Characterizing emergent representations in a space of candidate learning rules for deep networks

Yinan Cao*
Department of Experimental Psychology
University of Oxford
Oxford, UK
y.cao@uke.de

Christopher Summerfield
Department of Experimental Psychology
University of Oxford
Oxford, UK
christopher.summerfield@psy.ox.ac.uk

Andrew Saxe
Department of Experimental Psychology
University of Oxford
Oxford, UK
CIFAR Azrieli Global Scholar, CIFAR
andrew.saxe@psy.ox.ac.uk

Abstract

How are sensory representations learned via experience? Deep learning offers a theoretical toolkit for studying how neural codes emerge under different learning rules. Studies suggesting that representations in deep networks resemble those in biological brains have mostly relied on one specific learning rule: gradient descent, the workhorse behind modern deep learning. However, it remains unclear how robust these emergent representations in deep networks are to this specific choice of learning algorithm. Here we present a continuous two-dimensional space of candidate learning rules, parameterized by levels of top-down feedback and Hebbian learning. We show that this space contains five important candidate learning algorithms as specific points–Gradient Descent, Contrastive Hebbian, quasi-Predictive Coding, Hebbian & Anti-Hebbian. Next, we exhaustively characterize the properties of each rule during learning about hierarchically structured data, and identify zones within this space where deep networks exhibit qualitative signatures of biological learning. We find that while a large set of algorithms achieve zero training error at convergence, only a subset show hallmarks of human semantic development like progressive differentiation and illusory correlations. Further, only a subset adjust intermediate neural representations toward task-relevant representations, indicative of backpropagation-like behavior. Finally, we show that algorithms can dramatically differ in their learned neural representations and dynamics, providing experimentally testable hallmarks of different learning principles. Our findings provide a framework linking diverse neural representational geometries to learning principles which can guide future experiments, and offer evidence about the learning rules likely to be at work in biology.

1 Introduction

Understanding how neural representations are formed across the cortical hierarchy and how these emergent representations unfold over time to guide behavior are core goals of modern neuroscience.

*This author is now affiliated to University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

Neural representations in many cortical areas appear to be adapted to the performance of specific tasks, such as object recognition [34, 19], face identification [13], or semantic processing [35]. A growing range of work has sought to explain the link between tasks and neural representations within a deep learning framework [37]. Deep neural networks with task-optimized architectures [51] or trained with gradient descent to perform real-world tasks [17] recapitulate neural representations and categorical structure in a range of circumstances [19, 23]. Moreover, they exhibit known empirical features of learning trajectories, such as quasi-stage like transitions between different performance levels [41]. Most of these comparisons rely on supervised learning via gradient descent, corresponding to the highly effective standard practice in deep learning.

However, a long tradition in systems neuroscience has suggested other learning algorithms and principles which might contribute to neural representations. These algorithms could be unsupervised, such as Hebbian learning [29], or sparse coding [46]. Alternatively, they could be supervised but take different forms, such as contrastive Hebbian learning [50] or predictive coding [36, 2, 18]. It therefore remains unclear whether other algorithms might also be able to show the same empirical hallmarks as gradient descent; or conversely, whether even small modifications to gradient descent would radically alter predictions. Moreover, multiple learning algorithms might coexist, with representations emerging through lifelong exposure to a plethora of sensory statistics interacting with a mixture of supervised and unsupervised principles [26, 42].

Therefore a key challenge is to build a linking theory relating data from brains and behavior to a wide range of abstract machine learning algorithms. Here, we develop a systematic approach by introducing a parametric space of learning rules that contains a variety of proposed algorithms as special cases. We show how experimentally measurable aspects of training dynamics and learned representations depend on the choice of algorithm within this space. For the specific empirical domain of semantic development, we find that a region of algorithms centered on gradient descent are broadly consistent with existing experimental data, while algorithms with strong unsupervised Hebbian influences are not. More generally, our approach makes testable predictions about representational geometries and learning trajectories in neural circuits that can help guide future experiments.

2 Methods

What inferences about learning algorithms in biological systems are licensed by different experimental observations? Ideally, one could consider a conceptual space of richly-structured learning rules in which specific algorithms like gradient descent are single points. A given experimental constraint (such as stage-like learning) will be consistent with some subset of this space. As constraints are multiplexed, the set of candidate algorithms would narrow down to a feasible region that accurately represents the remaining uncertainty. Here, we carry out this approach for a simple two-dimensional parametric space of learning algorithms. We first explain the general model architecture, and then justify our construction by showing that it integrates five plausible learning rules.

As shown in Fig. 1a, we consider a simple network architecture with a single hidden layer and linear activation functions. Deep linear networks, while simple, nevertheless yield a nonconvex optimization problem with nonlinear dynamics reminiscent of their nonlinear counterparts [40]. While prior work has focused on the feedforward setting, a number of proposed learning rules make use of top-down feedback. We therefore add top-down feedback, but for simplicity, allow only one backwards sweep. In response to an input, neural activations thus have values at two time points, corresponding to a feedforward sweep that we denote as $\mathbf{h}^{\text{ff}} \in R^{N_h}$, and then one step of recurrent feedback that we denote as $\mathbf{h} \in R^{N_h}$. In response to an input pattern $\mathbf{x} \in R^{N_i}$, the feedforward activity pass yields hidden activity $\mathbf{h}^{\text{ff}}$ and output $\hat{\mathbf{y}} \in R^{N_o}$,

$$\mathbf{h}^{\text{ff}} = \mathbf{W}_1\mathbf{x}, \quad \hat{\mathbf{y}} = \mathbf{W}_2\mathbf{h}^{\text{ff}}, \tag{1}$$

and the feedback activity pass yields updated hidden activity $\mathbf{h}$

$$\mathbf{h} \quad = \quad \mathbf{W}_1\mathbf{x} + \gamma\mathbf{W}_2^T\hat{\mathbf{y}}, \tag{2}$$

where the scalar γ is a key parameter governing the strength of top-down feedback in the network. The matrices $\mathbf{W}_1 \in R^{N_h \times N_i}$ and $\mathbf{W}_2 \in R^{N_o \times N_h}$ denote trainable synaptic connections in the first and second layers respectively.[2]

