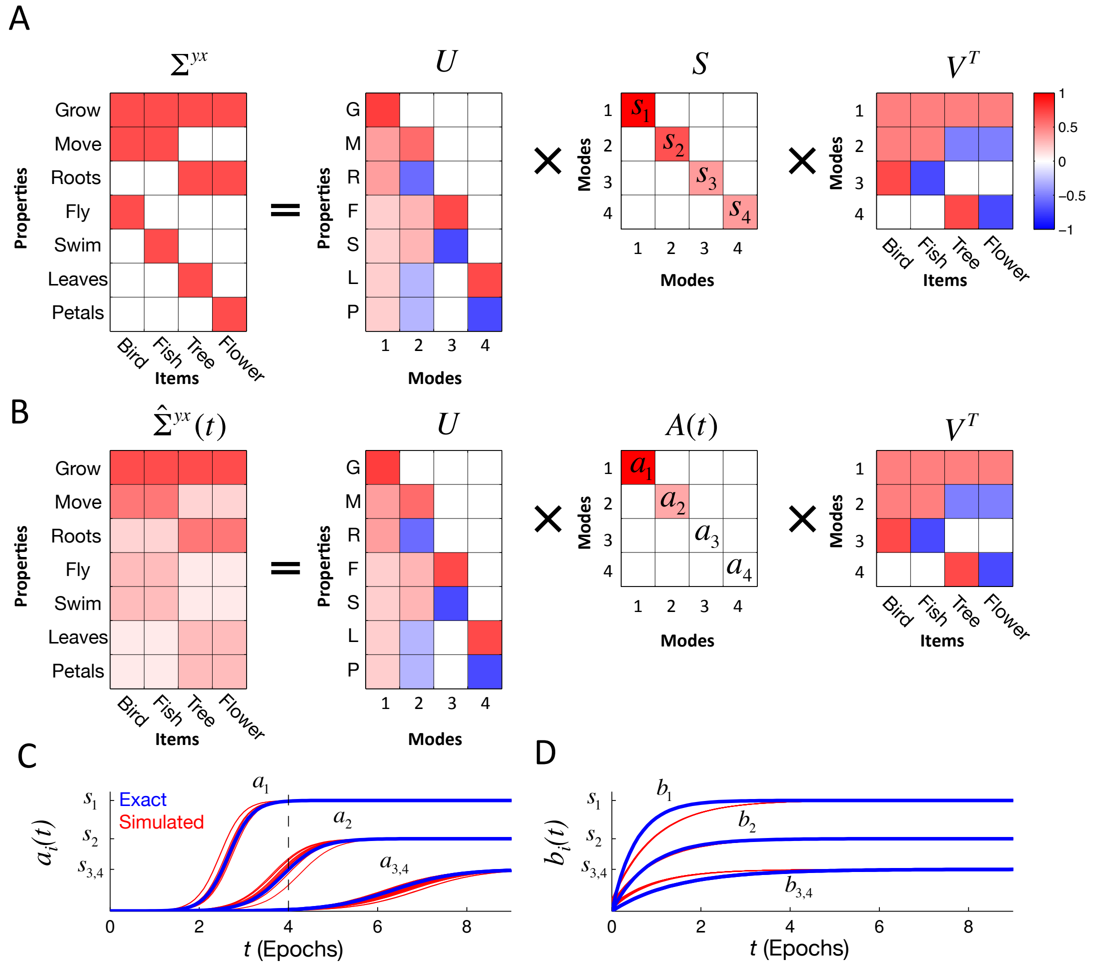

Figure 2: Exact solutions to the dynamics of learning, analyzed via the SVD. (A) Singular value decomposition (SVD) of input-output correlations for an example dataset with hierarchical structure. Associations between items and their properties are decomposed into modes. Each mode links a set of coherently covarying properties (a column of U) with a set of coherently covarying items (a row of V^T). The strength of the mode's covariation is encoded by the singular value of the mode (diagonal element of S). (B) Network input-output map. The effective singular values (diagonal elements of $A(t)$) evolve over time during learning. (C) Time-varying trajectories of the deep network's effective singular values $a_i(t)$. Black dashed line marks the point in time depicted in panel B. (D) Time-varying trajectories of a shallow network's effective singular values $b_i(t)$.

