Minimum perturbation theory of deep perceptual learning

Haozhe Shan and Haim Sompolinsky

Phys. Rev. E **106**, 064406 — Published 12 December 2022

DOI: 10.1103/PhysRevE.106.064406

A Minimum Perturbation Theory of Deep Perceptual Learning

Haozhe Shan*

Center for Brain Science, Harvard University, Cambridge, MA, United States and Program in Neuroscience, Harvard Medical School, Boston, MA, United States

Haim Sompolinsky

Center for Brain Science, Harvard University, Cambridge, MA, United States Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem, Israel and Racah Institute of Physics, Hebrew University of Jerusalem, Jerusalem, Israel

(Dated: November 28, 2022)

Abstract

Perceptual learning (PL) involves long-lasting improvement in perceptual tasks following extensive training and is accompanied by modified neuronal responses in sensory cortical areas in the brain. Understanding the dynamics of PL and the resultant synaptic changes is important for causally connecting PL to the observed neural plasticity. This is theoretically challenging because learning-related changes are distributed across many stages of the sensory hierarchy. In this work, we modeled the sensory hierarchy as a deep nonlinear neural network and studied PL of fine discrimination, a common and well-studied paradigm of PL. Using tools from statistical physics, we developed a mean-field theory of the network in the limit of large number of neurons and large number of examples. Our theory suggests that, in this "thermodynamic" limit, the input-output function of the network can be exactly mapped to that of a deep *linear* network, allowing us to characterize the space of solutions for the task. Surprisingly, we found that modifying synaptic weights in the first layer of the hierarchy is both sufficient and necessary for PL. To address the degeneracy of the space of solutions, we postulate that PL dynamics are constrained by a normative "minimum perturbation" (MP) principle, which favors weight matrices with minimal changes relative to their pre-learning values. Interestingly, MP plasticity induces changes to weights and neural representations in all layers of the network, except for the readout weight vector. While weight changes in higher layers are not necessary for learning, they help reduce overall perturbation to the network. In addition, such plasticity can be learned simply through slow learning. We further elucidate the properties of MP changes and compare them against experimental findings. Overall, our statistical mechanics theory of PL provides mechanistic and normative understanding of several important empirical findings of PL.

I. INTRODUCTION

Perceptual learning, the improvement of performance in perceptual tasks after practice, is one of the most basic forms of learning in the brain and has been extensively studied experimentally [1–12]. Physiologically, PL is accompanied by long-lasting changes to neuronal response properties in cortical areas. Connecting physiological changes to behavioral obser-

* hshan@g.harvard.edu; http://hzshan.github.io

vations has been challenging, in part due to the complex learning dynamics and processing in the sensory hierarchy, which is composed of multiple cortical regions. As a result, several important issues concerning the neural mechanisms of PL remain unresolved after decades of research.

First, which cortical areas undergo modifications and which of the changes causally drive PL? While behavioral specificity of PL [13, 14] points to an important role for plasticity in early sensory areas, single-unit response properties in early visual areas (V1, V2) show only minor changes after visual PL [3, 4]. In addition, PL induces significant changes to single-neuron properties in intermediate to late stages of visual processing, such as V4 [5, 6, 9, 10], LIP [15], and IT [16, 17]. Furthermore, it is unclear whether any of such changes necessarily causes PL. For example, PL of sound and tactile discrimination is correlated with substantial changes in respective primary sensory areas [1, 12], but such changes may not contribute to improved neural coding [11].

Second, what are the functional consequences of the observed changes? Analysis of changes in neuronal responses after PL indicates improved accuracy of the neural coding of the trained stimuli [8–10]. This appears to be inconsistent with the behavioral finding that PL does not transfer to a different task even when using the same stimuli [18–20]. The Reverse Hierarchy Theory [21] proposes that PL is initially driven by learning in high areas, which results in less specific learning; modifications of lower areas follow if the task is difficult, as for instance in fine perceptual discrimination tasks, leading to more specific learning. Analysis of a reduced model of perceptual learning has lent support for this theory [22]. However, recent experimental and computational studies questioned these predictions [8, 9, 23], providing evidence of changes in primary sensory areas already in the early stages of PL. On the other hand, experiments in random dot visual motion discrimination tasks found that PL is correlated with changes in decision-making areas (e.g., LIP) but not sensory areas (e.g., MT) [15, 24]. From a theoretical perspective, the hierarchical nature of the underlying sensory system implies that there is an enormous degeneracy of possible synaptic weight matrices that solve the task of PL.

Most existing theories of PL assume changes only in either the weights of the readout from a fixed sensory array [25–27] or the input layer to a single cortical circuit [23]. Such "shallow" models are inconsistent with the sensory hierarchy in the brain and do not address the neural correlates of PL in multiple cortical regions.

In the present work, we directly addressed the issue of PL in a deep network by studying PL of a fine-discrimination task in a deep neural network (DNN) model of the sensory hierarchy [28–30]. As learning dynamics in DNNs are in general challenging to study [31–39], we developed a mean-field theory of information propagation in the model in the limit of large numbers of neurons in every layer and large number of training examples. The theory reveals that during the perceptual task, the DNN effectively behaves like a deep *linear* neural network. This considerably simplifies the theoretical analysis of the space of solutions, as well as the emergent changes in neural representations. Surprisingly, we found that modifications of synaptic weights in the first level of the hierarchy are both sufficient and necessary for PL. To address the degeneracy of the space of solutions, we developed a *normative* theory of PL. Specifically, we postulated that in the brain, learning dynamics are constrained by a normative "minimum perturbation (MP)" principle, which favors weight matrices with minimal changes relative to their pre-learning values. Interestingly, MP learning induces changes in weights and neural representations in all layers of the networks, except for the readout weight vector. While weight changes in higher layers are not necessary for learning, they help reduce overall perturbation to the network. MP learning predicts changes to tuning properties of cortical neurons that are consistent with experimental observations and suggests that signal amplification, not noise reduction, is the primary driver of PL. Our theory makes the readily testable prediction that PL can simultaneously lead to positive and negative transfer to different untrained stimuli. Finally, we found that MP learning can be implemented through slow gradient-descent learning. Overall, leveraging the large size of the network involved in PL, we have developed a statistical mechanics theory of PL in deep neural networks which provides mechanistic and normative understanding of several important empirical findings of PL.

Put in a broader context, this work complements recent theoretical studies of learning in deep networks [31–39], contributing to the understanding of learning and computation in these important architectures. In particular, our setting where a deep, nonlinear network learns a linearly solvable task is a popular paradigm for understanding network learning in the so-called overparameterized regime [40, 41], where the network is vastly larger and richer than is required by the trained task [42]. Unlike standard analyses that focus on how the network starts from random initialization and learns a single task, our work introduces a continual-learning perspective where the impact of learning on previously learned tasks

needs to be minimized.

Our deep network model of PL is described in Section II. The mean field analysis is summarized in Section III. Section IV presents the MP principle and analyzes PL with minimum perturbation. Section V analyzes the use of gradient descent to learn MP plasticity. A discussion of the implications for the field of perceptual learning is presented in Section VI.

