

Supplementary Information for:

Title: Perceptual Learning as Improved Probabilistic Inference in Early

Sensory Areas

Authors: Vikranth R. Bejjanki, Jeffrey M. Beck, Zhong-Lin Lu and

Alexandre Pouget

Table of Contents

Supplementary Figure S1: Sensitivity to initial weights

Supplementary Figure S2: The effect of subsampling

Supplementary Note

Supplementary Table 1: Parameters for networks shown in Fig. 3 (in the main text)

Supplementary Table 2: Parameters for networks shown in Fig. 4 (in the main text)

Supplementary Figures

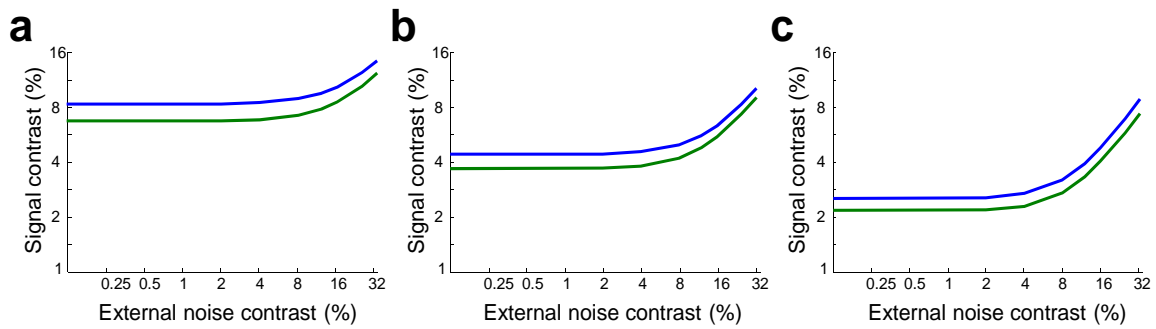


Figure S1: Sensitivity to initial weights. **a–c.** TVC curves before (blue) and after (green) learning for three different networks in each of which we used a different initial pattern of feed-forward thalamo-cortical weights but in all of which we moved the weights towards a matched filter. Moving towards a matched filter leads to a uniform shift in the TVC curve, independent of the precise pattern of initial weights. All TVC curves were obtained for the 79.3% correct performance criterion.

