Title Page

Title

Sub-optimality of the early visual system explained through biologically plausible plasticity

Authors

Tushar Chauhan[*1,2], Timothée Masquelier[1,2], Benoit R. Cottereau[1,2]

**Corresponding author*

Tushar Chauhan: Tushar.Chauhan@cnrs.fr

Affiliations

1. Centre de Recherche Cerveau et Cognition, Université de Toulouse, 31052 Toulouse, France
2. Centre National de la Recherche Scientifique, 31055 Toulouse, France

Abstract

The early visual cortex is the site of crucial pre-processing for more complex, biologically relevant computations that drive our perception and, ultimately, behaviour. In literature, this pre-processing is often viewed as an optimisation which enables the most efficient (in terms of information, spikes, energy, etc.) representation of the visual input. However, measurements in monkey and cat suggest that receptive fields in the primary visual cortex are often noisy, blobby, and symmetrical, making them sub-optimal for operations such as edge-detection. Here, we propose that this suboptimality occurs because the receptive fields do not emerge through a global minimisation of the generative error, but through locally operating biological mechanisms such as spike-timing dependent plasticity. Furthermore, using an orientation discrimination paradigm we show that while sub-optimal, such models offer a much better description of biology, both at a physiological and a perceptual level.

Introduction

The human visual system processes an enormous throughput of sensory data in successively convergent operations to generate percepts and behaviours necessary for biological functioning (Anderson, Van Essen, & Olshausen, 2005; Raichle, 2010). Due to the exclusively feed-forward (bottom-up) nature of thalamocortical inputs to the granular layer of area V1, computations in the early visual cortex are often explained through unsupervised normative models which, given an input dataset with statistics similar to those found in our surrounding space, carry out a global optimisation of criteria such as energy consumption and information-theoretic efficiency (Bell & Sejnowski, 1997; Bruce, Rahman, & Carrier, 2016; Hoyer & Hyvärinen, 2000; Olshausen & Field, 1996; van Hateren & van der Schaaf, 1998; Zhaoping, 2006). While such arguments do explain why many properties of the early visual system are closely related to characteristics of natural scenes (Bell & Sejnowski, 1997; Beyeler, Rounds, Carlson, Dutt, & Krichmar, 2019; Geisler, 2008; Hunter & Hibbard, 2015; Lee & Seung, 1999; Olshausen & Field, 1996), they are not equipped to answer questions such as how cortical structures which support complex computational operations implied by such optimisation may emerge, how these structures adapt, even in adulthood (Hübener & Bonhoeffer, 2014; Wandell & Smirnakis, 2010), and why some neurones possess receptive fields (RFs) which are sub-optimal (Jones & Palmer, 1987; Ringach, 2002).

It is now well established that locally-driven synaptic mechanisms such as spike-timing dependent plasticity (STDP) are natural processes which shape the computational architecture of the brain (Abbott, Song, & Miller, 2000; Bi & Poo, 1998; Caporale & Dan, 2008; Chauhan, Masquelier, Montlibert, & Cottereau, 2018; Gollisch & Meister, 2008; Masquelier, 2012). Therefore, it is only natural to hypothesise that locally operating, biologically plausible models of plasticity must offer a better description of the receptive fields in early visual cortex.

However, such line of reasoning leads to the obvious question: what exactly constitutes a ‘better description’ of a biological system, and more specifically, the early visual cortex. Here, we use a series of criteria spanning across electrophysiology, information theory, and machine learning, to investigate how descriptions of early visual RFs provided by a local, process-driven STDP model differ from two influential and widely employed normative schemes based on a generative paradigm – Independent component analysis (ICA), and Sparse coding (SC). Our results indicate that a process-based feedforward model of experience-driven plasticity is better suited to capturing the receptive-field (RF) diversity reported in the thalamocortical interface to V1 (Jones & Palmer, 1987; Ringach, 2002). Furthermore, using an orientation detection paradigm, we show that such a model is also consistent with orientation biases reported in human psychophysical studies – something that we do not observe for the normative schemes. While the information carrying capacity of STDP-driven local networks is not optimal when compared to generative schemes, our findings suggest that it is precisely this sub-optimality which makes process-based, local models ideal for describing the initial stages of sensory processing.