[2] N_i, N_h, and N_o are the number of input neurons, hidden neurons, and output neurons, respectively.

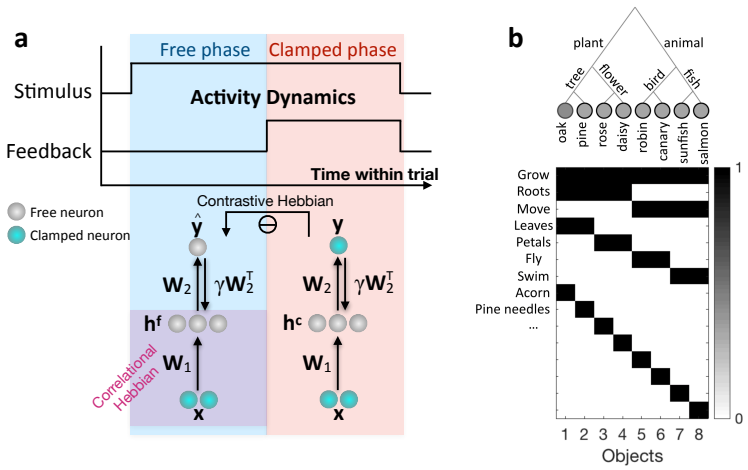

Figure 1: Model and dataset. (**a**) Network model architecture and learning phases. Contrastive Hebbian learning relies on contrasting a clamped state and a free state differing in whether the output layer is clamped at the desired output. (**b**) A hierarchical dataset (desired output matrix $\mathbf{y}$) comprising 8 objects (items).

In order to learn, weights are updated with a learning rule defined by two parameters, γ and η, which govern the nature of a contrastive Hebbian update $\Delta\mathbf{W}_i^{\text{CHL}}(\gamma), i = 1, 2$ and a standard Hebbian update $\Delta\mathbf{W}_1^{\text{HEBB}}(\eta)$, respectively, such that the total update is $\Delta\mathbf{W}_1 = \Delta\mathbf{W}_1^{\text{CHL}}(\gamma) + \Delta\mathbf{W}_1^{\text{HEBB}}(\eta)$ and $\Delta\mathbf{W}_2 = \Delta\mathbf{W}_2^{\text{CHL}}(\gamma)$.

The contrastive Hebbian learning (CHL) term is a standard error-driven learning technique that generalizes gradient descent [50]. It is defined by a Hebbian contrast between two states of the network, one in which the activity of the output neurons are left free, and one in which the output neurons are clamped to the correct target values $\mathbf{y} \in R^{N_o}$. In the 'free' (i.e., input-driven) state, the hidden layer activity after recurrence $\mathbf{h}^f$ is given by $\mathbf{h}^f = \mathbf{W}_1\mathbf{x} + \gamma\mathbf{W}_2^T\hat{\mathbf{y}}$ where the first term is the bottom-up input, and the second term is the top-down feedback. In the 'clamped' (i.e., target-driven) state, hidden layer activity $\mathbf{h}^c$ is given by $\mathbf{h}^c = \mathbf{W}_1\mathbf{x} + \gamma\mathbf{W}_2^T\mathbf{y}$. The weight update is then a contrast of these states (learning rate λ is fixed to a small value of 0.005):

$$\frac{1}{\lambda}\Delta\mathbf{W}_1^{\text{CHL}}(\gamma) = (\mathbf{h}^c\mathbf{x}^T - \mathbf{h}^f\mathbf{x}^T)/\gamma = \mathbf{W}_2^T(\mathbf{y} - \hat{\mathbf{y}})\mathbf{x}^T \tag{3}$$

$$\frac{1}{\lambda}\Delta\mathbf{W}_2^{\text{CHL}}(\gamma) = \mathbf{y}\mathbf{h}^{c^T} - \hat{\mathbf{y}}\mathbf{h}^{f^T} = (\mathbf{y} - \hat{\mathbf{y}})\mathbf{x}^T\mathbf{W}_1^T + \gamma\left(\mathbf{y}\mathbf{y}^T - \hat{\mathbf{y}}\hat{\mathbf{y}}^T\right)\mathbf{W}_2. \tag{4}$$