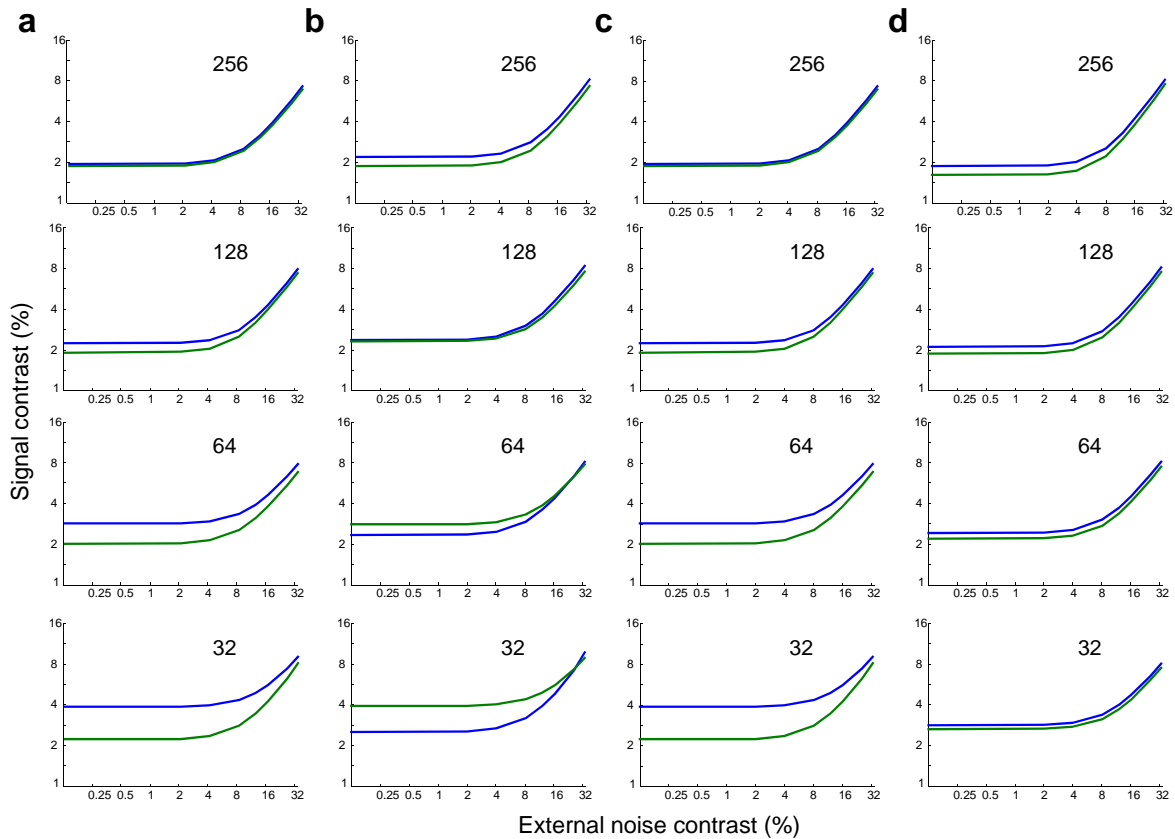


Figure S2: The effect of subsampling. **a–d.** TVC curves obtained from subsets of neurons, using the same procedure as in Fig. 7 in the main text, for the networks shown in Fig. 4 in the main text. Each column corresponds to the same column as in Fig. 4. The blue curves correspond to performance before learning, while the green curves correspond to performance after training session 1. With 256 neurons, the results are similar to the results obtained with the full networks (Fig. 4e–h). With only 32 neurons, the results start mimicking the results obtained with I_{shuffled} (Fig. 4i–l). This reflects the fact that neurons tend to be independent in small subsets. All TVC curves were obtained for the 79% correct performance criterion.

Supplementary Note

I. Replicating orientation discrimination with external noise

Stimulus design: As described in the main text, we replicated the experiment of Doshier & Lu (1999)¹. In Doshier & Lu (1999), subjects carried out a central fixation task while simultaneously being exposed to orientation discrimination stimuli in the periphery. It was the peripheral task that subjects became better at as a result of perceptual learning. In our study, we only considered the peripheral part of the task – in all our conditions, the stimulus consisted of an oriented Gabor pattern that was tilted θ_s° to the right or left of vertical. Formally, each stimulus image was created by assigning grayscale values to image pixels according to the following function:

$$Z(x, y, \theta) = Z_0 \left(1.0 + \left(c e^{\left[\frac{Cx^2}{2\sigma_x^2} + \frac{Cy^2}{2\sigma_y^2} \right]} \cos(2\pi KCx) \right) \right) \quad (S1)$$

where

$$Cx = x \cos \theta + y \sin \theta;$$

$$Cy = y \cos \theta - x \sin \theta;$$

$$\theta = \text{rad}(90 \pm \theta_s)$$

In the above formalism, x and y are the horizontal and vertical co-ordinates respectively, K is the spatial frequency of the Gabor pattern, σ_x and σ_y are the standard deviation (extent) of the Gabor in the x and y directions respectively, Z_0 is the mean, or background, grayscale value for the image and c is the maximum contrast of the Gabor

pattern as a proportion of the maximum achievable contrast. In all our experimental conditions, θ_s° was set to 12° , K was set to 0.75 cycles/deg, σ_x was set to 0.4, σ_y was set to 0.4 and Z_0 was set to 126.22 which was the equivalent mean grayscale value used in Doshier & Lu's studies^{1, 2}. The maximum contrast of the Gabor, labeled the signal contrast, varied depending on the experimental condition (see **Section III** below). Each Gabor pattern extended over $2.3^\circ \times 2.3^\circ$ of visual angle and was rendered on a 23×23 pixel grid. However, it should be noted that the actual stimulus image extended over a span of 45×45 pixels. In addition to the Gabor pattern, each stimulus image was padded with 22 extra pixels in the horizontal and vertical dimensions (11 at each end of the image), which were set to the background grayscale value. The reason for this padding was that the retinal cells in our model of orientation discrimination, in line with experimental data, had circular center-surround receptive fields (described in **Section II** below) and unless the image was padded, the receptive fields of the retinal cells at the corners of the square retinal array would extend beyond the extent of the image, leading to possible edge effects in the response of these cells. We added padding for an extent equivalent to two retinal surround field standard deviations, thereby accounting for the vast majority of the corner cells' receptive field extents.

Adding external noise: External noise was added to the Gabor stimulus in a manner similar to that used by Doshier & Lu. Pixel gray levels for the external noise were drawn from a Gaussian distribution with mean zero and standard deviation depending on the experimental condition (see **Section III** below). As in Doshier & Lu's studies, we used eight external noise levels in which the standard deviation of the external noise

distribution was varied as a proportion of the maximum achievable contrast. The effective noise levels we used were: 0.005%, 2%, 4%, 8%, 12%, 16%, 25% and 33%. Each noise element included a single pixel and the noise gray level values were added to the stimulus gray level values on a pixel by pixel basis to generate the noise-injected image.

II. Network model of orientation discrimination

Network architecture: We developed a network model of orientation discrimination containing four stages: retina, LGN, V1 and a decoder. Several aspects of the model are based on previous models of orientation discrimination, particularly the models in Somers et al. (1995)³ and Series et al. (2004)⁴. The first three stages, which take as input an image of a noisy, oriented Gabor pattern and which give as output a pattern of V1 orientation-tuned activity, model the major processing steps involved in early sensory processing in the visual system. The retina consists of uncoupled analog units that are driven by the image and output an analog firing rate. The retina feeds into the Lateral Geniculate Nucleus (LGN), which consists of a layer of uncoupled, spiking neurons. The LGN in turn feeds into V1, which also consists of spiking neurons, but which are coupled through lateral connections. The fourth stage of the network models the processing that occurs in later decision stages. This stage includes a single unit, with connections to each of the cortical cells, that takes as input the activities of the units in the V1 layer and which gives as output an estimate of the orientation of the stimulus. We describe each layer of the network in the following sections.