II. A DEEP NETWORK MODEL OF PL

A. Input channels

We assume N input channels (**Fig. 1A**, gray squares) representing a 1D stimulus. Neurons in the input channels are indexed by a preferred stimulus angle $\theta_i = \frac{i}{N}2\pi$ for the ith neuron. The collective response of input channels to a stimulus with angle θ is given by the N-dim vector

$$\boldsymbol{x}^0(\theta) = \boldsymbol{f}^0(\theta) + \boldsymbol{\epsilon}^0, \tag{1}$$

where $\boldsymbol{\epsilon}^0$ is i.i.d. Gaussian noise with zero mean and variance σ^2. The noise averaged response of each input neuron is given by a bell-shaped tuning curve centered on its preferred stimulus

$$f_i^0(\theta) = Z_s^{-1} \exp\left(\frac{\cos(\theta_i - \theta) - 1}{\sigma_s^2}\right), \tag{2}$$

where Z_s ensures $\|\boldsymbol{f}^0(\theta)\| = \sqrt{N}$, making the firing rate of each neuron $O(1)$. σ_s controls the input selectivity, assumed to be the same for all channels (**Fig. 1B, C**). Tuning and noise properties of input channels are not affected by learning.

B. Model architecture and pre-PL weights

Our model of the sensory system is a feedforward network with L hidden layers and a linear readout from the top layer (**Fig. 1A**). Each hidden layer is composed of N rectified linear (ReLU) neurons ("cortical neurons"). Let $\boldsymbol{x}^l(\theta)$ denote the noisy population response vector of neurons in layer l, and $\boldsymbol{f}^l(\theta)$ its average over noise. $\{\boldsymbol{x}^l(\theta)\}_{l=1,.,...,L}$ are recursively given by

$$\boldsymbol{x}^l(\theta) = \Phi(\boldsymbol{W}^l \boldsymbol{x}^{l-1}(\theta)), \tag{3}$$

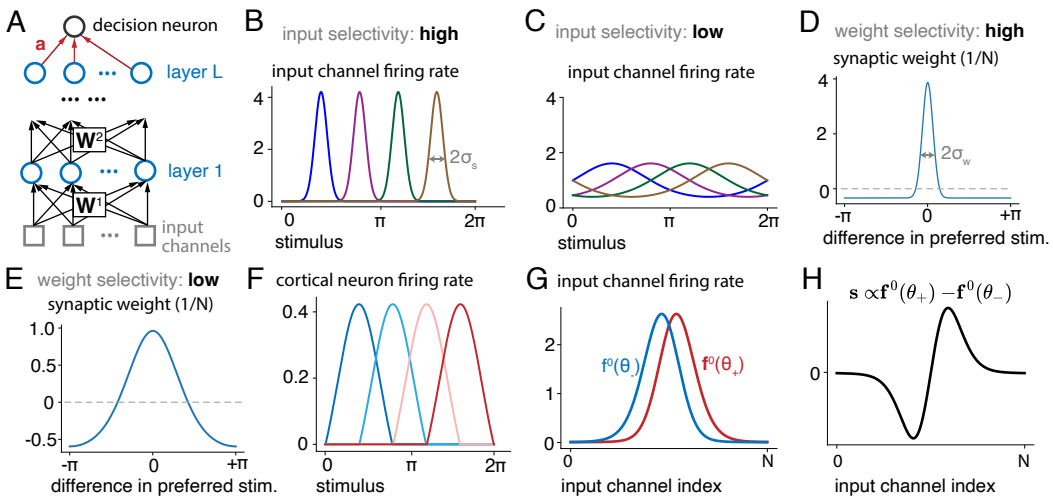

FIG. 1. **Model of perceptual learning.**

A Diagram of the deep network model of the sensory hierarchy. The output of input channels (gray squares) is passed through L layers of cortical neurons with ReLU nonlinearity (blue circles) before getting read out by a linear readout ($\boldsymbol{a}$).

B Example tuning curves of input channels. Each curve represents a channel with a different preferred stimulus. The preferred stimuli of input channels uniformly tile $[0, 2\pi]$. This panel shows the regime of high input selectivity and hence narrow tuning curves.

C Same as B, but for the scenario of low input selectivity.

D Example feedforward weight structure before learning. Weights connecting neurons with similar preferred stimuli tend to be excitatory (positive) and strong while those connecting neurons with dissimilar preferred stimuli tend to be weak and inhibitory (negative). This panel shows the regime of high weight selectivity.

E Same as D, but for low weight selectivity.

F Bell-shaped tuning curves of input channels and the initial weight patterns lead to bell-shaped tuning curves for all cortical neurons before PL.

G Noise-averaged activity of the input channels ($\boldsymbol{f^0}(\theta)$) in response to the two presented stimuli, $\theta_\pm$. The difference between them is exaggerated here for illustration purposes.

H The signal $\boldsymbol{s}$ is in the direction of the difference between $\boldsymbol{f^0}(\theta_\pm)$.

where $\Phi(\cdot)$ is the element-wise rectified linear function. The linear behavioral readout $\boldsymbol{a}$ produces a scalar network output from the activity in the last layer

$$r(\theta) = \boldsymbol{a}^T \boldsymbol{x}^L(\theta). \tag{4}$$

Pre-PL weights are modelled after feedforward synaptic connections between visual areas in the brain: we chose a circulant structure which is appropriate for propagating angular signals [43, 44]. Every neuron receives strong, excitatory input from neurons in the previous layer with similar preferred stimuli and weak, inhibitory input from neurons with dissimilar preferred stimuli. For simplicity, we assume pre-PL weights to be identical across layers. We do not expect qualitative predictions of our analysis to differ if initial weights have different σ_w across layers. Concretely, pre-PL weights $\{\boldsymbol{W}^l\}_{l=1,2,\ldots,L}$ are given by

$$W^l_{ij,\mathrm{pre}} = Z_w^{-1} \exp\left(\frac{\cos(\theta_i - \theta_j) - 1}{\sigma_w^2}\right) + b_w, \tag{5}$$

where Z_w is chosen such that each row of $\boldsymbol{W}^l_{\mathrm{pre}}$ has norm $1/\sqrt{N}$ (i.e. each weight is $O(N^{-1})$). This normalization ensures that the input to any hidden neuron is of magnitude $O(1)$. The offset b_w is chosen such that each row sums to 0. The parameter σ_w controls selectivity of the pre-learning weights; small σ_w leads to a high-selectivity weight structure where a few weights dominate the input (**Fig. 1D**) and vice versa (**Fig. 1E**). As a result of the input tuning curves and the feedforward weight structure, all cortical neurons are tuned to the 1D stimulus and have bell-shaped tuning curves before learning (**Fig. 1F**).