The Hebbian update term $\Delta\mathbf{W}_1^{\text{HEBB}}(\eta)$ is a standard correlation rule in which co-active neurons potentiate their connections, and is unsupervised. Because simple Hebbian learning updates of the form $\Delta\mathbf{W} = \mathbf{h}\mathbf{x}^T$ are known to be unstable, we include a norm constraint (known as Oja's rule) for positive coefficients η, and normalize the update using the Euclidean norm $\|\cdot\|_2$ of the weights for negative η [44]. Denoting the n^{th} row of $\Delta\mathbf{W}_1^{\text{HEBB}}(\eta)$ as $\Delta\mathbf{w}(\eta)_n^T$, corresponding to the update for weights into the n^{th} hidden unit, we have

$$\Delta\mathbf{w}(\eta)_n^T = \begin{cases} \eta\mathbf{h}_n^{\text{ff}}(\mathbf{x}^T - \mathbf{h}_n^{\text{ff}}\mathbf{w}_n^T) & \text{if } \eta > 0 \\ \eta\mathbf{h}_n^{\text{ff}}\mathbf{x}^T/(1 + \|\mathbf{w}_n\|_2^2) & \text{otherwise} \end{cases} \tag{5}$$

We built this specific two-dimensional continuous space of learning rules, first, so that zero-error solutions ($\hat{\mathbf{y}} = \mathbf{y}$) are stable fixed points of the CHL dynamics (as revealed clearly in Eqs. (3)-(4)); and second, so that by adjusting γ and η it is possible to implement five important and widely considered candidate algorithms. We now describe these connections.

Contrastive Hebbian Learning ($\gamma = 1, \eta = 0$): CHL uses positive top-down feedback with and without clamped targets to update weights, and has been used as a potentially more "biologically plausible" alternative to gradient descent in neuroscience [50, 5].

Gradient Descent ($\gamma \to 0, \eta = 0$): The workhorse algorithm behind deep learning lies at the origin in the 2D space, where additional top-down feedback is infinitesimal and there is no Hebbian contribution. This fact has been shown by [50] and can be seen directly from Eqs. (3)-(4) when $\gamma \to 0$ (i.e., the standard gradient descent equations for deep linear networks [41]).

Quasi-Predictive Coding ($\gamma = -1, \eta = 0$): Predictive coding theories propose that higher processing layers attempt to explain away the activity of lower layers [36, 2, 12, 48, 18]. A version of this idea is instantiated by selecting $\gamma = -1$, such that top-down connectivity is inhibitory. As we show in the Supplementary Material, when trained as an autoencoder, minimum norm weight configurations that achieve zero training error exactly cancel feedforward activity with top-down inhibition. We call our version "quasi"-Predictive Coding because it differs in detail from the canonical model by Rao and Ballard [36] (cf. a diverse range of other specific instantiations in previous work [24]).

Hebbian Learning ($\gamma \to 0, \eta > 0$): Correlational Hebbian learning implements the foundational idea that neurons that fire together wire together [14, 28], justified by many studies probing synaptic plasticity. While the preceding three algorithms are supervised, Hebbian learning is unsupervised and on its own cannot drive error to zero. It extracts the largest principal component of the input data [30]. Here, diversity in neural tuning, and eventual convergence to zero error, is achieved by the CHL term.

Anti-Hebbian Learning ($\gamma \to 0, \eta < 0$): Anti-Hebbian learning, in which co-active neurons tend to unwire, is a hallmark of efficient coding theories of learning. These typically require different neurons to decorrelate such that they represent different aspects of the input. For instance the update in independent component analysis (ICA) is $\Delta \mathbf{W} \propto -\mathbf{h}\mathbf{x}^T + \mathbf{W}^{-T}$ [20] (with traditionally nonlinear $\tanh$ hidden neurons). The first term is anti-Hebbian, while the second term promotes orthogonality of the weight matrix to prevent it from degenerating to zero. In our framework, degeneration to zero is prevented by the CHL term, which requires active neurons to implement the input-output mapping.

Hence five theoretically important learning algorithms exist as points in this two-dimensional space. Alongside them, however, there are an infinite number of candidates which interpolate between these alternatives. Notably, the space allows mixtures of error-driven and unsupervised learning by taking both γ and η to be nonzero.

These learning rules can be applied to a variety of environments and datasets. One important structure of the world considered in prior work is a hierarchy, which organizes semantic knowledge about the natural kinds (e.g., a Ragdoll is a type of cat, which is a type of animal) [38, 41]. Following this line of work, we train networks on an explicitly hierarchical dataset with 15 output properties and 8 input objects (Fig. 1b). The task is to link each object, encoded by a one-hot input vector, to its set of associated binary properties (Fig. 1b; see Supplementary Material for further discussion). To perform semantic tasks, these input objects must be linked to the properties they possess, for instance, whether they can grow, have roots, or can fly. We selected this dataset for two reasons: first, similar datasets have been used to model features of human semantic development [38, 41], providing a rich empirical domain for comparison; and second, because the one-hot inputs contain no statistical structure, any hierarchical similarity emerging in feedforward hidden representations indicates an influence of the target output on lower layers, reflecting an important hallmark of end-to-end training and backpropagation-like behavior. Networks were initialized with small weights drawn randomly from independent Gaussian distributions at both layers (standard deviation = 10^{-10}; see Supplementary Material for simulations for networks initialized with large Gaussian weights), and contained 32 hidden units.

3 Results

We characterize the behavior of algorithms throughout this space of learning rules along three broad dimensions. First, we probe their learning trajectories. While a large region of our 2D space of learning algorithms can drive training error to zero (see Supplementary Material), algorithms can take diverse trajectories. We ask whether there is stage-like learning of hierarchical data, and transient errors on specific properties during learning known as illusory correlations. Second, we investigate internal representations in the hidden layer of the networks over the course of learning. Different algorithms yield distinct neural representations and exhibit specific patterns of one-shot generalization. Third, we determine the degree to which internal representations reflect task-driven or unsupervised influences, separating algorithms which yield end-to-end 'feature learning' from those that do not.