Retina: The retina is modeled after Series et al. (2004) (in turn modeled after Somers et al. 1995) and contains units in two layers. One difference between the retina modeled in this network and that modeled in previous networks is that we only model the spatial receptive field properties of retinal cells, and not the temporal receptive field properties. We feel justified in doing so since the stimuli we use are temporally stationary. One retinal layer consists of ON center-surround cells and the other OFF center-surround cells. Each layer contains 529 cells arranged in a 23 by 23 array, and the center-to-center spacing between cells, expressed in degrees of visual angle, is 0.1° .

The firing rate of a cell at location (x,y) is determined by the firing rates of the associated center and surround subfields. Specifically,

$$\begin{aligned} r_{ON}(x, y) &= G[r_{baseline} + r_{center}(x, y) - r_{surround}(x, y)] \\ r_{OFF}(x, y) &= G[r_{baseline} - r_{center}(x, y) + r_{surround}(x, y)] \end{aligned} \tag{S2}$$

The expressions in (S2) highlight another difference between the retina modeled in this network and that modeled in previous networks. In Series et al. (2004), the firing rate of a retinal cell was rectified using a threshold-linear function in a manner that set the firing rate of the cell to zero, when the expressions within the square brackets on the right hand side of (S2) were less than zero. In our model, we implemented a smooth rectification function G which had the following form:

$$G(x) = \frac{\mu}{\lambda} \ln(1 + e^{\lambda(x-\theta)}) \tag{S3}$$

We used $\mu = 1.0$, $\lambda = 0.2$ and $\theta = 0.0$

The center and surround subfield responses were generated by convolving the image with a spatial receptive field. Letting $\alpha = \{\text{center, surround}\}$, the subfield responses are given by:

$$r_{\alpha}(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F_{\alpha}(x - x', y - y') I(x', y') dx' dy' \quad (\text{S4})$$

The center and surround receptive fields $F_{\alpha}(x, y)$ are modeled as circularly symmetric Gaussians,

$$F_{\alpha}(x, y) = \frac{K_{\alpha}}{2\pi\sigma_{\alpha}^2} e^{-\frac{x^2+y^2}{2\sigma_{\alpha}^2}} \quad (\text{S5})$$

We used $\sigma_{\text{center}} = 0.176^{\circ}$, $\sigma_{\text{surround}} = 0.53^{\circ}$, $K_{\text{center}} = 16$, $K_{\text{surround}} = 16.64$, and $r_{\text{baseline}} = 15$ spk/s.

As described in **Section I**, the stimulus is a stationary, Gabor pattern of width w , length l and percent contrast c . Without loss of generality, we center the Gabor pattern at the origin (0,0) and compute $r_{\alpha}(x, y)$ via equations (S4) and (S5) above as:

$$r_{\alpha}(x, y) = q(c) K_{\alpha} \left(\int_{-\frac{w}{2}}^{\frac{w}{2}} \frac{1}{\sqrt{2\pi\sigma_{\alpha}^2}} e^{-\frac{(x-x')^2}{2\sigma_{\alpha}^2}} dx' \right) \left(\int_{-\frac{l}{2}}^{\frac{l}{2}} \frac{1}{\sqrt{2\pi\sigma_{\alpha}^2}} e^{-\frac{(y-y')^2}{2\sigma_{\alpha}^2}} dy' \right) \quad (\text{S6})$$

where $q(c)$ is the effective intensity of the stimulus at contrast c .

Note that in the above expression $r_{\alpha}(x, y)$ is implicitly dependent on the orientation of the Gabor pattern because the image co-ordinates $I(x', y')$ are a function of the Gabor orientation. In all of our experimental conditions, as described in **Section I**, the Gabor pattern's dimensions in degrees of visual angle were $w = 2.3^{\circ}$ and $l = 2.3^{\circ}$. The effective intensity $q(c)$ was defined to be

$$q(c) = \left(\frac{\beta [\log_{10}(c)]}{c} \right) \quad (S7)$$

with $\beta = 0.275$.

The expression in (S7) was chosen to account for the contrast dependence of LGN responses (see below).

LGN: Following Series et al. (2004), we assume a one-to-one correspondence between retinal ganglion cells and LGN cells, so that the response of each ganglion cell is passed on to one LGN cell of the same center polarity. The firing rate of an LGN cell at location (x, y) is either $r_{\text{ON}}(x, y)$ or $r_{\text{OFF}}(x, y)$, depending on whether the LGN cell has ON or OFF polarity. Again, as with the retina, we only model the spatial response properties of LGN cells and not the temporal properties, owing to the temporally stationary nature of our stimuli. Given the parameters used in modeling the retina, the peak response of the LGN cells (that is, the LGN cell with the largest firing rate) versus percent contrast (c), denoted R_{LGN} , is well fit by:

$$R_{LGN}(c) = r_{baseline} + 25[\log_{10}(c)] \quad (S8)$$

where, as in the previous section, $r_{baseline} = 15$ spk/s is the spontaneous firing rate. This relation is consistent with Somers *et al.* (1995) and with experimental data.

Note that this LGN model is simplified in several ways. It does not account for the mild orientation bias that has been reported in LGN responses or for the precise firing statistics and bursting in LGN. These properties are likely to influence the information available in the input to the cortex. However, it is unlikely that the fraction of this information that is transmitted to the cortical stage will depend critically on these assumptions, at least for the network regimes that we explored in this study.