C. Fine discrimination

We focus on learning a fine-discrimination task, where one out of two similar visual stimuli is presented to the subject, who must correctly indicate which one is presented. In our model, the task consists of discriminating two values of the stimulus, $\theta_\pm = \theta_{\mathrm{tr}} \pm \delta\theta$, where the center stimulus θ_{tr} is called the trained stimulus and $\delta\theta \sim O(N^{-1/2})$. This choice of scaling ensures that the total signal-to-noise ratio (SNR) in the input layer is $O(1)$. In each trial, one of $\theta_\pm$ is presented and generates a noisy activation of the input array (Eq. 1, **Fig. 1G**). In each trial, the decision neuron activity r indicates whether the input comes from the θ_+ stimulus or from θ_- with $r > 0$ or $r < 0$, respectively. Stimuli are presented with equal probability; the optimal performance in the task is thus given by performing

maximum likelihood discrimination (MLD [25]). Importantly, since the noise is Gaussian, the task can be performed optimally by a linear discriminator reading out directly from the input channels and using weights parallel to the signal (**Fig. 1H**), defined as the unit vector

$$\boldsymbol{s} = \left(\boldsymbol{f}^0(\theta_+) - \boldsymbol{f}^0(\theta_-)\right) \|\boldsymbol{f}^0(\theta_+) - \boldsymbol{f}^0(\theta_-)\|^{-1}. \tag{6}$$

Thus the output in this scenario equals $\boldsymbol{s}^T\boldsymbol{x}^0(\theta)$ which leads to optimal performance in this setup [45].

D. Pre-PL readout

We assume the pre-PL value of the readout weight vector $\boldsymbol{a}_{\text{pre}}$ to be optimized for this task when reading out the pre-PL top-layer representations. Thus, we initialize the pre-PL readout such that it minimizes the loss function between the network readout and the optimal output (see below and S. M. Sec. I). The rationale for non-random initialization of the readout weights is to provide the network with well-above-chance but generally suboptimal performance (as shown below, it is suboptimal because the top-layer representations may be suboptimal). In the context of animal experiments, this mimics the situation where animals understand the task but have not yet acquired the expert skills required for near optimal performance.

E. Learning

We model the process of PL as modifying weights in order to minimize the discrimination error. Since the optimal output for this task is given by $\boldsymbol{s}^T\boldsymbol{x}^0(\theta)$ it is convenient to use a mean-squared error objective function

$$E(\Theta) = \langle(\boldsymbol{a}^T\boldsymbol{x}^L(\theta, \boldsymbol{W}^1, ..., \boldsymbol{W}^L) - \boldsymbol{s}^T\boldsymbol{x}^0(\theta))^2\rangle_{\theta=\theta_{\text{tr}}\pm\delta\theta,\boldsymbol{\epsilon}^0}. \tag{7}$$

where $\Theta = (\boldsymbol{W}^1, \boldsymbol{W}^2, ..., \boldsymbol{W}^L, \boldsymbol{a})$ denotes the vector of all weights of the networks and angular brackets denote averaging over the two stimuli and noise. This cost function measures the deviation of the input-output function of the system from the optimal one.

Python scripts used for analysis can be found at `https://github.com/hzshan/perceptual_learning`.

III. A MEAN-FIELD THEORY OF PL OF THE FINE-DISCRIMINATION TASK

In this section, we describe our mean-field approach to studying sensory processing and PL in the deep, nonlinear network by approximating it with an equivalent linear network. We describe the approximation (Sec. III A) and the insights it provides into why pre-PL cortical representations can be suboptimal (Sec. III B). We discuss the space of possible solutions that this theory revealed in Sec. III C.

A. Equivalent linear networks

First, we note that, during the fine-discrimination task, signal and noise-induced fluctuations in the input to any neuron are small (they both scale as $N^{-1/2}$). This can be seen by considering the scaling relations $\delta\theta \sim O(N^{-1/2})$, $\sigma^2 \sim O(1)$ and $W^l_{ij} \sim O(N^{-1})$. In the large N limit, we can expand activities of cortical neurons around their average inputs by writing (using $\odot$ to denote the Hadamard product and $\boldsymbol{f}^l$ to denote $\boldsymbol{f}^l(\theta_{\rm tr})$; this is similar to the approximation done in [46] for recurrent networks)

$$\boldsymbol{x}^l(\theta_\pm) = \Phi\left(\boldsymbol{W}^l\boldsymbol{x}^{l-1}(\theta_\pm)\right) \tag{8}$$

$$\approx \Phi\left(\boldsymbol{W}^l\boldsymbol{f}^{l-1}\right) + \Phi'\left(\boldsymbol{W}^l\boldsymbol{f}^{l-1}\right) \odot \left[\pm\delta\theta \boldsymbol{W}^l d_\theta \boldsymbol{f}^{l-1} + \boldsymbol{W}^l\boldsymbol{\epsilon}^{l-1}\right], \tag{9}$$

where

$$d_\theta \boldsymbol{f}^l = \frac{\left[\boldsymbol{f}^l(\theta_+) - \boldsymbol{f}^l(\theta_-)\right]}{2\delta\theta} \tag{10}$$

and $\boldsymbol{\epsilon}^l = \frac{1}{2}\left[\boldsymbol{x}^l(\theta_+) + \boldsymbol{x}^l(\theta_-) - \boldsymbol{f}^l(\theta_+) - \boldsymbol{f}^l(\theta_-)\right]$ are the signal-induced and noise-induced fluctuations in layer l, respectively. At large N, by the central limit theorem, the components of noise are Gaussian (though correlated). For the ReLU nonlinearity, the activation slope Φ'_i is 1 for an *active* neuron and zero for an *inactive* one. In the limit of large N, fluctuations in the input to each neuron are small compared to the mean. Thus, inactive neurons remain quiescent for most of the trials and do not contribute to the network output. For a similar reason, activities of active neurons are $[\boldsymbol{W}^l\boldsymbol{f}^{l-1}]_i$ i.e, they are *linear* functions of activities of neurons in the previous layer. Thus, we can define effective weight matrices , $\boldsymbol{W}^l_{\rm eff}$, as

$$\left(W^l_{\rm eff}\right)_{ij} = W^l_{ij}, \text{ iff } \; f^l_i > 0 \; \text{ and } \; f^{l-1}_j > 0 \tag{11}$$

and zero otherwise. Importantly, this approximation holds only during fine-discrimination around a fixed $\theta_{\rm tr}$ because of the strong similarity between the different inputs. Inputs with

angles very different from $\theta_{\rm tr}$ will be processed by different sets of effective weights. Given the analysis above, during the fine-discrimination task around $\theta_{\rm tr}$ the input output function of the deep network is effectively linear,

$$r(\theta) = \boldsymbol{a}^T \boldsymbol{P} \boldsymbol{x}^{\boldsymbol{0}}(\theta), \tag{12}$$

where

$$\boldsymbol{P} = \boldsymbol{W}^L_{\rm eff} \boldsymbol{W}^{L-1}_{\rm eff} ... \boldsymbol{W}^1_{\rm eff}. \tag{13}$$

We call P the *processing matrix* (**Fig. 2A**, right). We proceed to consider how the properties of P affect task performance.