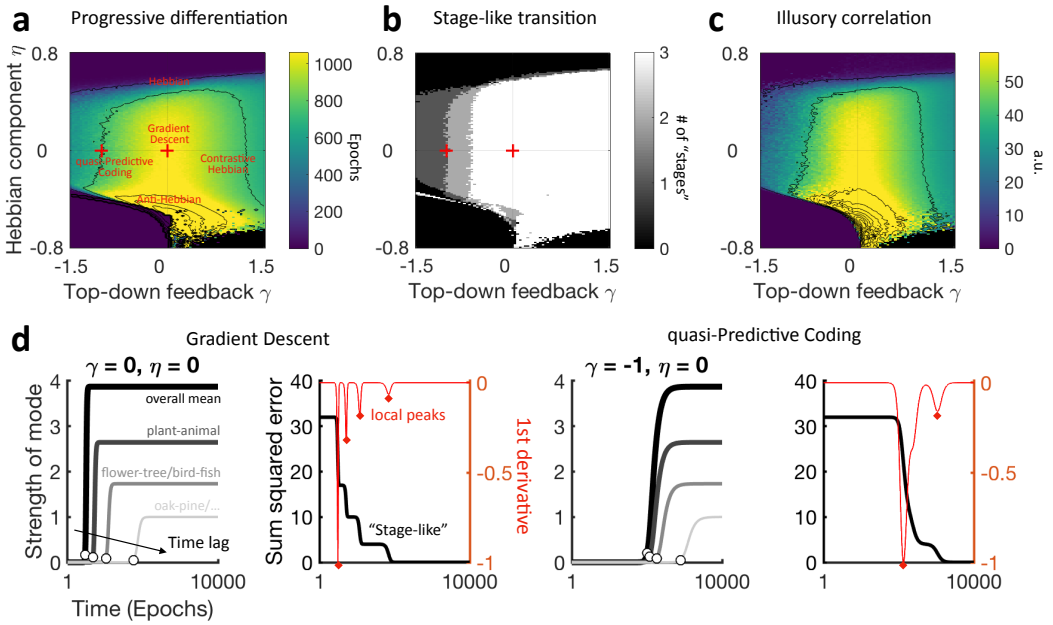

Figure 2: Characterizing learning dynamics under different learning rules throughout the 2D space of algorithms controlled by top-down feedback parameter γ and Hebbian parameter η. (**a**) Mean time lag between learning adjacent hierarchy levels. (**b**) Number of stage-like transitions during learning (quantified as number of local peaks in the time derivative of the error, minus one). (**c**) Illusory correlation strength, measured as the integral of a network's erroneous predictions for a feature/object combination that should have been zero (see Supplementary Material for detailed description). (**d**) Illustrative learning trajectories for two algorithms. Gradient descent (left) exhibits strong progressive differentiation corresponding to each hierarchy level, quantified by large time differences between different effective singular values reaching 5% of their maximal value (open circles), and passes through four distinct stages of error. By contrast, quasi-predictive coding (right) learns all but the finest hierarchical level at a similar time. $\gamma = 0$ represents $\gamma \to 0$, i.e., the limit where top-down feedback is infinitesimal.

3.1 Learning dynamics

During the course of acquiring semantic knowledge, children generally learn broader categorical distinctions, such as between animate and inanimate objects, more rapidly than finer-grained distinctions [25]. Their knowledge can transition through quasi-discrete stages characterized by abrupt reorganizations of concepts [16, 45], and they sometimes attest to false beliefs (e.g., worms have bones) that they should not have learned in real experience [9]. These three signatures of semantic development in biological systems–progressive differentiation, stage-like transitions, and illusory correlations–are known to be emergent properties of deep networks trained with gradient descent on richly structured data [38, 41]. Below, we identify zones within our two-dimensional space of learning rules where deep networks exhibit these signatures of biological learning.

Progressive differentiation of hierarchical statistics To investigate when each hierarchy level in the dataset is learned by different algorithms, we measure the strength of the network's input-output mapping for each hierarchy level. In particular, we use the singular value decomposition of the dataset's input-output correlation matrix to extract the component of the network's input-output mapping in each hierarchical level [41] (see Supplementary Material for more details). A network's learning progress can be described by the trajectory of its effective singular values, one for each hierarchical distinction[3]. These effective singular values start near zero when the network has not encoded a particular distinction, and rise to the true singular value in the dataset when learning converges, as depicted in sample trajectories for two algorithms in Fig. 2d.

[3]Here we assume diagonal one-hot input vectors for simplicity, but see Supplementary Material for further discussion of richer input correlations.

We plot the dependence of learning speed on hierarchy level in Fig. 2a, quantified as the mean time span between learning successive levels. A subset of algorithms centered on gradient descent exhibit prominent progressive differentiation, whereas those with large positive Hebbian or negative top-down feedback components do not.

Stage-like transitions Beyond differences in learning timescale captured by progressive differentiation, we also examine a qualitative change in the shape of the learning trajectory. We measure the number of stage-like drops in error by counting the number of peaks in the time derivative of the total training error. As shown in Fig. 2b, only a subset of algorithms exhibit many stage-like transitions in learning, and these largely overlap with those exhibiting progressive differentiation.