VI – Feed forward thalamo-cortical connections: The cortical simple cell receptive field structure is established through a segregation of ON and OFF LGN inputs into 3 main subfields (OFF-ON-OFF). We model the receptive field of each cortical cell, with respect to the LGN, using a Gabor function $gab(x, y, \theta)$ defined by:

$$gab(x, y, \theta) = e^{-\left[\frac{C_x^2}{2\sigma_x^2} + \frac{C_y^2}{2\sigma_y^2}\right]} \cos(2\pi k C_x) \quad (S9)$$

where

$$\begin{aligned} C_x &= x \cos \theta + y \sin \theta; \\ C_y &= y \cos \theta - x \sin \theta; \end{aligned}$$

The parameters σ_x and σ_y determine the size (or extent) of the receptive field for each cortical cell in the horizontal and vertical dimensions respectively. The anisotropy of the receptive fields is controlled by the parameter $\gamma = \frac{\sigma_y^2}{\sigma_x^2}$. The parameter k determines the preferred spatial frequency of the receptive field for each cortical cell. The receptive fields of all cortical cells are centered at the same position in space; they differ only by their orientation θ . Positive regions of the Gabor function correspond to ON subfields; negative regions correspond to OFF subfields. Each cortical cell receives connections from all the LGN cells within a subfield boundary, with ON-subfields yielding connections from all ON-center LGN cells and OFF subfields yielding connections from all OFF-center LGN cells. The strength of each connection was set to $\alpha * |gab(x, y, \theta)|^2$ where α represents a uniform gain or amplification parameter. The baseline parameters (σ_x , σ_y , k and α) for the thalamo-cortical connectivity were set as $\sigma_x = 0.36$, $\sigma_y = 0.2$, $k = 0.7$, and $\alpha = 0.7$. These parameters were changed in the conditions that involved changes to thalamo-cortical feed forward connectivity as a model of perceptual learning (see **Section IV**).

VI – Lateral Connections: Units in the cortical layer are modeled as spiking units that are coupled to each other through lateral connections. In previous models of orientation selectivity^{3, 4}, the cortical layer consisted of both excitatory and inhibitory neurons, reflecting cortical layers *in vivo*. In such models, a cortical cell received lateral connections from both excitatory and inhibitory neurons and the net polarity of recurrent connectivity, to a particular cell, was determined by the proportion and strength of lateral

connections from excitatory neurons, relative to lateral connections from inhibitory neurons. In contrast to these previous models, we only include excitatory cortical cells in the cortical layer of our model. However, we are nevertheless able to implement the full range of excitatory and inhibitory lateral connectivity to a particular cell by allowing the strength of a recurrent connection between two cells to be either positive or negative (in previous models, and *in vivo*, the strength of a connection between two cells could only be positive). This allows us to model both excitatory and inhibitory lateral connections without including inhibitory neurons in the cortical layer.

We implement full lateral connectivity – every cell is coupled with every other cell in the cortical layer. We model all the lateral connections as being inhibitory in polarity by making the baseline connection strength significantly negative. However, the pattern of connection strengths between neurons versus the difference in their preferred orientations is chosen so that the connection strengths form a “Mexican hat” function^{3, 5}, relative to baseline. This implies that:

1. A cell is connected to other cells with similar preferred orientations, with connection strengths that are less inhibitory than baseline.
2. A cell is connected to other cells with significantly dissimilar preferred orientations, with connection strengths that are more inhibitory than baseline.

The polarity of the lateral connection weights, relative to the baseline weight, from a particular cell to all other cells is described mathematically as a difference between two Von-Mises functions of preferred orientations. The width of one function represents the

width of the short-range excitatory (relative to baseline) connection weights and the width of the other function represents the width of the long-range inhibitory (relative to baseline) connection weights.

Formally, the strength of the lateral connection between two cells x and y , with preferred orientations P_x and P_y (in radians) can be written as follows:

$$W(x, y) = \frac{G_w}{N_{out}} \left[e^{K_e(\cos(P_y - P_x) - 1)} - A_i e^{K_i(\cos(P_y - P_x) - 1)} \right] + DC_w \quad (S10)$$

The strength of a lateral connection between two neurons is described by five parameters:

1. The width of the excitatory (relative to baseline) connections is represented by the Von-Mises concentration parameter K_e
2. The width of the inhibitory (relative to baseline) connections is represented by the Von-Mises concentration parameter K_i
3. The overall baseline weight is represented by DC_w . As we mentioned previously, this value is set to be significantly negative (inhibitory).
4. The overall gain or amplification factor applied to the lateral connection weights is represented by G_w . It should be noted that G_w is scaled by N_{out} which represents the total number of units in the cortical layer thereby making $\frac{G_w}{N_{out}}$ the effective amplification factor.

5. The relative amplitude of the inhibitory Von-Mises distribution, with respect to the excitatory Von-Mises distribution is represented by A_i . For example, when $A_i=1$, both the distributions have the same amplitude.

The baseline parameters used were: $K_e=1$, $K_i=0.5$, $DC_w=-1.0$, $G_w=100$ and $A_i=0.4$.

These parameters were changed in the conditions that involved changes to cortical lateral connectivity, as a model of perceptual learning (see **Section IV**).