Methods

Dataset

The Hunter-Hibbard dataset of natural images (available under the MIT license at https://github.com/DavidWilliamHunter/Bivis) was used. It consists of 139 images of natural scenes, captured using a realistic acquisition geometry (Hunter & Hibbard, 2015) and a 20 degree field of view. In order to make the pre-processing comparable between the various encoding schemes, the images were resized to a resolution of 5 pixels/degree along both horizontal and vertical directions. During the learning phase, inputs to all encoding schemes were 3×3 degree patches (15×15 pixels) sampled randomly from the dataset.

Encoding models

Sparse Coding

We used the classical SC scheme based on L2-norm regularisation (Olshausen & Field, 1996, 1997). The pre-processing in this algorithm consists of an initial whitening of the input using low pass filtering, followed by a trimming of higher frequencies. The latter is employed to counter artefacts introduced by high frequency noise, and the effects of sampling across a uniform square grid. In the frequency domain the pre-processing filter is given by a zero-phase filter:

$$H(f) = f \cdot e^{-\left(\frac{f}{f_0}\right)^4}$$

Here, $f_0 = 10$ cycles/degree is the cut-off frequency. After the pre-processing, $3° \times 3°$ grids were sampled at random locations from random images in the dataset. These patches were used as inputs to the algorithm. Since we constrained each SC unit to have a $3° \times 3°$ RF at 5 pixels / degree, this led to a total of 225 units.

Due to the non-linear filtering, the retinal projections of the converged RFs could not be recovered by direct linear estimations. An approximate reverse-correlation algorithm derived from a linear stability analysis of the sparse coding objective function about its operating point (Ringach, 2002; Ringach & Shapley, 2004) was used. The RFs (denoted as columns of a matrix, say $\boldsymbol{\xi}$) are given by:

$$\boldsymbol{\xi} = \boldsymbol{A}[\boldsymbol{A}^T\boldsymbol{A} + \lambda S''(0)\boldsymbol{I}]^{-1}$$

Here, $\boldsymbol{A}$ is the matrix containing converged sparse components as column vectors, λ is the regularisation parameter (for the reconstruction, it is set to 0.14σ, where σ^2 is the variance in the input dataset), and $S(x)$ is the shape-function for the prior distribution of the sparse coefficients (this implementation uses $\log(1 + x^2)$).

Independent Component Analysis

If $\boldsymbol{X}$ is a matrix describing n samples of d-dimensional data, ICA-based models attempt to decompose $\boldsymbol{X}$ such that it can be interpreted as the activity $\boldsymbol{S}$ of $\tilde{d}$ independent units. This decomposition takes the form:

$$f(\boldsymbol{X}) = \boldsymbol{AS}$$

Here, $\boldsymbol{A}$ is the so-called mixing matrix. If $\boldsymbol{X}$ is assumed to describe n signals from the retina, f is often interpreted as the function which describes processing in the relaying thalamic circuitry. In this study we used a model (Hyvärinen & Oja, 2000) which implements f as a truncated principal component analysis (PCA) transform. In this model, the data is first centred ($\boldsymbol{X_C}$), and then projected to an LGN mapping:

$$\boldsymbol{L} = \boldsymbol{U^T X_C}$$

Here, $\boldsymbol{L}$ is interpreted as the activation of $\tilde{d}$ $(< d)$ LGN units, and the columns of $\boldsymbol{U}$ are the first $\tilde{d}$ principal components describing $\boldsymbol{X_C}$. This leads to low-pass filtering, and local decorrelation akin to centre-surround processing reported in the LGN. These LGN responses are then used to derive the activity of $\tilde{d}$ independent V1 units:

$$\boldsymbol{\Sigma} = \boldsymbol{WL}$$

Here, $\boldsymbol{W}$ is the un-mixing matrix and $\boldsymbol{\Sigma}$ is the activity of the V1 neurones. The recovery of the RFs for ICA is relatively straight forward, as, in our implementation, they are assumed to be equivalent to the filters which must be applied to a given input to generate V1 activity. The RFs (denoted as columns of a matrix, say $\boldsymbol{\xi}$) are given by:

$$\boldsymbol{\xi} = \boldsymbol{UW}^T$$

Spike timing dependent plasticity

We used a feedforward model based on an abstract spike-timing dependent plasticity (STDP) rule. The pre-processing in the model consisted of ON/OFF filtering using difference-of-Gaussian filters based on properties of magno-cellular LGN cells. The outputs of these filters were then converted to relative-latencies using a monotonically decreasing function(Masquelier & Thorpe, 2007; Rullen & Thorpe, 2001; Shriki et al., 2012) ($1/x$ was used), and only the earliest 10% spikes were allowed to propagate further. These spikes were used to train an unsupervised network of 225 (non-leaky) integrate-and-fire neurones endowed with a simplified version of the STDP rule proposed by Gütig, Aharonov, Rotter, & Sompolinsky (2003) given by:

$$\Delta w = \begin{cases} -\alpha^- \cdot w^{\mu^-}, & \Delta t \leq 0 \\ \alpha^+ \cdot (1-w)^{\mu^+}, & \Delta t > 0 \end{cases}$$

Here, Δt is the difference between the post- and pre-synaptic spike times. The weight w of a synapse was increased if a presynaptic spike occurred before the postsynaptic spike (causal firing), and decreased if the presynaptic neurone fired after the postsynaptic neurone (acausal firing). The constants $\alpha^{\pm}$ and $\mu^{\pm}$ describe the rate and non-linearity of the process respectively. During learning, the population followed a winner-take-all inhibition wherein the firing of one neurone reset the membrane potentials of all other neurones. After learning, this inhibition was no longer active and multiple units were allowed to fire for each input. The RFs of the converged neurones were recovered using a linear approximation. If w_i denotes the weight of the synapse connecting a given neurone to the ith LGN filter, the RF $\boldsymbol{\xi}$ of the neurone was given by:

$$\boldsymbol{\xi} = \sum_{i \in LGN} w_i \boldsymbol{\psi_i}$$

Here, $\boldsymbol{\psi_i}$ is the RF of the ith LGN neurone. The robustness of this model to the ratio of long-term potentiation (LTP) rate to long-term depression (LTD) rate, LGN noise, as well to external noise has been shown elsewhere (Chauhan et al., 2018). Furthermore, the results presented in this article were also found to hold for a purely additive network where the LTP and LTD nonlinearities are reduced to zero (i.e., $\mu^{\pm} = 0$).

Evaluation metrics

Gabor fitting

The linear approximations of RFs obtained by each encoding strategy were fitted using 2-D Gabor functions. This is motivated by the fact that all the encoding schemes considered here lead to linear, simple-cell-like RFs. A 2-D Gabor function $\boldsymbol{\zeta}$ is given by:

$$\boldsymbol{\zeta}(k, f, \theta, \phi, \boldsymbol{c}, \boldsymbol{\sigma}) = N_o + k \cdot \boldsymbol{S}(f, \theta, \phi) \cdot \boldsymbol{G}(\boldsymbol{c}, \boldsymbol{\sigma}, \theta)$$

Here, k is the sensitivity of the Gabor, $\boldsymbol{S}$ is a 2-D sinusoid propagating with frequency f and phase ϕ, at an angle θ to the positive x-axis, $\boldsymbol{G}$ is a 2-D Gaussian centred at $\boldsymbol{c}$, with size parameter $\boldsymbol{\sigma}$, also oriented at an angle θ, and N_o is a constant term. For each RF, the goodness-of-fit parameter (R^2) was also stored. This provides an intuitive measure of how Gabor-like a given RF is. The fitting was carried out using an adapted version of the code available at https://uk.mathworks.com/matlabcentral/fileexchange/60700-fit2dgabor-data-options, shared under an open-source license by Gerrit Ecke (University of Tübingen).