B. Pre-learning suboptimal representations

Optimizing $\boldsymbol{a}_{\rm pre}$ amounts to optimizing a linear readout from an input $\boldsymbol{P}\boldsymbol{x}^{\boldsymbol{0}}(\theta)$ which contains a signal and an additive (correlated) noise. In such a system, the probability of error under optimal readout is given by $H(\delta\theta\sqrt{J_L})$ where $H(x) = (2\pi)^{-1/2}\int_x^\infty e^{-z^2/2}dz$ and J_L is the linear Fisher information [47]. It is defined as

$$J_L = (d_\theta \boldsymbol{f}^L)^T \boldsymbol{\Sigma}_L^{-1} d_\theta \boldsymbol{f}^L, \tag{14}$$

where the matrix $\boldsymbol{\Sigma}_L$ is the noise covariance matrix in the top layer. Given the linear approximation, it is given by $\boldsymbol{\Sigma}_L = \sigma^2 \boldsymbol{P}\boldsymbol{P}^T$ and the top layer signal is $d_\theta \boldsymbol{f}^L = \boldsymbol{P} d_\theta \boldsymbol{f}^0$. Even if $\boldsymbol{P}$ is low-rank (see below), the (pseudo)-inverse $(\boldsymbol{P}\boldsymbol{P}^T)^{-1}$ is well defined when multiplied by $\boldsymbol{P}$. Hence,

$$J_L = \sigma^{-2} (d_\theta \boldsymbol{f}^0)^T \boldsymbol{P}^T (\boldsymbol{P}\boldsymbol{P}^T)^{-1} \boldsymbol{P} d_\theta \boldsymbol{f}^0. \tag{15}$$

Note that $\boldsymbol{P}^T(\boldsymbol{P}\boldsymbol{P}^T)^{-1}\boldsymbol{P}$ is a projection matrix. It is identity if $\boldsymbol{P}$ is full-rank. Otherwise, it projects inputs onto the low-rank subspace spanned by its rows. Thus, Eq. 15 states that $\sigma^2 J_L$ is the squared norm of the projection of the signal vector $d_\theta \boldsymbol{f}^0 \propto \boldsymbol{s}$ onto the subspace spanned by $\boldsymbol{P}$. The network is optimal if $\boldsymbol{s}$ resides in the span of $\boldsymbol{P}$, yielding $J_L = J_0 = \sigma^{-2}\|d_\theta \boldsymbol{f}^0\|^2$.

We now ask whether the pre-PL weights are already optimal for the present task. We computed the singular value decomposition of the pre-PL $\boldsymbol{P}$ and found that it has a low-rank structure (Fig. S2). There are two sources of the reduced rank of $\boldsymbol{P}$ depending on the system parameter regime: a "selective-input-unselective-weights" regime characterized

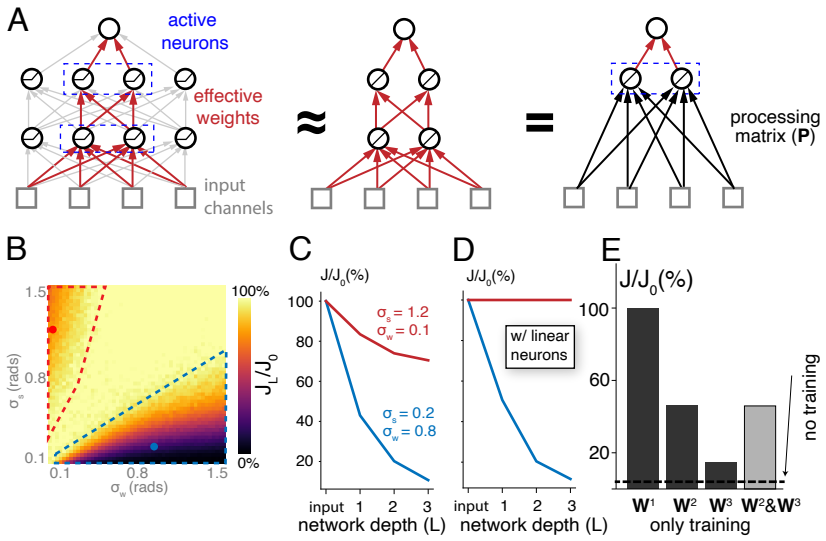

FIG. 2. **Suboptimal neural representations before learning.**

A Schematics showing the relationship between weights (all arrows, left), effective weights (red arrows, left and center), and the processing matrix (black arrows, right).

B Information for the trained stimulus in the last layer (J_L) divided by the input information (J_0), for different input and weight selectivity. The ratio is low for large σ_s, small σ_w(red polygon) or small σ_s, large σ_w (blue polygon). Dots: example parameters used in C, D. $N = 1000$ in all panels. $L = 1$ in this panel. See Fig. S1 for deeper networks.

C Information for the trained stimulus in the last layer of networks of different depths, divided by the input-layer information (J_0).

D Same as C, but assuming that all neurons are in the linear regime.

E Best last-layer information achievable if plasticity is restricted to some weight matrices in a three-layer network. Dashed line: performance if no weight matrix is modified. Modifying any weight matrix improves the performance, but only modifying $\boldsymbol{W}^1$ is sufficient and necessary for optimizing it.

by small σ_s and large σ_w (**Fig. 2B**, blue) and an "unselective-input-selective-weights" regime, characterized by large σ_s and small σ_w (**Fig. 2B**, red). In the selective-input-unselective-weights regime, the pre-PL network weights $\boldsymbol{W}^l_{\text{pre}}$ are low-rank (even before rectification) due to the smoothness of circulant weights, implying that they project to subsequent layers only part of the signal in the input. In this regime, information loss occurs regardless of the rectification of representation neurons (**Fig. 2C, D**, blue line). On

the other hand, in the unselective-input-selective-weights (large σ_s, small σ_w) regime, the original weight matrices project the full signal. However, due to firing-rate rectification, a substantial fraction of the neurons are inactive for essentially all training stimuli. Thus, the effective weights are low-rank. In this regime, the low-rank structure of $\boldsymbol{P}$ disappears if we remove neuronal rectification (**Fig. 2C, D**, red line). In both cases, the signal contains a substantial component perpendicular to the low-rank span of pre-PL $\boldsymbol{P}$, as evidenced by computing J_L/J_0 (**Fig. 2B**). Hence the pre-PL network exhibits suboptimal performance.

C. Space of solutions

We derived the following *necessary and sufficient condition* on post-learning effective weights that solve the PL task by making the performance optimal (derivations in S. M. Sec. II; hereafter we use $\boldsymbol{W}^l$ to denote $\boldsymbol{W}^l_{\text{eff}}$ for brevity): for any values of $\boldsymbol{W}^2, ..., \boldsymbol{W}^L$ and $\boldsymbol{a}$ that satisfy $\tilde{\boldsymbol{a}} \equiv (\boldsymbol{W}^2)^T ... (\boldsymbol{W}^L)^T \boldsymbol{a} \neq \boldsymbol{0}$, the task can be performed optimally if and only if $\boldsymbol{W}^1$ satisfies

$$\boldsymbol{W}^1 = \|\tilde{\boldsymbol{a}}\|^{-2} \tilde{\boldsymbol{a}} \boldsymbol{s}^T + \boldsymbol{W}_\perp, \tag{16}$$

where $\boldsymbol{W}_\perp^T \tilde{\boldsymbol{a}} = 0$. This result implies that the learning problem can be solved for essentially arbitrary (non-zero) higher-layer weights as long as $\boldsymbol{W}^1$ is adjusted accordingly. Conversely, restricting the plasticity to higher-layer weights while freezing the first layer weights to their pre-PL values do not obey this condition; thus this is insufficient for optimal performance, as discussed earlier (**Fig. 2E**). This result underscores the critical role played by early sensory areas in PL.