Illusory correlations Finally, we ask whether algorithms exhibit transient errors on specific features during learning. For instance, a child might say that a worm has bones, even though they never could have seen this, because they have overgeneralized from the fact that most animals have bones. These illusions are known to exist under gradient descent training in deep but not shallow networks [41]. We consider an individual feature *m* for item *i*, which should be zero, but has strong positive contributions on the top three hierarchy levels and negative contributions for the third (see detailed description in the Supplementary Material). This can result in a U-shaped trajectory for this feature over the course of learning (see Supplementary Material). We plot the integral w.r.t time of the feature trajectory to quantify the prominence of illusory correlations in Fig. 2c. These semantic illusions emerge in a similar region of algorithms that exhibit progressive differentiation and stage-like transitions.

Hence, the behavioral dynamics observed in a region of this space of algorithms centered on the canonical case of gradient descent recreates experimental results showing that children exhibit progressive differentiation of hierarchical category structure [25], pass through sequences of developmental stages [16], and exhibit semantic illusions during learning [9]. The feasible set of algorithms includes gradient descent, but also extends to CHL and modestly strong quasi-Predictive Coding, Hebbian, and Anti-Hebbian learning. Notably, however, very strong Hebbian learning is inconsistent with these learning dynamics.

3.2 Geometry of emergent neural representations

Because of the redundancy in neural networks with many hidden neurons, any given task can be implemented with a variety of different hidden representations. Here we systematically investigate the neural representations that emerge from different learning algorithms. Comparisons between representations in deep networks and the brain have frequently made use of representational similarity analysis (RSA) [19], which considers two representations to be equivalent when the pairwise distances between neural representations for different inputs are identical. We therefore compute the neural representational similarity matrix (RSM) $\mathbf{\Sigma}^h_{ij} = \mathbf{h}_i^T \mathbf{h}_j$, where $\mathbf{h}_i$ is the hidden representation of item i. This neural similarity pattern can be considered in relation to the similarity between output features in our hierarchical dataset, which has a characteristic ultrametric structure of blocks within blocks (see Supplementary Material). We also visualize time-dependent representational geometry by performing a multi-dimensional scaling (MDS) embedding of the hidden-layer activity for each item at every time point throughout learning. To show systematic changes within the two dimensional space of algorithms, we plot representations along the horizontal and vertical axis in Figs. 3-4, respectively.

Top-down feedback axis Along the horizontal axis as the top-down feedback γ changes from strongly negative to strongly positive, all models learn to perform the task perfectly (Fig. 3a). However the algorithms shift from RSMs emphasizing the representation of superordinate categories (the broadest distinction, animals vs. plants) to emphasizing the representation of subordinate categories (e.g., oak vs. pine) (Fig. 3b). Similarly, the MDS plots in Fig. 3c trace out a generally tree-shaped trajectory over learning, indicating progressive differentiation of the hierarchical structure in the dataset over time; but the distance between superordinate and subordinate distinctions decreases as γ increases.

Hebbian axis Along the vertical axis, most networks (except for strong anti-Hebbian networks) converge to zero training error Fig. 4a. Strong positive values of η yield particularly fast convergence reminiscent of shallow networks [41], but in this regime neural representations no longer reflect the hierarchical task structure (Fig. 4b). By contrast, strong anti-Hebbian learning only differentiates broad semantic categories within the same training time as other algorithms.

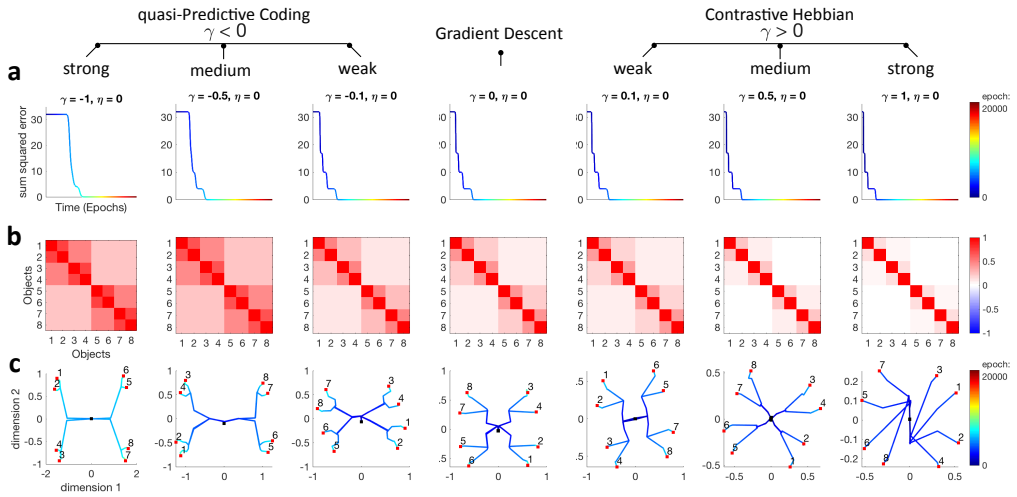

Figure 3: Emergent neural representations with varying top-down feedback γ. (**a**) Training error. (**b**) Neural RSMs at the end of learning. (**c**) MDS embeddings of neural representations for each input over the course of learning. $\gamma = 0$ represents $\gamma \to 0$, i.e., the limit where feedback is infinitesimal.