Frequency-normalised spread vector

The shape of the RFs approximated by each encoding strategy was characterised using frequency-normalised spread vectors (FSVs) (Chauhan et al., 2018; Ringach, 2002). For a RF fitted by a Gabor-function with sinusoid carrier f and envelope size $\boldsymbol{\sigma} = \begin{bmatrix} \sigma_x \\ \sigma_y \end{bmatrix}$, the FSV is given by:

$$\begin{bmatrix} n_x \\ n_y \end{bmatrix} = \begin{bmatrix} \sigma_x \\ \sigma_y \end{bmatrix} f$$

While n_x provides an intuition of the number of cycles in the RF, n_y is a cycle-adjusted measure of the elongation of the RF perpendicular to the direction of sinusoid propagation. The FSV serves as a compact, intuitive descriptor of the RF shape invariant to affine operations such as translation, rotation and isotropic scaling.

Dispersion of unit activations

The variability in the membrane potentials (activations) of the converged units in response to natural images was quantified using the coefficient of variation (CV) – a measure of the sample

dispersion. For a given neurone with response $\boldsymbol{r}$ (where each element of $\boldsymbol{r}$ denotes the activation in response to one input stimulus), it is given by:

$$CV = \frac{\sigma_{\boldsymbol{r}}}{\mu_{\boldsymbol{r}}}$$

Here, σ_r and μ_x are the standard-deviation and mean over the set of responses (elements of $\boldsymbol{r}$).

Sparsity: Gini Coefficient

The sparseness of the encoding was evaluated using the Gini coefficient (GC). GC is a measure which characterises the deviation of the population-response from a uniform distribution of activity across the samples. It is 0 if all units have the same response and tends to 1 as the responses become sparser (being equal to 1 if only one unit responds, while others are silent). It is invariant to the range of the responses within a given sample, and robust to variations in sample-size (Hurley & Rickard, 2009). Formally, the GC (denoted here as Λ) is given by:

$$\Lambda(\boldsymbol{x}) = 1 - 2\int_0^1 L(F)dF$$

Here, L is the Lorenz function defined on the cumulative probability distribution F of the neural activity (say, $\boldsymbol{x}$). For a discrete sample $\boldsymbol{x}$ with non-negative activity, this can be simplified as (Hurley & Rickard, 2009):

$$\Lambda(\boldsymbol{x}) = 1 - \frac{2}{N\|\boldsymbol{x}\|_1}\sum_{n=1}^{N}\left\{\left(N - n + \frac{1}{2}\right)\tilde{x}_n\right\}$$

Here, N is the number of neurones in the sample, $\|\cdot\|_1$ denotes the L1 norm, and $\tilde{\boldsymbol{x}}$ is the activity vector $\boldsymbol{x}$ sorted in ascending order. We defined two variants of the GC which measure the spatial (Λ_s) and temporal sparsity (Λ_t) of an ensemble of encoders (see Figure 3A for a schematic). Given a sequence of M inputs to an ensemble of N neurones, and denoting the activity of the

nth neurone in response to the mth input as x_m^n, the spatial sparsity of the ensemble response to the mth stimulus is given by:

$$\Lambda_s(m) = \Lambda(\{x_m^1, x_m^2, \dots, x_m^N\})$$

Similarly, the temporal sparsity of the nth neurone over the entire sequence of inputs is given by:

$$\Lambda_t(n) = \Lambda(\{x_1^n, x_2^n, \dots, x_M^n\})$$

Fisher Information

The information content in the activity of the converged units was quantified by using approximations of the Fisher Information (FI, denoted here by the symbol Φ). If $\boldsymbol{x} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_3}, \dots, \boldsymbol{x_N}\}$ is a random variable describing the activity of a system of N independent units, the FI of the population with respect to a parameter θ is given by:

$$\Phi(\theta) = \sum_{i=1}^{N} E\left[\left\{\frac{\partial \ln P(x_i|\theta)}{\partial \theta}\right\}^2\right]_{P(x_i|\theta)}$$