contribution of deep layered structure on learning dynamics. Real neuronal networks are of course nonlinear; yet understanding the simpler linear case is an important prerequisite to developing a theory for nonlinear networks. And as we shall see, deep linear networks in fact exhibit a variety of complex nonlinear learning behaviors which are also observed in nonlinear networks.

Despite their simple input-output map, the learning dynamics are nonlinear: the error between the network's output and the target pattern provided by the environment is the squared error across all inputs, and learning consists in performing the following minimization,

$$\min_{W^{D-1},\cdots,W^1} \sum_{\mu=1}^{P} ||y^\mu - \hat{y}^\mu||_2^2, \tag{3}$$

which for deep networks ($D > 2$) is both nonlinear and nonconvex. Hence deep linear networks have a linear input-output map, but a nonlinear error function, resulting in a nonconvex learning problem.

The gradient descent dynamics, obtained by differentiating (3), are nonlinear and coupled, and depend on the second order statistics of the training dataset, encoded by the input correlation matrix, $\Sigma^{xx} \equiv E[xx^T]$, and input-output correlation matrix $\Sigma^{yx} \equiv E[yx^T]$ as well as the network's depth.

Chapter 2 solves these dynamics for a class of decoupled, balanced initial conditions which provide good approximation to learning dynamics from small random weights. As shown in Fig. 2, the network learns a time-dependent SVD of the input-output correlations. In a deep network, each mode of the SVD is learned according to a sigmoidal trajectory (Fig. 2C), whereas shallow networks exhibit simple exponential approach (Fig. 2D).

These solutions, which describe the trajectory of every weight in the network over time, allow the calculation of other quantities of interest. A key result is an expression relating depth, the initial configuration of the network, and the statistical structure of the training environment to learning speed. The learning dynamics of deep networks are exquisitely sensitive to the initial configuration of weights. Starting from small random weights, deep networks take exponentially more iterations to train than a shallow network, clearly revealing the potential slowdown in learning due to depth (H2). Yet when a deep network is initialized using the unsupervised layerwise pretraining strategy, a deep network is only a finite number of iterations slower than a shallow network. Hence the theory reveals how the unsupervised pretraining strategy speeds learning (H3) in a domain general way (H6).

Chapter 2 constitutes the theoretical core of the dissertation, providing a simple, tractable account of the main ramifications of depth. The results show that depth itself–apart from neural nonlinearities–exerts a strong impact on the learning process, transforming simple learning trajectories into irregular, nonlinear progress, and making learning heavily dependent on the initial knowledge embedded in the network. Chapter 2 provides general solutions that link network depth, training environment statistics, and weight initializations to learning dynamics. The remainder of the thesis specializes these solutions to specific experimental paradigms, by placing further assumptions on the training environment. In this way, deep linear neural networks may provide a new analytical framework with which to investigate a variety of other neural network modeling efforts.

Chapter 3: Domain general unsupervised learning

Chapter 3 focuses on the phenomenon of critical period plasticity in primary sensory cortices, a brief epoch early in an organism's life when neural receptive fields change in response to the statistics of their inputs [3, 15, 22].[2] As remarked earlier, deep networks can be difficult to train unless they are suitably initialized using an unsupervised layerwise pretraining scheme. This chapter suggests that critical period plasticity fulfills this purpose, establishing suitable initial synaptic weights such that subsequent learning is fast. This provides a new computational rationale for critical period plasticity, and gives a clear theoretical justification for how it accelerates learning in the brain's deep layered structure.