VI – Neural properties: The cortical layer contains 256 neurons which are modeled as Linear Non-Linear Poisson (LNP) neurons. LNP neurons represent a mathematical description that provides a very good model of neural data⁶⁻⁸ – they represent a good model of integrate and fire neurons in the physiologically realistic high-noise limit. Furthermore, LNP neurons have the advantage of being analytically tractable thereby allowing for good analytical descriptions of neural behavior.

Each cortical cell is modeled as a single point process and the input-output relationship in the cell is composed of three distinct operations.

1. **Linear step:** LNP neurons linearly combine their input spike trains, both feed-forward and recurrent, to obtain a “membrane potential proxy” $u(t)$. The membrane potential proxy $u_i(t)$, for neuron i , is given by:

$$\tau \frac{du_i(t)}{dt} = -u_i(t) + \left[\sum_j M_{ij} h_j(t) + \sum_j W_{ij} r_j(t) \right] \quad (\text{S11})$$

In this expression, τ represents the membrane time constant which was set to 20 msec, M represents the matrix of feed-forward connection weights from the LGN to the cell (described earlier), W represents the matrix of lateral connection weights from all other cortical cells to the cell (described earlier), $h_j(t)$ represents the input spikes at time t from the j^{th} pre-synaptic LGN cell and $r_j(t)$ represents the input spikes at time t from the j^{th} pre-synaptic cortical cell.

2. **Non-linear step:** The “membrane potential proxy” $u_i(t)$ is then passed through a static non-linearity $g(u)$,

$$g(u_i(t)) = \frac{\mu}{\lambda} \ln(1 + e^{\lambda(u_i(t) - \theta)}) \quad (\text{S12})$$

We used $\mu = 1.0$, $\lambda = 0.07$ and $\theta = 50.0$.

This in turn gives the instantaneous probability that the neuron emits a spike.

$$\rho_i(t | u_i(t)) = g(u_i(t)) \quad (\text{S13})$$

3. **Poisson step:** Finally, a Poisson process is instantiated with the instantaneous spike probability, leading to a variable number of spikes being generated by the neuron at a given time step. The variability of the generated spikes is thus guaranteed to be of the Poisson form which is a good description of biological

neural variability. The spikes emitted by the neuron in a small time interval dt is given by:

$$r_i(t) = \text{Poisson}[dt * \rho_i(t | u_i(t))] \quad (\text{S14})$$

These spikes are in turn transmitted to all the other cortical cells through lateral connections thereby influencing the post-synaptic cells' membrane potential proxies at the next time step.

V1 – Single-cell Response Properties: The response properties of individual cells in the cortical layer of our network closely match the response properties of V1 neurons *in vivo*. First, since we model cortical cells as LNP units, the variability of generated spikes is guaranteed to be of the Poisson form which is a good description of neural variability *in vivo*. Second, the response of cortical cells in our network shows the characteristic contrast-invariance that has been reliably demonstrated in orientation selective cells in the primary visual cortex⁹⁻¹¹. Specifically, as shown in Figure 2b in the main text, the amplitude of the response of V1 cells in our network increases with an increase in signal contrast but the width of the response does not change – in line with responses observed *in vivo*. Furthermore, *in vivo* the amplitude of the response of orientation selective cells in the primary visual cortex has been shown to be logarithmically related to linear changes in signal contrast¹⁰. The cortical cells in our network reliably show this behavior in a manner that closely mimics realistic cortical neurons.

Decoder: The final stage of the network involves connections from all the cortical cells to a single decision unit that outputs an estimate for the orientation of the stimulus (right or left of center, for example). This stage thus represents the decoding step in which the V1 activities are translated into a task-relevant estimate. The overarching goal in this study was to show that early changes in network properties can lead to the uniform shift of TVC curves that is observed during perceptual learning. Accordingly, in simulating the effects of perceptual learning, we kept the connection weights between the cortical cells and the decision unit constant. This was to ensure that any shift in the TVC curve we observed by making early changes in our network, could in fact be linked directly to those early changes. This was particularly critical given that previous work¹² has already shown that changes to the decision weights (the weights from V1 onto the decision unit) can shift the TVC curve uniformly. Had we allowed the weights to change anywhere, and found a uniform shift, we would have been unable to determine whether the uniform shift was the result of an early change (such as a change in the LGN-V1 weights) as opposed to being the result of a change in the decision weights.

We chose to use a linear classifier as the decoder in our network. Formally, the pattern of connection weights W_{dec} from the cortical cells to the decision unit can be written as:

$$W_{dec} = \Sigma^{-1}(s) f'(s) \tag{S15}$$

where $f'(s)$ represents the derivative of the tuning curve of the cortical cells. The tuning curve is a concise representation of the response properties of cortical cells and represents

the mean activation of the cell in response to a stimulus s . $\Sigma^{-1}(s)$ represents the inverse of the cortical covariance matrix.