Here, $E[.]_{P(x_i|\theta)}$ denotes expectation value with respect to the firing-state probabilities of the ith neurone in response to the stimuli corresponding to parameter value θ. In our simulations, θ was the orientation (defined as the direction of travel) of a set of sine-wave gratings (SWGs) with additive Gaussian noise and random phase values between 0° and 360°. The ground-truth values of θ were sampled at intervals of 4° in the range $[0°, 180°)$. Here, we present the results of simulating SWGs with Gaussian noise at 3dB SNR. Since this approach allows for the activity of the entire population to be driven by the same set of ground-truth stimuli, it permits empirical estimates of the population FI without the need for the calculation of tuning curves for each neurone.

Decoding using a linear classifier

While the estimates of FI give us a rough measure of the information about the parameter contained in the population response, it does not necessarily follow that all the information is directly usable by downstream processes to draw inferences about the parameter value. To show this, we use a simple linear decoder on the population responses obtained in the FI analysis. The ground-truth values of the orientation θ were used as the class labels, and we present the accuracy of the decoder for each ground-truth label (the diagonal of the confusion matrix).

Post-convergence threshold variation in STDP

To test how post-learning changes in the threshold affect the specificity of a converged network, we tested an STDP network trained using a threshold Θ_{training} by increasing or decreasing its threshold (to say, $\Theta_{testing}$) and presenting it with SWGs (same stimuli as the ones used to calculate the FI). In this article, we report the results of seven simulations where the relative change in threshold was given by:

$$\frac{\Theta_{\text{testing}} - \Theta_{\text{training}}}{\Theta_{training}} = \{0, \pm 0.25, \pm 0.50, \pm 0.75\}$$

Code

The code for ICA was written using the sklearn library in python. The code for SC was based on the C++ and Matlab code generously made available by Prof. Bruno Olshaussen at www.rctn.org/bruno/sparsenet/. The STDP code will be made available on a public gitlab library.

Results

We trained a feed-forward STDP model with 225 integrate-and-fire neurones (Figure 1B) using 10^5 patches ($3° \times 3°$ visual angles) sampled from a database of realistic natural scenes (Figure 1A). The units, as expected (Delorme, Perrinet, & Thorpe, 2001; Masquelier, 2012), converged to oriented, edge-detector like receptive fields (RFs) (Figure 1C).