IV. LEARNING WHILE MINIMIZING NETWORK PERTURBATION

The large space of solutions makes it hard to predict the pattern of changes in the circuit induced by learning. To remove this degeneracy, we propose an optimality criterion, "minimum perturbation" (MP), that favors a solution with small perturbations to pre-PL weights. According to this criterion, the optimal weights are

$$\Delta\Theta = \Theta_{\text{post}} - \Theta_{\text{pre}} \tag{17}$$

$$\Theta_{\text{opt}} = \arg\min_{\Theta_{\text{post}}} \|\Delta\Theta\|^2 \text{ s.t. } E(\Theta_{\text{post}}) = 0, \tag{18}$$

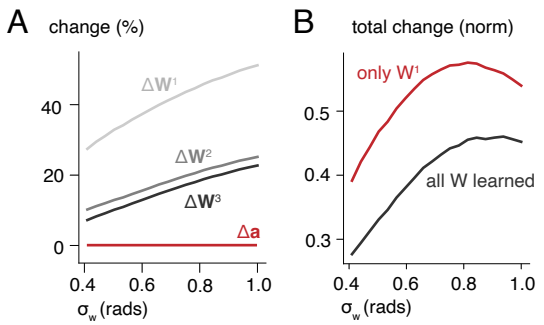

FIG. 3. **Minimum-perturbation plasticity of perceptual learning.**

A The magnitude of synaptic changes to each matrix and the readout vector $\boldsymbol{a}$, for networks initialized with different σ_w. Percent change is defined as the Frobenius norm of synaptic changes divided by that of the pre-PL weight matrix. In both panels, $\sigma_s = 0.2, L = 3, N = 1000$.

B Restricting learning to $\boldsymbol{W}^1$ leads to more network-wide perturbation (measured by the sum of matrix norms of $\Delta\boldsymbol{W}^1, \Delta\boldsymbol{W}^2, \Delta\boldsymbol{W}^3$) than unrestricted learning. In either case, the readout $\boldsymbol{a}$ is also allowed to learn but does not change significantly following PL.

where "post" indicates post-PL weights and Θ refers to all weights in the network. This principle is motivated by the brain's need to maintain relatively stable representations while learning a new task. We analytically solved this optimization problem with Lagrange multipliers by first setting up the Lagrangian

$$\mathcal{L} = \sum_{l=1}^{L} \|\Delta\boldsymbol{W}^l\|^2 + \|\Delta\boldsymbol{a}\|^2 - \boldsymbol{\lambda}^T[{\boldsymbol{W}^1_{\text{post}}}^T {\boldsymbol{W}^2_{\text{post}}}^T \ldots {\boldsymbol{W}^L_{\text{post}}}^T \boldsymbol{a}_{\text{post}} - \boldsymbol{s}]. \tag{19}$$

Extremizing the Lagrangian w.r.t. weight changes reveals a general rank-1 structure for MP $\Delta\boldsymbol{W}^l$

$$\Delta\boldsymbol{W}^l = ({\boldsymbol{W}^{l+1}_{\text{post}}}^T \cdots {\boldsymbol{W}^L_{\text{post}}}^T \boldsymbol{a}_{\text{post}})(\boldsymbol{W}^{l-1}_{\text{post}} \cdots \boldsymbol{W}^1_{\text{post}} \boldsymbol{\lambda})^T. \tag{20}$$

Solving the above equations requires introducing $2L$ scalar order parameters, which obey $2L$ self-consistent equations that need to be solved numerically. Expressions of the order parameters and self-consistent equations for $L = 1, 2, 3$, as well as the numerical procedure for solving the self-consistent equations, are given in S. M. Sec. III.

We discuss features of the solutions below.

Distribution of MP plasticity: We quantified the magnitude of MP modification to weights in each layer by computing $\|\Delta\boldsymbol{W}^l\|/\|\boldsymbol{W}^l_{\text{pre}}\|$ (**Fig. 3A**; see Fig. S6 for the unselective-input-selective-weights regime). The analysis reveals two salient features that

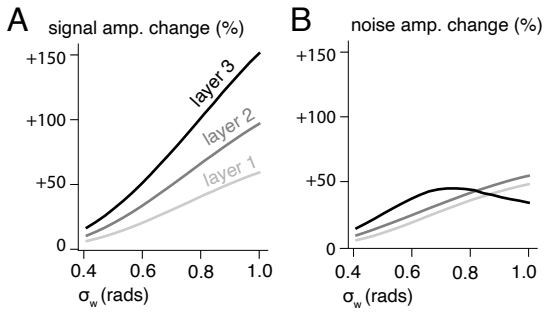

FIG. 4. **MP learning-induced changes to signal and noise.**

A, B PL-induced changes to signal (A) or noise (B) amplitude across layers for different weight selectivity. Changes are generally greater in higher layers and in networks with initial weights that are less selective (larger σ_w). In both panels, $\sigma_s = 0.4, N = 1000, L = 3$.

are consistent across system parameters. First, MP plasticity predominantly affects lower-layer weights. Second, surprisingly, MP plasticity does not appreciably alter the readout (**Fig. 3A**, red line). This suggests that hidden-layer representation changes, rather than readout changes, drive performance improvement. Plasticity in higher-layer weights $\boldsymbol{W}^{l\geq 2}$ plays the important role of reducing the overall perturbation to the network. Indeed, if we restrict learning to $\boldsymbol{W}^{\mathbf{1}}$, the total perturbation is greater (**Fig. 3B**).

Performance improvement is driven by signal amplification: An important and long-standing debate in PL research is whether behavioral improvement is driven by signal amplification, noise suppression, or both [3, 5, 6, 9, 48–51].

To address this question within the framework of MP learning, we define the signal and noise contributions via $J_L = (\text{signal/noise})^2$, where the signal amplitude is $\|d_\theta \boldsymbol{f}^L\|$ and the noise amplitude is defined via $\text{noise}^{-2} = (d_\theta \boldsymbol{f}^L)^T \boldsymbol{\Sigma}_{L-1} d_\theta \boldsymbol{f}^L / \|d_\theta \boldsymbol{f}^L\|^2$. The network after MP learning exhibits a pronounced amplification of signal (**Fig. 4A**), with the effect being stronger in higher layers. Surprisingly, we found that PL also *amplifies* noise across all layers, although to a weaker extent than signal amplification (**Fig. 4B**; further details in S. M. Sec. VII; effects on noise correlation are shown in Fig. S5). Thus, MP learning improves perceptual performance by strengthening the signal rather than weakening the noise. This analysis also reveals that signal/noise changes are generally greater in higher layers even though weight changes are greater in lower layers, highlighting the difference between distribution of weight changes and distribution of representation changes.