Furthermore, the chapter highlights that unsupervised learning can be a domain general strategy for learning useful representations. The core of the chapter is an extensive comparison of five unsupervised learning methods to electrophysiology data from nine different empirical studies across three sensory modalities. In particular, we fed naturalistic visual, auditory, or somatosenory inputs to several unsupervised learning algorithms to obtain predicted artificial neural receptive fields. Notably, we used the exact same algorithm regardless of the modality of the sensory inputs. We then extracted several summary statistics from these artificial receptive fields, and compared them to the distribution of these summary statistics derived experimentally. We find that a wide range of unsupervised learning methods provide indistinguishable quantitative fits to the experimental data, consistent with the hypothesis that a single functional plasticity principle may operate across multiple sensory cortices.

Chapter 4: A deep learning theory of perceptual learning

Chapter 4 investigates the dynamics of simple visual perceptual learning in deep networks, based on the hypothesis that tuning changes follow the gradient of task performance.[3] With

[2] This chapter is based on work published as A. Saxe, M. Bhand, R. Mudur, B. Suresh, & A.Y. Ng. (2011). Unsupervised learning models of primary cortical receptive fields and receptive field plasticity. In *Advances in Neural Information Processing Systems 25*.

[3] Elements of this chapter have appeared in poster format as R. Lee, A.M. Saxe, & J. McClelland (2014). "Modeling Perceptual Learning with Deep Networks." Poster at the 36th annual meeting of the Cognitive Science Society. Quebec City; R. Lee & A.M. Saxe (2015). "The Effect of Pooling in a Deep Learning Model of Perceptual Learning." Poster at the Computational and Systems Neuroscience Conference. Salt Lake City; and A.M. Saxe (2015). "A deep learning theory of perceptual learning dynamics." Poster at the Computational and Systems Neuroscience Conference. Salt Lake City.

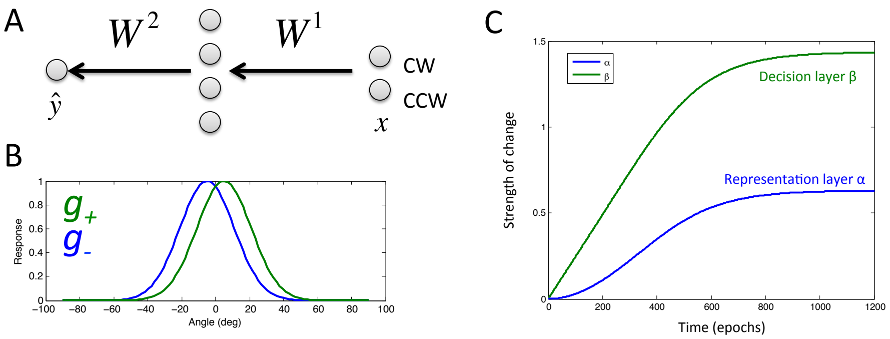

Figure 3: Deep perceptual learning model. (A) Three layer neural network to be trained on an orientation discrimination task. (B) First layer weights are initially orientation-tuned (pictured), while the second layer weights start untuned ($W^2(0) = 0$). (C) Size of weight changes in the representation and decision layers during gradient descent learning. The decision layer changes earlier and more than the representation layer, revealing a reverse hierarchy of learning.

practice, humans and other organisms can dramatically improve their accuracy in simple perceptual discriminations [35, 34, 12, 2]. In the case of visual orientation discrimination [1, 9, 20, 26], experiments have reported learning-induced changes in neural tuning at many levels of the cortical visual hierarchy [11, 28, 33, 36], and the magnitude of the changes within an area has been found to depend strongly on its level within this hierarchy. Generally, larger changes have been found in higher layers, a striking finding given that low layers like V1 exhibit the sharpest orientation tuning [33, 36, 7]. A fundamental challenge for theory is to understand the distribution of neural tuning changes across brain areas that underly the behavioral improvements seen in perceptual learning.