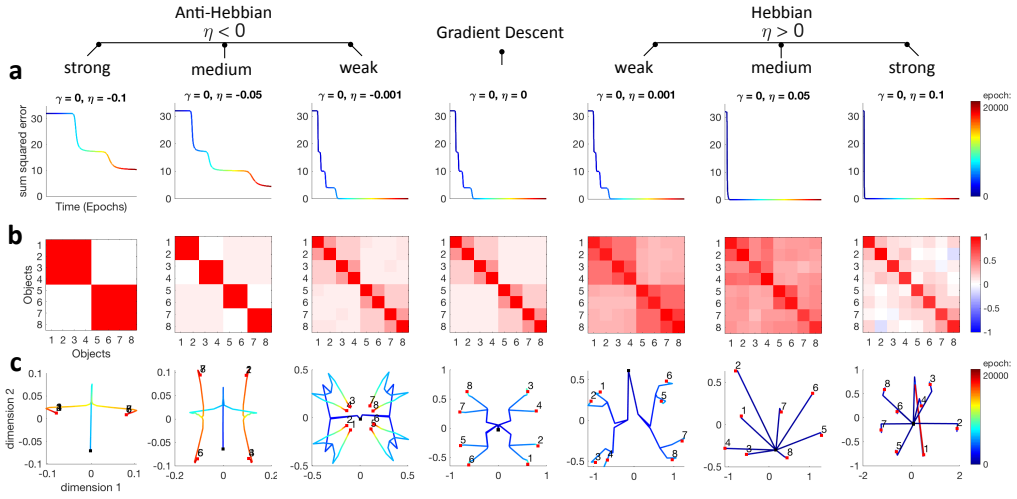

Figure 4: Emergent neural representations with varying Hebbian component η. Same conventions as Fig. 3.

One-shot inductive generalization While many algorithms drive training error to zero, the preceding results show that they do so with different hidden representations. As a result, networks will generalize their knowledge differently on the basis of these representations. Here we evaluate one-shot learning of a new property for a particular item, following the procedure in [39, 27, 41].

Suppose that a novel property m is observed for a familiar object i (e.g. an oak has property m). We instantiate an additional output neuron $\hat{\mathbf{y}}_m$ in the output layer of the neural network, and learn only the weights from the hidden layer to this new output neuron, so as to prevent this fast learning of a single property from interfering with the structural knowledge already stored in the network. Then the network predicts the value of m for the rest of the items. For gradient descent, as learning progresses, the novel feature is first extended broadly to all items before being progressively restricted, and eventually follows the hierarchical structure in the dataset as shown in Fig. 5a. However, this is not the case for algorithms which produce anticorrelated hidden-unit activity patterns during learning (see Supplementary Material). For a network with positive top-down feedback $\gamma = 1$ and a small anti-Hebbian component $\eta = -0.001$), for instance, the network learns anticorrelations between neighboring items (e.g., a pair of 'siblings' like oak and pine). As a result, its one-shot generalizations

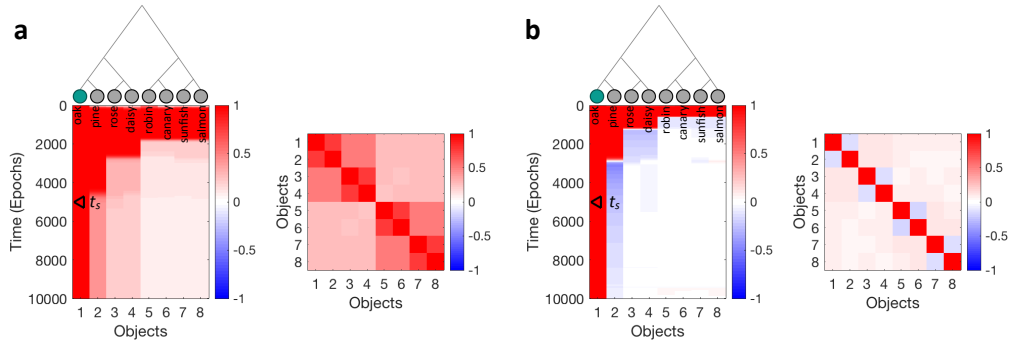

Figure 5: Linking hidden representations to inductive generalization. A novel feature (m; value $y_m = 1$) is observed for oak. Here we ask: "if an oak has property m, do other objects have property m?" (**a**) Under gradient descent, neural representations of objects undergo progressive differentiation, yielding progressively restricted projection of the novel feature to other items, and generalizations which eventually mirror the hierarchy. (**b**) For the algorithm with $\gamma = 1, \eta = -0.001$, the network prioritizes semantic distinctions at the subordinate levels (e.g., oak vs. pine) within a selected training epoch t_s). The RSMs show the output-layer similarity matrix $\Sigma^{\hat{y}}$ after the network rapidly learns 10 novel properties of each of the 8 objects within t_s.

do not conform to the hierarchy in the dataset, even at the end of learning (Fig. 5b). In the domain of semantic development, children are capable of rich generalizations that conform to the structure of a domain, consistent with the behavior of algorithms near gradient descent [9, 16].