We optimized the linear classifier for the pre-learning network condition – with the parameters set to the baseline values for each of the layers (described in the sections above) – and with the signal and noise contrast of the stimulus set to an arbitrary value within the range of values used in Doshier & Lu’s experiments. We then kept the classifier weights constant across all of our network and experimental manipulations. It was important to use the optimal weights to read out the activity in V1 for the pre-learning network condition, because had we used suboptimal weights, it would have been difficult to pinpoint the basis for any improvement in network performance. For instance, there might have been a way to adjust the LGN-V1 weights such that the Fisher information in V1 does not increase, but such that the performance of the network increases because the sensory representation in V1 is read out more efficiently by the decision weights. This would be a repeat of what Lu and Doshier have already published. With the method we used, on the other hand, the shift in the TVC curves observed in our network can be related solely to the increase of information in the V1 layer.

III. Computing discrimination performance

Relating neural responses to discrimination performance: In **Section II**, we described the architecture and response properties of our network model of orientation discrimination. We now describe the procedure by which we compute orientation discrimination performance in our network, when presented with the noise-injected stimuli described in

Section I. One standard way to relate neural response properties to discrimination performance is to consider the information-theoretic quantity of Fisher information which directly predicts performance in discriminations tasks. Fisher information is an upper bound on performance in discrimination tasks as it is proportional to the square of the discrimination threshold of an ideal observer of neural activity. Therefore, the greater the Fisher information, the better the performance.

We have recently derived an analytic expression for the Fisher information in a population of LNP neurons driven to a noise-perturbed steady state, with network properties similar to those of the network used here¹³. This expression takes advantage of the analytic tractability of LNP neurons and is able to consider the influence of network correlations, which are present in biologically realistic networks, on Fisher information. This expression can be written as follows:

$$I = (Mh')^T [M\Gamma_{hh}M^T + D^{-1}GD^{-1}]^{-1} (Mh') \quad (\text{S16})$$

The above expression contains several terms each of which map onto specific network and response properties of our network. They are:

1. **M** represents the matrix of thalamo-cortical feed-forward connections. As described earlier, this depends on four network parameters – σ_x , σ_y , k and α .
2. **h** represents the mean input firing rates from the thalamus. It is directly dependent on the percent signal contrast c of the stimulus (described in **Section I**).

3. Γ_{hh} represents the covariance matrix of the input firing rates from the thalamus.
The input covariance is a combination of the variance of the external noise in the stimulus and the variance resulting from the Poisson spiking of the LGN cells.
4. \mathbf{G} is a diagonal matrix whose entries give the mean response of the LNP neurons.
It is obtained from the non-linear function which transforms the “membrane potential proxy” $u(t)$ to the firing rate for the LNP neurons.
5. \mathbf{D} is a diagonal matrix which gives the derivative or slope of the activation function \mathbf{G} .

Readers are referred to¹³ for an in-depth description of the techniques used to derive the above expression, including the precise assumptions that were made. It is important to note that the expression in (S16) derives the linear Fisher information in a population of LNP neurons with realistic variability. Linear Fisher information is the fraction of Fisher information that can be recovered by a locally optimal linear estimator. In practice, linear Fisher information has been found to provide a tight bound on total Fisher information, both in simulations⁴ and in vivo¹⁴. Consequently, in the rest of this document, we consider the linear Fisher information in our network and any reference to Fisher information refers to the linear Fisher information. Another point to note is that the original derivation of linear Fisher information in¹³ was based on considering a discrimination task in which the difference in stimulus orientation was small. In our case, we are concerned with a difference in orientation of ± 12 degrees which is not a small difference. However, linear Fisher Information can still be computed using the analytic

expression by considering the difference between the mean activities induced by the two stimuli and the average covariance induced by the two stimuli.

Another point that should be noted is that the analytic expression as it is written in equation (S16) is based on the assumption that the decoder that is used to read out the activity of the cortical layer is the optimal decoder for the specific network. This implies that a change in network parameters, across learning for instance, might lead to a change in the optimal decoder. However, in this study, we are interested in highlighting the changes in early response properties that could lead to the observed behavioral improvements, while keeping the decoder constant. Thus, based on prior work by Wu et al (2001)¹⁵, we derived another form of the expression in (S16) above, for the Fisher information in our network using a fixed decoder, which looks as follows:

$$I(W_{dec}) = \frac{(W_{dec}^T \mu')^2}{W_{dec}^T \Gamma W_{dec}} \quad (\text{S17})$$

where

$$\begin{aligned} \mu' &= (D^{-1} - W)^{-1} M h' \\ \Gamma &= (D^{-1} - W)^{-1T} [M \Gamma_{hh} M^T + D^{-1} G D^{-1}] (D^{-1} - W)^{-1} \end{aligned}$$

W_{dec} represents the pattern of connection weights from the cortical layer to the decision stage, i.e. the fixed decoder described earlier and W represents the matrix of cortical recurrent connections which, as described earlier, depends on five network parameters – K_e , K_i , DC_w , G_w and A_i .

Deriving TVC curves: In Doshier & Lu's studies, given a level of external noise contrast, a staircase method was used to determine the level of signal contrast that was needed to elicit a criterion level of accuracy. This process allowed the experimenters to compute each subject's discrimination threshold for a particular level of external noise contrast, given a criterion level. By repeating the staircase procedure at multiple levels of external noise contrast for the same criterion level, they were then able to derive a curve of discrimination threshold as a function of external noise contrast known as a threshold-versus contrast or TVC curve. This curve depicts the change in discrimination threshold, for a given criterion level, as a function of the change in external noise contrast. In order to be able to relate the results of our network to the results of Doshier & Lu, we need to be able to generate TVC curves of our network's performance similar to the TVC curves of their subjects' performance.