I advance the view that all levels of the cortical hierarchy change during perceptual learning to improve task performance, but that they do so within two critical constraints: first, the cortical levels anatomically lie in a deep, serial hierarchy; and second, this hierarchy is initially configured with the most orientation selective neurons in its lower levels.

I develop a quantitative theory of perceptual learning in this layered structure by considering the simplest incarnation of the deep learning problem: a three layer neural network. This is the minimal model that permits studying the relative distribution of changes in a V1-like representation layer as opposed to readout neurons. At the start of the learning process, the representation layer contains neurons already sensitive to orientation; the decision layer, by contrast, is untuned. Learning occurs by repeatedly adjusting all synaptic weights via gradient descent when the network makes errors.

Using methods similar to those in Chapter 2, I develop an exact reduction of the gradient descent learning dynamics to two variables, one encoding the size of synaptic changes in the V1-like "representation" layer, and one encoding the size of the changes in a higher-level "decision" layer. As shown in Fig. 3, the decision layer weights change more, and earlier, than the V1-layer weights. I prove that this is in fact a more general feature of gradient learning dynamics in deep structure: weaker layers change more during learning than stronger layers. In the case of perceptual learning, in which lower layers are more orientation tuned than higher layers, this results in a reverse hierarchy of learning [2, 1, 16]. The results thus 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.'

In the domain of visual perceptual learning, 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 versus low precision tasks; and the transfer of performance to untrained locations. Further, it consolidates insights from a variety of previous theoretical models, providing a unified quantitative account of the basic features of perceptual learning dynamics and task transfer.

Chapter 5: A theory of semantic development

Finally, Chapter 5 studies the acquisition of human semantic knowledge, knowledge that is typically richly structured.[4] Our knowledge about the natural kinds, for example, sits naturally in a hierarchy, while our knowledge of cities on the globe resides in a spatial structure. A wide array of psychology experiments have revealed remarkable regularities in the developmental time course of human semantic 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 [25, 4, 23, 30, 21]. What are the theoretical principles underlying the ability of neuronal networks to discover abstract structure from experience?

To address this question, the chapter considers training a deep neural network on data drawn from structured probabilistic graphical models of different structural forms. Combining the analytic results of Chapter 2 with the structured domains represented by these probabilistic models yields an integrated theory of the acquisition of abstract structured domains from incremental online experience by a deep neural network (Fig. 4). Depth controls not only the learning dynamics, but also the kinds of generalizations and transient errors made by the network over the course of learning. The learning dynamics in this network exhibit complex developmental phenomena such as stage-like transitions and

[4] A version of this work is published as A.M. Saxe, J.L. McClelland, and S. Ganguli. (2013) Learning hierarchical category structure in deep networks. In M. Knauff, M. Paulen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society.* (pp. 1271-1276). Austin, TX: Cognitive Science Society.