3.3 End-to-end learning versus local learning

Deep learning emphasizes task-driven end-to-end representation learning [7], rather than local unsupervised learning. A similar distinction has been the focus of a variety of empirical work in neuroscience. Several theories propose substantial task-driven reorganization of low-level representations during learning [43, 47]. By contrast, other accounts have suggested that learning mainly changes readout weights from a relatively stable internal representation [11, 33]. Here we consider the degree to which first-layer weights are determined by CHL updates, reflecting task-driven representational change, as compared to the Hebbian component, reflecting unsupervised representational change. As shown in Fig. 6, task-driven contributions are substantial in a region of the space, reflecting an end-to-end learning regime that centers on gradient descent and largely overlaps with the rules that exhibit progressive differentiation. Strong Hebbian algorithms, by contrast, yield first-layer weights that are dominated by unsupervised Hebbian updates (see Supplementary Material for further discussion). Fig. 6 shows the degree to which hidden-layer features are learned via error backpropagation vs. unsupervised Hebbian learning, a key distinction in many theories of neural learning. We note that γ and η do scale the update size, but this does not directly translate to the integrated changes. E.g., if error is driven near zero, the CHL component will stop learning even with large γ. The integrated synaptic strength changes in the network model correspond to important training-induced plasticity that can be measured in electrophysiological experiments [1]. Hence our space of rules interpolates between task-driven end-to-end feature learning and local unsupervised theories [6, 10, 4, 15, 21, 3], opening new avenues for testing these theories in future experiments.

4 Conclusion

Biological learning exhibits striking phenomena like progressive differentiation, stage-like transitions, and semantic illusions. Using a 2D space of learning rules, we have shown that these signatures are not limited to gradient descent alone. Instead, a region of learning algorithms exhibit these phenomena, though with meaningful subtleties. In particular, the amount of Hebbian learning is constrained, consistent with prior work [28, 6]. Of note, we do not claim that Hebbian learning never yields progressive differentiation. In other tasks (particularly those where the unsupervised statistics are hierarchical too), progressive differentiation could occur in Hebbian learning. Our goal here was to start with a task environment well-studied in prior work, that specifically does not have the

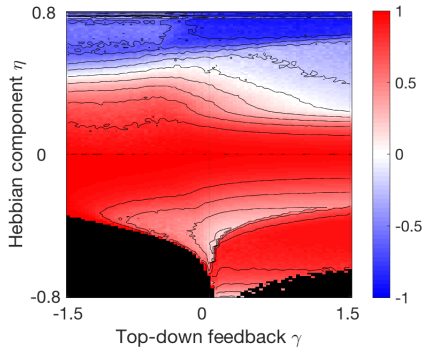

Figure 6: Competing contributions from CHL versus Hebbian term to synaptic weight update in the first layer: $\left(\int_t \left\|\Delta \mathbf{W}_1^{\text{CHL}}\right\|_F dt - \int_t \left\|\Delta \mathbf{W}_1^{\text{HEBB}}\right\|_F dt\right) / \int_t \left\|\Delta \mathbf{W}_1\right\|_F dt$. Red: greater contribution from CHL. Blue: greater contribution from Hebbian. Black: non-convergence in simulation.

target structure embedded in the inputs. Future work could extend our rule space to cover second-order methods [8], nonlinear networks, and the growing number of approximate backpropagation alternatives [22, 49], as well as other ways of combining error-correcting and Hebbian learning (e.g., Leabra; [31, 32]). A major contribution of our work is proposing a minimalist space that nevertheless encompasses five commonly discussed learning rules. The metrics we propose provide a methodology that can be used to characterize any desired learning rule. Our findings offer a roadmap for further constraining the learning mechanisms at work in biology, by comparing the representations that emerge through learning, the generalizations they support, and the distribution of task-driven changes across brain structures.

Broader Impact

We hope the work presented here could be of interest to neuroscientists and cognitive scientists who use deep networks to model and understand how biological brains encode, compose and generalise abstract knowledge. Our work aims to characterize basic learning processes in the brain, and any applications are likely to lie far in the future with substantial uncertainty. If the methods in this paper did lead to identification of learning principles operating in parts of the brain, this could aid design of optimal curricula, with possible benefits for educational and medical settings (for instance, restoration of vision after stroke). Conversely, however, quantitative theories of human learning could allow design of adversarial curricula that rapidly induce false beliefs. Our findings on illusory correlations show that different learning algorithms make errors in which broad correlations in the dataset are over-extended to specific inputs to which they do not apply. These findings could have implications for susceptibility to biases in data. If these learning mechanisms describe human learning, this could provide insight into the development of implicit biases. At present our work is far from these impacts.

Acknowledgments and Disclosure of Funding

We thank Xavier Roberts-Gaal for contribution to code and simulations, and Sarah Armstrong, Hannah Sheahan, Stephanie Nelli, and Adam Harris for helpful comments on early versions of this paper. This work was funded by the European Research Council (award 725937 to C.S.) and the Human Brain Project (Special Grant Agreement 3; to C.S.). A.M.S. is supported by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (Grant Number 216386/Z/19/Z), and the CIFAR Azrieli Global Scholars program.

References

[1] M. Ahissar and S. Hochstein. The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, 8(10):457–464, 2004.

[2] L. Aitchison and M. Lengyel. With or without you: Predictive coding and Bayesian inference in the brain. *Current Opinion in Neurobiology*, 46:219–227, 2017.

[3] S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, pages 477–502, 2019.

[4] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 2020.

[5] P. Baldi and F. Pineda. Contrastive learning and neural oscillations. *Neural Computation*, 3(4):526–545, 1991.

[6] P. Baldi and P. Sadowski. A theory of local learning, the learning channel, and the optimality of backpropagation. *Neural Networks*, 83:51–74, 2016.