The analytic expression in equation (S17) allows us to compute the Fisher information in the network at the decision stage, when presented with a stimulus generated using a particular signal and noise contrast. In order to then derive TVC curves of our network's performance, using this analytic expression, we used the following approach:

1. We computed the information at the decision stage (using eq.(S17)) when the network was presented with stimuli including a range of signal contrasts and a particular external noise contrast. We used 15 signal contrast levels. They were: 1.25%, 1.5%, 2%, 2.5%, 3%, 3.5%, 4%, 5%, 6%, 7%, 8%, 10%, 12%, 14% and 16%.

2. We repeated this process with the eight levels of external noise contrast used by Doshier & Lu.
3. As a result of steps 1 and 2, we generated a matrix which represents the information at the decision stage as a function of signal contrast and external noise contrast. As we move across columns of this matrix we have the change in information as a function of external noise contrast and as we move across rows we have the change in information as a function of signal contrast.
4. Since the criterion accuracy level used by Doshier & Lu was in terms of percent correct performance and our network's performance was in terms of Fisher information, we derived the Fisher information that was equivalent to the percent correct criterion used by Doshier & Lu using techniques from classical signal detection theory¹⁶. For example, the 79.3% correct criterion used by them corresponds to a Fisher information value of 0.0046 deg^{-2} for the stimuli used.
5. Finally, we computed the iso-information contour through the matrix of information, for the information matching the chosen criterion (0.0046 deg^{-2} in the above example). This gave us a curve of discrimination threshold as a function of external noise contrast, for the criterion accuracy level – a TVC curve.

By following the above approach we have the equivalent orientation discrimination performance curves from our network as Doshier & Lu had for their human subjects. We can now make changes to the network parameters and measure the changes to the TVC curves in an effort to model the neural basis of perceptual learning. The test will be

whether early network changes lead to similar changes in the TVC curves as the changes observed by Doshier & Lu, due to perceptual learning by their subjects.

IV. Network models tested

Parametric study: In the preceding sections, we have described the manner in which we replicated the external-noise inclusion experiments of Doshier and Lu, our network model for orientation discrimination and the analytic approach that we used to compute the orientation discrimination performance in a given network, when presented with the noise-injected stimuli of Doshier & Lu. We now discuss the specific changes to network parameters that we simulated in order to test our hypothesis that the behavioral improvements observed during perceptual learning can be obtained through changes in the neural representations in early sensory areas. We considered two possible mechanisms through which neural representations could be changed in early sensory areas – through changes to cortical lateral connections and through changes to the thalamo-cortical feed forward connections.

1. **Changes to cortical lateral connections:** As described in **Section II**, the cortical lateral connections in our network model were defined by five parameters – K_e , K_i , DC_w , G_w and A_i . The baseline values for these parameters were set to $K_e=1$, $K_i=0.5$, $DC_w=-1.0$, $G_w=100$ and $A_i=0.4$. To test the effect of changing cortical lateral connections on the behavioral performance of the system, we modified each of these parameters across a range of values and observed the

resultant changes in TVC curves. Specifically, $K_e \in (0.05, 4)$, $K_i \in (0.05, 4)$, $DC_w \in (-2.5, 0)$, $G_w \in (0.1, 200)$ and $A_i \in (0, 1.5)$.

2. Changes to thalamo-cortical feed forward connections: As described in **Section II**, the thalamo-cortical feed forward connections in our network model were defined by four parameters $-\sigma_x$, σ_y , k and α . The baseline values for these parameters were set to $\sigma_x = 0.36$, $\sigma_y = 0.2$, $k = 0.7$, and $\alpha = 0.7$. To test the effect of changing thalamo-cortical feed forward connections on the behavioral performance of the system, we modified each of these parameters across a range of values and observed the resultant changes in TVC curves. Specifically, $\sigma_x \in (0.2, 0.4)$, $\sigma_y \in (0.2, 0.4)$, $k \in (0.5, 1.25)$ and $\alpha \in (0.2, 2.0)$.

Modeling Perceptual Learning: As discussed in the main text (see Results section and Fig. 3), we were able to replicate the full range of observed behavioral changes, by only making changes to the thalamo-cortical feed forward connections, in our model. The specific parameter values that led to the reported results (Fig. 3) are as follows:

Supplementary Table 1: Parameters for networks shown in Fig. 3 (in the main text)

	σ_x	σ_y	k	α	K_e	K_i	DC_w	G_w	A_i
Before Learning (baseline)	0.36	0.2	0.7	0.7	1	0.5	-1.0	100	0.4
Training Session 1	0.36	0.23	0.67	0.6	1	0.5	-1.0	100	0.4
Training Session 2	0.36	0.27	0.62	0.5	1	0.5	-1.0	100	0.4

Exploring the role of amplification and sharpening: In addition to a uniform shift in TVC curves, changing the thalamo-cortical feed forward connections also led to a modest amount of amplification and sharpening in cortical tuning curves (Fig. 3b in the main text). We simulated several network models which allowed us to demonstrate that amplification and sharpening were neither necessary nor sufficient for the observed shift in TVC curves. The parameters for the models used in these demonstrations (Fig. 4 in the main text) are as follows:

Supplementary Table 2: Parameters for networks shown in Fig. 4 (in the main text)

	σ_x	σ_y	k	α	K_e	K_i	DC_w	G_w	A_i
Amp. not sufficient									
Before Learning	0.4	0.3	0.75	0.5	2	1	-1.5	1	0.75
After Learning	0.29	0.33	1.00	1.1	2	1	-1.5	1	0.75
Amp. not necessary									
Before Learning	0.26	0.29	1.25	1.6	2	1	-1.5	1	0.75
After Learning	0.4	0.3	0.75	0.5	2	1	-1.5	1	0.75
Sharp. not sufficient									
Before Learning	0.4	0.3	0.75	0.5	2	1	-1.5	1	0.75
After Learning	0.29	0.33	1.00	1.1	2	1	-1.5	1	0.75
Sharp. not necessary									
Before Learning	0.32	0.3	0.75	1.0	0.1	0.75	-1.2	40	1.0
After Learning	0.32	0.4	0.6	0.63	0.05	0.4	-1.0	60	1.3

Quantifying the contribution of correlations: In addition to demonstrating that learning need not depend on tuning curve changes like amplification and sharpening, we also quantified the improvement in performance we would have observed in the network shown in Figure 3 in the main text, if we had not seen any tuning curve changes across

training. In other words, we quantified the contribution of only changing the inter-neuronal correlations to the overall improvement in performance that was observed in our simulations. To do this, we used the analytic relationship in (S17) to compute the contribution of a change in correlations to the training induced increase in Fisher information. Specifically, we compared the increase in information across training sessions for the network shown in Figure 3, against the change in information in a virtual population of neurons with the same change in correlations but with identical tuning curves across training sessions. We call this population ‘virtual’ because it is not clear that such spike statistics could be generated by an actual neural network since tuning curves and correlations cannot easily be decoupled in our model. Nonetheless, we can still compute Fisher information for such spike statistics.

This method was inspired by a metric that was recently proposed by Cohen and Maunsell¹⁷. However, in the case of Cohen and Maunsell, their metric was based on an approximate computation of the Fisher Information before and after learning. Here, since we have access to the analytic relationship between network parameters and the true Fisher Information, we were able to quantify the contribution of a change in correlations to the training-induced increase in the true Fisher Information.

Supplementary References

1. Doshier, B.A. & Lu, Z.L. Mechanisms of perceptual learning. *Vision Research* **39**, 3197-3221 (1999).
2. Doshier, B.A. & Lu, Z.L. 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**, 13988-13993 (1998).
3. Somers, D.C., Nelson, S.B. & Sur, M. An emergent model of orientation selectivity in cat visual cortical simple cells. *J. Neurosci.* **15**, 5448-5465 (1995).
4. Series, P., Latham, P.E. & Pouget, A. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nature Neuroscience* **7**, 1129-1135 (2004).
5. Sompolinsky, H. & Shapley, R. New perspectives on the mechanisms for orientation selectivity. *Current Opinion in Neurobiology* **7**, 514–522 (1997).
6. Gerstner, W. & Kistler, W. *Spiking Neuron Models: An Introduction* (Cambridge University Press New York, NY, USA, 2002).
7. Paninski, L. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems* **15**, 243–262 (2004).
8. Plesser, H.E. & Gerstner, W. Noise in integrate-and-fire neurons: from stochastic input to escape rates. *Neural Computation* **12**, 367–384 (2000).
9. Ferster, D. & Miller, K.D. Neural Mechanisms of Orientation Selectivity in the Visual Cortex. *Annual Reviews in Neuroscience* **23**, 441-471 (2000).

10. Sclar, G. & Freeman, R.D. Orientation selectivity in the cat's striate cortex is invariant with stimulus contrast. *Experimental Brain Research* **46**, 457-461 (1982).
11. Troyer, T.W., Krukowski, A.E., Priebe, N.J. & Miller, K.D. Contrast-Invariant Orientation Tuning in Cat Visual Cortex: Thalamocortical Input Tuning and Correlation-Based Intracortical Connectivity. *J. Neurosci.* **18**, 5908-5927 (1998).
12. Lu, Z.-L., Liu, J. & Doshier, B.A. Modeling mechanisms of perceptual learning with augmented Hebbian re-weighting. *Vision Research* **50**, 375-390 (2010).
13. Beck, J.M., Bejjanki, V.R. & Pouget, A. Insights from a simple expression for linear Fisher information in a recurrently connected population of spiking neurons. *Neural Computation* (In Press).
14. Averbach, B.B., Latham, P.E. & Pouget, A. Neural correlations, population coding and computation. *Nat Rev Neurosci* **7**, 358-366 (2006).
15. Wu, S., Nakahara, H. & Amari, S.I. Population coding with correlation and an unfaithful model. *Neural Computation* **13**, 775-797 (2001).
16. Green, D.M. & Swets, J.A. *Signal detection theory and psychophysics* (John Wiley and Sons, Los Altos, California, USA, 1966).
17. Cohen, M.R. & Maunsell, J.H.R. Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience* **12**, 1594-1600 (2009).