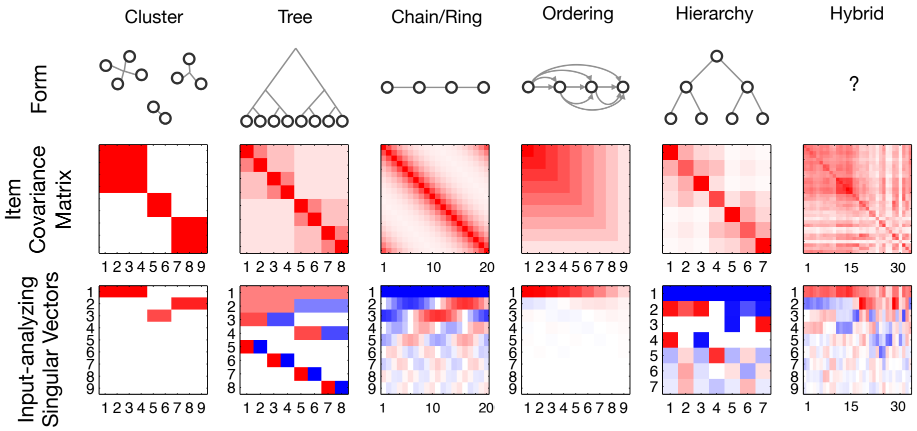

Figure 4: Representation of explicit structural forms in a neural network. Each column shows a different structural form. First row: The structure of the generating probabilistic graphical model. Second row: The resulting item covariance matrix arising from data drawn from the PGM. Third row: The input-analyzing singular vectors that will be successively learned by the linear neural network.

U-shaped learning trajectories, and show how knowledge of explicit abstract structures comes to be encoded in the internal representations of the neural network. The ability to arrange items within an internal representational space allows deep models to exhibit phenomena that go beyond simple associationist accounts, such as illusory correlations during learning. In the case of hierarchically structured domains, these effects combine into a dynamic of progressive differentiation, with waves of progress beginning at the highest levels and eventually reaching the lowest. These results recapitulate analytically several key claims of previous neural network models of these phenomena [30]. Taken together, the model provides the first analytic theory of the acquisition, inductive projection, and underlying neural representation of semantic property knowledge.

Conclusion

This thesis offers a theory of learning in deep layered structure, and argues that the ramifications of depth are evident in many neuroscientific and cognitive phenomena ranging from critical period plasticity to the development of semantic cognition. The theory showcases the direct role of depth itself–as opposed to neural nonlinearities–in qualitatively altering both the learning dynamics of a neural network model and the learned representations.

Overall, deep linear networks occupy a new spot on the continuum between mathematical tractability and expressiveness. They are rare among psychological and neural models in permitting tractable solutions for the entire dynamics of the learning process. From two simple assumptions–deep network structure and gradient descent learning–spring a remarkable array of emergent phenomena, from stage-like transitions in learning, to illusory correlations, U-shaped learning curves, and the progressive differentiation of hierarchical structure. They interlink computation, neural representation, and behavior, providing a framework well-suited to crossing levels of analysis that yields diverse predictions for experiments. The methods developed in this thesis may enable a recapitulation of other neural network models in formal terms.

Neural network models have been criticized for their lack of transparency; while they may make use of sophisticated background knowledge, the exact form this takes can be hard to specify and difficult to understand. By giving an exact account of how explicit structures are represented implicitly in neural networks, our theory provides a clear window into these simple neural network models.

The applications in this thesis span levels of analysis from single neurons to psychological phenomena, showing the potentially broad relevance of depth to learning dynamics in the brain and mind. As an organizing computational principle, deep learning offers an interlocking set of hypotheses that can relate disparate phenomena, from critical period plasticity to semantic development.

References

[1] M. Ahissar and S. Hochstein. Task difficulty and the specificity of perceptual learning. *Nature*, 387(6631):401–6, May 1997.

[2] M. Ahissar and S. Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in cognitive sciences*, 8(10):457–64, October 2004.

[3] N. Berardi, T. Pizzorusso, and L. Maffei. Critical periods during sensory development. *Current Opinion in Neurobiology*, 10(1):138–145, 2000.

[4] S.E. Carey. *Conceptual Change In Childhood.* MIT Press, Cambridge, MA, 1985.

[5] R.L. De Valois, D.G. Albrecht, and L.G. Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision research*, 22(5):545–59, January 1982.

[6] R.L. De Valois, E.W. Yund, and N. Hepler. The orientation and direction selectivity of cells in macaque visual cortex. *Vision research*, 22(5):531–544, 1982.

[7] B.A. Dosher and Z.L. Lu. Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences of the United States of America*, 95(23):13988–93, November 1998.

[8] R.J. Douglas and K.A.C. Martin. Neuronal circuits of the neocortex. *Annual review of neuroscience*, 27:419–51, 2004.