[7] Y. Bengio and Y. LeCun. Scaling learning algorithms towards AI. In L. Bottou, O. Chapelle, D. DeCoste, and J. Weston, editors, *Large-Scale Kernel Machines*. MIT Press, 2007.

[8] A. Bernacchia, M. Lengyel, and G. Hennequin. Exact natural gradient in deep linear networks and application to the nonlinear case. In *NeurIPS*, pages 5941–5950, 2018.

[9] S. Carey. Précis of 'the origin of concepts'. *Behavioral and Brain Sciences*, 34(3):113–24, 2011.

[10] L. Chizat, E. Oyallon, and F. Bach. On lazy training in differentiable programming. In *NeurIPS*, 2018.

[11] B. Dosher and Z. 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, 1998.

[12] T. Egner and C. Summerfield. Grounding predictive coding models in empirical neuroscience research. *Behavioral and Brain Sciences*, 36(3):210–211, 2013.

[13] J. Haxby, M. Gobbini, M. Furey, A. Ishai, J. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[14] D. Hebb. *The organization of behavior*. John Wiley & Sons, New York, 1949.

[15] A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, pages 8571–8580, 2018.

[16] F. Keil. *Semantic and conceptual development: An ontological perspective*. Harvard University Press, Cambridge, MA, 1979.

[17] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[18] G. Keller and T. Mrsic-Flogel. Predictive processing: A canonical cortical computation. *Neuron*, 100(2):424–435, 2018.

[19] N. Kriegeskorte, M. Mur, D. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, and P. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008.

[20] D. Krotov and J. J. Hopfield. Unsupervised learning by competing hidden units. *Proceedings of the National Academy of Sciences*, 116(16):7723–7731, 2019.

[21] J. Lee, L. Xiao, S. Schoenholz, Y. Bahri, R. Novak, J. Sohl-Dickstein, and J. Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In *NeurIPS*, 2019.

[22] T. Lillicrap, A. Santoro, L. Marris, C. Akerman, and G. Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(June):335–346, 2020.

[23] G. Lindsay. Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, pages 1–15, 2020.

[24] W. Lotter, G. Kreiman, and D. Cox. A neural network trained to predict future video frames mimics critical properties of biological neuronal responses and perception. *arXiv preprint arXiv:1805.10734*, 2018.

[25] J. Mandler and L. McDonough. Concept formation in infancy. *Cognitive Development*, 8:291–318, 1993.

[26] A. Marblestone, G. Wayne, and K. Kording. Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, 10:1–61, 2016.

[27] J. McClelland. A connectionist perspective on knowledge and development. In T. Simon and G. Halford, editors, *Developing cognitive competence: New approaches to process modeling*. Erlbaum, Hillsdale, NJ, 1995.

[28] J. L. McClelland. How far can you go with hebbian learning, and when does it lead you astray. *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI*, 21:33–69, 2006.
[29] Y. Miyashita. Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, 335(6193):817–820, 1988.
[30] E. Oja. Principal components, minor components, and linear neural networks. *Neural Networks*, 5(6):927–935, 1992.
[31] R. C. O'Reilly. Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2(11):455–462, 1998.
[32] R. C. O'Reilly and Y. Munakata. *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press, 2000.
[33] A. Petrov, B. Dosher, and Z. Lu. The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, 112(4):715–43, 2005.
[34] R. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107, 2005.
[35] M. L. Ralph, E. Jefferies, K. Patterson, and T. Rogers. The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, 18(1):42–55, 2017.
[36] R. P. Rao and D. H. Ballard. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999.
[37] B. Richards, T. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, and K. P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, 2019.
[38] T. Rogers and J. McClelland. *Semantic cognition: A parallel distributed processing approach*. MIT Press, Cambridge, MA, 2004.
[39] D. Rumelhart and P. Todd. Learning and connectionist representations. In D. Meyer and S. Kornblum, editors, *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*. MIT Press, Cambridge, MA, 1993.
[40] A. Saxe, J. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Y. Bengio and Y. LeCun, editors, *International Conference on Learning Representations*, Banff, Canada, 2014. Oral presentation.
[41] A. Saxe, J. McClelland, and S. Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.
[42] A. Saxe, S. Nelli, and C. Summerfield. If deep learning is the answer, then what is the question? *arXiv preprint arXiv:2004.07580*, 2020.
[43] A. Schoups, R. Vogels, N. Qian, and G. Orban. Practising orientation identification improves orientation coding in V1 neurons. *Nature*, 412(6846):549–53, 2001.
[44] N. N. Schraudolph and T. J. Sejnowski. Competitive anti-hebbian learning of invariants. In *Advances in Neural Information Processing Systems*, pages 1017–1024, 1992.
[45] R. Siegler. Three aspects of cognitive development. *Cognitive Psychology*, 8:481–520, 1976.
[46] E. P. Simoncelli and B. A. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001.
[47] L. Wenliang and A. Seitz. Deep neural networks for modeling visual perceptual learning. *The Journal of Neuroscience*, 38(27):6028–6044, 2018.
[48] J. Whittington and R. Bogacz. An approximation of the error backpropagation algorithm in a predictive coding network with local Hebbian synaptic plasticity. *Neural Computation*, 29:1229–1262, 2017.
[49] J. Whittington and R. Bogacz. Theories of error back-propagation in the brain. *Trends in Cognitive Sciences*, 23(3):235–250, 2019.
[50] X. Xie and H. Seung. Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation*, 15(2):441–454, 2003.
[51] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.