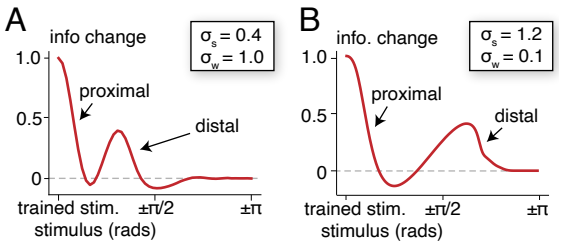

FIG. 5. **Transfer of PL to untrained stimuli.**

A, B Information changes in the last layer for different stimuli after PL, normalized by change for the trained stimulus. The change for the trained stimulus is 1 by definition. Information gain is prominent for stimuli close to the trained one ("proximal"), and those dissimilar from the trained one ("distal"). In all panels, $N = 1000, L = 3$, and the last layer is analyzed. For the selective-input-unselective-weights regime (**A**), $\sigma_s = 0.4, \sigma_w = 1.0$. For the unselective-input-selective-weights regime (**B**), $\sigma_s = 1.2, \sigma_w = 0.1$.

Impact on discrimination around untrained stimuli: MP plasticity breaks the symmetry of pre-PL representations w.r.t. θ, thus altering the representations of untrained stimuli. To assess how these changes affect the discrimination ability of angles around untrained values we define a normalized information gain, $[J_{L,\text{post}}(\theta) - J_{L,\text{pre}}(\theta)]/[J_{L,\text{post}}(\theta_{\text{tr}}) - J_{L,\text{pre}}(\theta_{\text{tr}})]$ for an untrained stimulus θ. Our analysis revealed a rich, non-monotonic pattern of transfer arising from MP plasticity. Consistent with experimental findings, PL increases information for stimuli similar to the trained one ("proximal transfer", **Fig. 5**). In addition, PL also transfers to distal stimuli, where the distance between trained and test stimuli is intermediate ("distal transfer"). Importantly, PL can also decrease information for certain untrained stimuli (negative transfer), as revealed by the dips below 0 in **Fig. 5**. Finally, as expected, representations for stimuli far away from the trained one are unaffected by learning.

V. MP LEARNING WITH GRADIENT DESCENT

So far, our analysis has focused on properties of MP plasticity without addressing the important question of *how* such plasticity is learned. We modeled the process of learning by studying gradient descent (GD), which has been shown to reproduce physiological features of PL in deep network models [22, 52]. We used GD to optimize Θ for a regularized loss

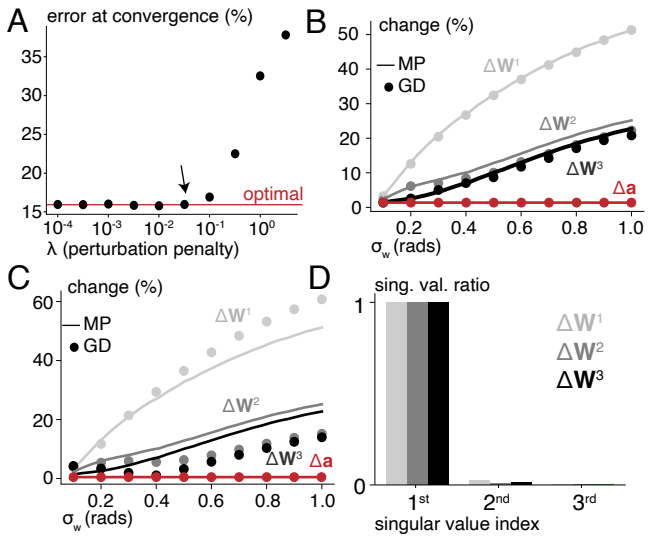

FIG. 6. **MP learning with gradient descent.**

A Discrimination error rate at convergence after regularized gradient descent with different perturbation penalty strength. Arrow: maximum λ with optimal performance. In all panels $\sigma_s = 0.2, L = 3$; $\sigma_w = 0.8$ except in B, C.

B Magnitude of changes from regularized gradient descent (dots), compared against that of MP plasticity (line).

C Same as B, but for slow GD without explicit regularization.

D Leading singular values of slow GD induced changes to weight matrices (normalized by the top singular value). That the first singular value is overwhelmingly large suggests that induced changes are close to rank-1.

function, defined as

$$E_{\text{reg}}(\Theta) = E(\Theta) + \lambda \|\Delta\Theta\|^2, \tag{21}$$

where the second term imposes a penalty on weight perturbation; strength of the penalty is controlled by the hyperparameter λ. We implemented GD by iterating $\Theta(t+1) = \Theta(t) - \eta \nabla_{\Theta(t)} E_{\text{reg}}(\Theta)$ until convergence, where η is the learning rate. At convergence, larger λ results in smaller weight perturbations but potentially suboptimal post-PL performance. To realize MP learning, λ should be as large as possible without making the final performance suboptimal (**Fig. 6A**). GD with such λ results in weight changes that are fully consistent with MP plasticity (data not shown).

In the deep learning literature, it has been suggested that small changes in learned pa-

rameters can also be realized through implicit regularization by using small learning rates [42, 53]. We performed GD on the unreguarlized loss function, $E(\Theta)$, with a small learning rate η. The resultant plasticity agrees reasonably well with MP plasticity (**Fig. 6C)** in terms of magnitude. It also has the same salient features as MP plasticity: changes to the readout are negligible, weight changes are very close to being rank-1 (**Fig. 6D**) and the identity of active neurons does not change over learning (S. M. Sec. V), as is the case for MP plasticity. These results point to the possibility that the slow progression of PL could be normatively explained as a mechanism to minimize perturbation during PL.

VI. DISCUSSION

We have presented a theory of PL of fine discrimination in a deep network. The theory leverages similarity of all inputs relevant to the task, large network size and structured weight initialization to establish the effective linearity of the network input-output function during training and performance of the task. This input-output function is expressed by a processing matrix $\boldsymbol{P}$ which has been shown to be low-rank, hence leading to a sub-optimal representation of the stimulus that cannot be resolved by adapting the readout weight only. We further derived the space of post-learning weights that resolve the suboptimality by fully spanning the task-relevant signal direction. Motivated by the brain's need to strike a balance between plasticity (acquiring new skills) and stability (preventing previously learned skills to be affected) in sensory areas [54], we propose the normative minimum perturbation principle that favors a specific solution. The favored solution, which we call MP plasticity, induces physiological and behavioral changes largely consistent with current experimental findings (for a detailed comparison, see S. M. Sec. VIII). It also predicts that PL improves the sensory code for some untrained stimuli while degrading the representation of others, a readily testable prediction. We discuss some prominent features of MP plasticity and their implications for neural mechanisms of PL.