[9] M. Fahle. Specificity of learning curvature, orientation, and vernier discriminations. *Vision research*, 37(14):1885–95, July 1997.

[10] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1(1):1–47, 1991.

[11] G.M. Ghose, T. Yang, and J.H.R. Maunsell. Physiological correlates of perceptual learning in monkey V1 and V2. *Journal of neurophysiology*, 87(4):1867–88, April 2002.

[12] R.L. Goldstone. Perceptual learning. *Annual review of psychology*, 49:585–612, January 1998.

[13] K. Grill-Spector and K.S. Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

[14] M.J. Hawken and A.J. Parker. Spatial properties of neurons in the monkey striate cortex. *Proceedings of the Royal Society of London. Series B*, 231(1263):251–88, July 1987.

[15] T.K. Hensch. Critical period regulation. *Annual review of neuroscience*, 27:549–579, 2004.

[16] S. Hochstein and M. Ahissar. View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*, 36(5):791–804, December 2002.

[17] D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. In *J. Physiology*, 1962.

[18] A.G. Huth, S. Nishimoto, A.T. Vu, and J.L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–24, December 2012.

[19] Y. Iwamura. Hierarchical somatosensory processing. *Curr. Opin. Neurobiol.*, 8(4):522–8, August 1998.

[20] P.E. Jeter, B.A. Dosher, S.H. Liu, and Z.L. Lu. Specificity of perceptual learning increases with increased training. *Vision research*, 50(19):1928–40, September 2010.

[21] C. Kemp and J.B. Tenenbaum. The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10687–92, August 2008.

[22] E.I. Knudsen. Sensitive periods in the development of the brain and behavior. *Journal of cognitive neuroscience*, 16(8):1412–1425, 2004.

[23] J.M. Mandler and L. McDonough. Concept Formation in Infancy. *Cognitive Development*, 8:291–318, 1993.

[24] A. Mohamed, G.E. Dahl, and G. Hinton. Acoustic Modeling Using Deep Belief Networks. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):14–22, January 2012.

[25] G.L. Murphy and D.L. Medin. The role of theories in conceptual coherence. *Psychological review*, 92(3):289–316, 1985.

[26] T. Poggio, M. Fahle, and S. Edelman. Fast perceptual learning in visual hyperacuity. *Science*, 256(5059):1018–21, May 1992.

[27] R.Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(June):1102–1107, 2005.

[28] S. Raiguel, R. Vogels, S.G. Mysore, and G.A. Orban. Learning to see the difference specifically alters the most informative V4 neurons. *The Journal of neuroscience*, 26(24):6589–602, June 2006.

[29] H.L. Read, J.A. Winer, and C.E. Schreiner. Functional architecture of auditory cortex. *Current Opinion in Neurobiology*, 12(4):433–440, 2002.

[30] T.T. Rogers and J.L. McClelland. *Semantic cognition: A parallel distributed processing approach.* MIT Press, Cambridge, MA, 2004.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *arXiv:1409.0575*, 2014.

[32] J. Schmidhuber. Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61:85–117, 2015.

[33] A. Schoups, R. Vogels, N. Qian, and G. Orban. Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412(6846):549–53, August 2001.

[34] A. Seitz and T. Watanabe. A unified model for perceptual learning. *Trends in cognitive sciences*, 9(7):329–34, 2005.

[35] R. Vogels. Mechanisms of Visual Perceptual Learning in Macaque Visual Cortex. *Topics in Cognitive Science*, 2(2):239–250, April 2010.

[36] T. Yang and J.H.R. Maunsell. The Effect of Perceptual Learning on Neuronal Responses in Monkey Visual Area V4. *The Journal of Neuroscience*, 24(7):1617–1626, 2004.

[37] A. Yaron, I. Hershenhoren, and I. Nelken. Sensitivity to Complex Statistical Regularities in Rat Auditory Cortex. *Neuron*, 76(3):603–615, November 2012.