First, MP plasticity predominantly modifies the lowest-level weights while leaving the readout essentially unchanged. This points to the importance of involving low-level cortical areas in PL of fine discrimination, consistent with recent numerical experiments with deep convolutional networks [52]. That the readout is unchanged critically depends on our assumption that the pre-PL readout is already optimized w.r.t. pre-learning representations,

in contrast to most neural network models of learning where the initial weights are random. We argue that random initialization is not biologically plausible when considering naturalistic tasks where subjects perform well above chance with little-to-no training. Note that while synaptic plasticity is greater in lower layers, the resultant representational changes are greater in *higher* layers.

Second, MP plasticity makes rank-1 modifications to weights. While rank-1 weight changes are sufficient for optimizing neural representations for the trained task, such changes can be highly task-specific. To demonstrate this, we analyzed performance of the post-PL network on a width discrimination task where two input-layer patterns with the same θ but different σ_s are presented; PL does not improve this performance, despite the fact that width discrimination and angle discrimination involve the same mean input-layer pattern (S. M. Sec. IV). Importantly, this absence of cross-task transfer reconciles the apparent inconsistency of the observed improved sensory representations by PL [3, 7, 8, 10] and the psychophysical findings that PL for one task did not transfer to another task using the same stimuli [18–20], which was interpreted as evidence that population codes for these stimuli did not improve [55]. Our results suggest that the improvement of representations does not equally benefit all tasks even if they share the same stimuli. Thus, cross-task transfer is not a reliable indicator of whether representations improve after PL.

Finally, from the perspective of signal and noise, MP plasticity improves task performance by amplifying the signal. This result is inconsistent with [23], who found that amplification is not necessary for PL. Their conclusion may be confined to the regime where performance is dominated by neural noise, not input noise as in ours. Additionally, their plasticity model differs from ours in that it assumes circularly invariant weights both before and after learning, which forces a global change of synaptic weights. In contrast, in our model, PL plasticity is localized to the neurons responding to the stimulus (if we require post-PL weights to be circularly invariant in our model, post-PL tuning curves have very unnatural multi-modal shapes. See S. M. Sec. IX). Finally, we note that our prediction of signal amplification stems from the fact that the readout layer remains essentially unchanged under MP learning. If the readout were adapted in ways that violate the MP principle, signal amplification is not always necessary (S. M. Sec. X).

Our current theory can be extended in several interesting directions. First, our plasticity model does not include a mechanism of unsupervised learning, namely, plasticity triggered

by the mere exposure to the stimulus, independent of task. Thus, including considerations for task-irrelevant plasticity, observed in some PL studies [56, 57], is an interesting topic for future work. Second, the scenario where the input contains stimulus-dependent noise correlation, which requires nonlinear readouts for optimal performance [58], is worth further analysis. We expect some qualitative findings from our work to generalize. In particular, when the first layer weight matrix $\boldsymbol{W^1}$ is low-rank and not modified by learning, it can cause irrecoverable loss of the linear component of the input Fisher information [58], making the network performance suboptimal despite changes to higher layers. Nevertheless, changes in the higher layers may also be necessary (depending on hyperparameters such as the width of the first layer) to recover the nonlinear components of the task. We also conjecture that, MP changes in the higher-layer weights will be more extensive than what is necessary to solve the task, in order to minimize the overall perturbations in the weights. Finally, our formulation of the MP principle implicitly assumes that neural representations are stable in the absence of learning. While this appears to hold true for artificial stimuli commonly used in PL experiments [59, 60], representations of naturalistic stimuli are known to drift over time in the absence of explicit training [61–63]. It has been suggested that representational drift may induce adaptive processes in downstream populations that essentially compensate for the drift, keeping the underlying computation intact [63]. If this is true, relative to a given state of the readout population, MP learning may be an adequate strategy.

ACKNOWLEDGMENTS

The authors would like to thank Andrew Saxe and Ravid Ziv for very helpful discussions. This research was partially supported by the Swartz Program in Theoretical Neuroscience at Harvard University, the Gatsby Charitable Foundation, the National Institute of Neurological Disorders and Stroke (Grant No.1U19NS104653), and the National Science Foundation (Grant No.1806818). This paper is dedicated to the memory of Mrs. Lily Safra, a great supporter of brain research.

[1] G. H. Recanzone, M. M. Merzenich, W. M. Jenkins, K. A. Grajski, and H. R. Dinse, Journal of Neurophysiology **67**, 1031 (1992).

[2] G. H. Recanzone, C. E. Schreiner, and M. M. Merzenich, Journal of Neuroscience **13**, 87 (1993).

[3] A. Schoups, R. Vogels, N. Qian, and G. Orban, Nature **412**, 549 (2001).

[4] G. M. Ghose, T. Yang, and J. H. Maunsell, Journal of neurophysiology **87**, 1867 (2002).

[5] T. Yang and J. H. Maunsell, Journal of Neuroscience **24**, 1617 (2004).

[6] S. Raiguel, R. Vogels, S. G. Mysore, and G. A. Orban, Journal of Neuroscience **26**, 6589 (2006).

[7] Y. Gu, S. Liu, C. R. Fetsch, Y. Yang, S. Fok, A. Sunkara, G. C. DeAngelis, and D. E. Angelaki, Neuron **71**, 750 (2011).

[8] Y. Yan, M. J. Rasch, M. Chen, X. Xiang, M. Huang, S. Wu, and W. Li, Nature neuroscience **17**, 1380 (2014).

[9] M. Sanayei, X. Chen, D. Chicharro, C. Distler, S. Panzeri, and A. Thiele, Nature communications **9**, 1 (2018).

[10] A. M. Ni, D. A. Ruff, J. J. Alberts, J. Symmonds, and M. R. Cohen, Science **359**, 463 (2018).

[11] I. Maor, R. Shwartz-Ziv, L. Feigin, Y. Elyada, H. Sompolinsky, and A. Mizrahi, Frontiers in neural circuits **13**, 82 (2020).

[12] H. R. Dinse, P. Ragert, B. Pleger, P. Schwenkreis, and M. Tegenthoff, Science **301**, 91 (2003).

[13] A. Karni and D. Sagi, Proceedings of the National Academy of Sciences **88**, 4966 (1991).

[14] A. Fiorentini and N. Berardi, Nature **287**, 43 (1980).

[15] C.-T. Law and J. I. Gold, Nature neuroscience **11**, 505 (2008).

[16] H. Z. Adab, I. D. Popivanov, W. Vanduffel, and R. Vogels, Journal of cognitive neuroscience **26**, 2187 (2014).

[17] H. O. de Beeck, J. Wagemans, and R. Vogels, Neuroscience **145**, 775 (2007).

[18] M. Ahissar and S. Hochstein, Proceedings of the National Academy of Sciences **90**, 5718 (1993).

[19] R. E. Crist, M. K. Kapadia, G. Westheimer, and C. D. Gilbert, Journal of neurophysiology **78**, 2889 (1997).

[20] M. Fahle, Vision research **37**, 1885 (1997).

[21] M. Ahissar and S. Hochstein, Trends in cognitive sciences **8**, 457 (2004).

[22] A. M. Saxe, *Deep linear neural networks: A theory of learning in the brain and mind* (Stanford University, 2015).

[23] V. R. Bejjanki, J. M. Beck, Z.-L. Lu, and A. Pouget, Nature neuroscience **14**, 642 (2011).

[24] S. A. Chowdhury and G. C. DeAngelis, Neuron **60**, 367 (2008).

[25] H. S. Seung and H. Sompolinsky, Proceedings of the national academy of sciences **90**, 10749 (1993).

[26] A. A. Petrov, B. A. Dosher, and Z.-L. Lu, Psychological review **112**, 715 (2005).

[27] B. A. Dosher, P. Jeter, J. Liu, and Z.-L. Lu, Proceedings of the National Academy of Sciences **110**, 13678 (2013).

[28] J. J. DiCarlo and D. D. Cox, Trends in cognitive sciences **11**, 333 (2007).

[29] D. L. Yamins and J. J. DiCarlo, Nature neuroscience **19**, 356 (2016).

[30] A. J. Kell and J. H. McDermott, Current opinion in neurobiology **55**, 121 (2019).

[31] J. Lee, Y. Bahri, R. Novak, S. S. Schoenholz, J. Pennington, and J. Sohl-Dickstein, arXiv preprint arXiv:1711.00165 (2017).

[32] A. Jacot, F. Gabriel, and C. Hongler, Advances in neural information processing systems **31** (2018).

[33] B. Bordelon and C. Pehlevan, arXiv preprint arXiv:2205.09653 (2022).

[34] Q. Li and H. Sompolinsky, Physical Review X **11**, 031059 (2021).

[35] A. M. Saxe, J. L. McClelland, and S. Ganguli, arXiv preprint arXiv:1312.6120 (2013).

[36] S. Mei, A. Montanari, and P.-M. Nguyen, Proceedings of the National Academy of Sciences **115**, E7665 (2018).

[37] L. Zdeborová, Nature Physics **16**, 602 (2020).

[38] Y. Bahri, J. Kadmon, J. Pennington, S. S. Schoenholz, J. Sohl-Dickstein, and S. Ganguli, Annual Review of Condensed Matter Physics **11** (2020).

[39] D. A. Roberts, S. Yaida, and B. Hanin, arXiv preprint arXiv:2106.10165 (2021).

[40] R. Dietrich, M. Opper, and H. Sompolinsky, Physical review letters **82**, 2975 (1999).

[41] M. S. Advani, A. M. Saxe, and H. Sompolinsky, Neural Networks **132**, 428 (2020).

[42] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, Communications of the ACM **64**, 107 (2021).

[43] E. P. Simoncelli and D. J. Heeger, Vision research **38**, 743 (1998).

[44] N. C. Rust, V. Mante, E. P. Simoncelli, and J. A. Movshon, Nature neuroscience **9**, 1421 (2006).

[45] C. R. Rao, Journal of the Royal Statistical Society. Series B (Methodological) **10**, 159 (1948).

[46] J. Beck, V. R. Bejjanki, and A. Pouget, Neural computation **23**, 1484 (2011).

[47] P. Seriès, P. E. Latham, and A. Pouget, Nature neuroscience **7**, 1129 (2004).

[48] J. Poort, A. G. Khan, M. Pachitariu, A. Nemri, I. Orsolic, J. Krupic, M. Bauza, M. Sahani, G. B. Keller, T. D. Mrsic-Flogel, *et al.*, Neuron **86**, 1478 (2015).

[49] E. Zohary, S. Celebrini, K. H. Britten, and W. T. Newsome, Science **263**, 1289 (1994).

[50] H. Z. Adab and R. Vogels, Current biology **21**, 1661 (2011).

[51] J. Gold, P. Bennett, and A. Sekuler, Nature **402**, 176 (1999).

[52] L. K. Wenliang and A. R. Seitz, Journal of Neuroscience **38**, 6028 (2018).

[53] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro, in *International Conference on Machine Learning* (PMLR, 2018) pp. 1832–1841.

[54] W. C. Abraham and A. Robins, Trends in neurosciences **28**, 73 (2005).

[55] B. Dosher and Z.-L. Lu, Annual Review of Vision Science **3**, 343 (2017).

[56] T. Watanabe, J. E. Nanez, and Y. Sasaki, Nature **413**, 844 (2001).

[57] A. R. Seitz and T. Watanabe, Vision research **49**, 2604 (2009).

[58] M. Shamir and H. Sompolinsky, Neural computation **16**, 1105 (2004).

[59] B. B. Jeon, A. D. Swain, J. T. Good, S. M. Chase, and S. J. Kuhlman, Scientific Reports **8**, 1 (2018).

[60] T. D. Marks and M. J. Goard, Nature communications **12**, 1 (2021).

[61] Y. Ziv, L. D. Burns, E. D. Cocker, E. O. Hamel, K. K. Ghosh, L. J. Kitch, A. E. Gamal, and M. J. Schnitzer, Nature neuroscience **16**, 264 (2013).

[62] D. Deitch, A. Rubin, and Y. Ziv, Current biology **31**, 4327 (2021).

[63] M. E. Rule, T. O'Leary, and C. D. Harvey, Current opinion in neurobiology **58**, 141 (2019).

[64] L. M. Vaina, J. W. Belliveau, E. B. Des Roziers, and T. A. Zeffiro, Proceedings of the National Academy of Sciences **95**, 12657 (1998).

[65] M. Sigman, H. Pan, Y. Yang, E. Stern, D. Silbersweig, and C. D. Gilbert, Neuron **46**, 823 (2005).

[66] J. F. Jehee, S. Ling, J. D. Swisher, R. S. van Bergen, and F. Tong, Journal of Neuroscience **32**, 16747 (2012).

[67] Y. Wang, W. Wu, X. Zhang, X. Hu, Y. Li, S. Lou, X. Ma, X. An, H. Liu, J. Peng, *et al.*, Frontiers in behavioral neuroscience **10**, 42 (2016).

[68] R. E. Crist, W. Li, and C. D. Gilbert, Nature neuroscience **4**, 519 (2001).

[69] B. Pleger, A.-F. Foerster, P. Ragert, H. R. Dinse, P. Schwenkreis, J.-P. Malin, V. Nicolas, and M. Tegenthoff, Neuron **40**, 643 (2003).

[70] S. Schwartz, P. Maquet, and C. Frith, Proceedings of the National Academy of Sciences **99**, 17137 (2002).

[71] C. S. Furmanski, D. Schluppeck, and S. A. Engel, Current Biology **14**, 573 (2004).

[72] C. Schiltz, J. Bodart, S. Dubois, S. Dejardin, C. Michel, A. Roucoux, M. Crommelinck, and G. Orban, Neuroimage **9**, 46 (1999).

[73] A. A. Schoups, R. Vogels, and G. A. Orban, The Journal of physiology **483**, 797 (1995).

[74] M. Fahle and S. Edelman, Vision research **33**, 397 (1993).

[75] L.-P. Shiu and H. Pashler, Perception & psychophysics **52**, 582 (1992).

[76] P. E. Jeter, B. A. Dosher, A. Petrov, and Z.-L. Lu, Journal of vision **9**, 1 (2009).