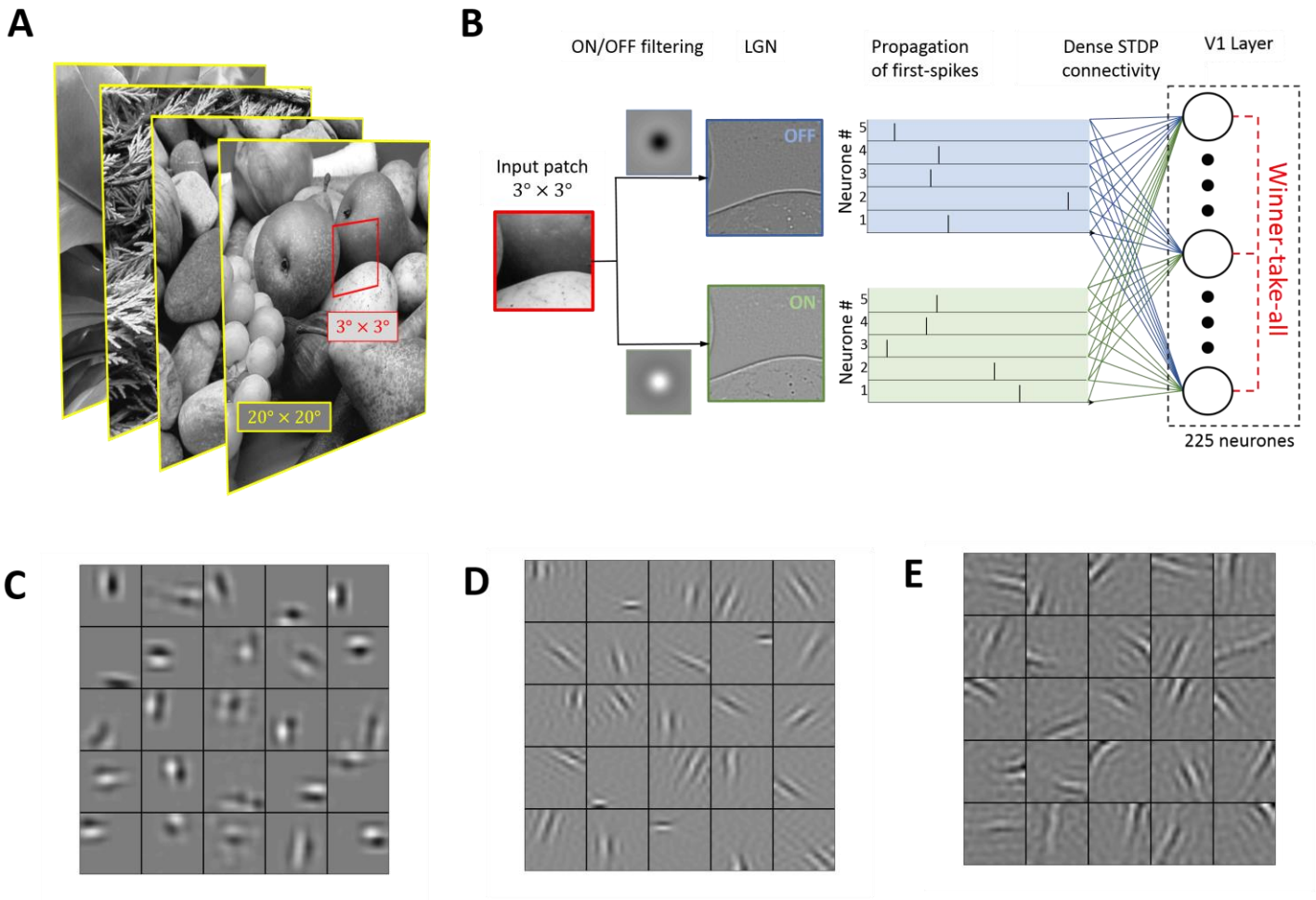

Figure 1: Emergence of Gabor-like receptive fields (RFs) **A.** Training data. The dataset (Hunter-Hibbard) consisted of natural images with a $20° \times 20°$ field of view. Patches of size $3° \times 3°$ were sampled from random locations from the images (overlap allowed). The same set of 100,000 randomly sampled patches was used to train all the models. **B.** STDP model pipeline. Each input patch was filtered using difference-of-Gaussian ON/OFF filters to generate LGN representations. First-spike ranks were obtained from the LGN representations using a

monotonically decreasing $1/x$ intensity to latency conversion (Rullen & Thorpe, 2001; Shriki, Kohn, & Shamir, 2012). A threshold was applied such that only the first 10% spikes were fed forward to a network of 225 neurones endowed with an abstract spike-timing dependent plasticity (STDP) rule. During training, a winner-take-all inhibition was implemented to ensure convergence. **C.** STDP RFs. RFs of randomly chosen neurones in the converged STDP population. **D.** ICA RFs. RFs of converged units from an Independent component analysis (ICA) model(Hoyer & Hyvärinen, 2000; van Hateren & van der Schaaf, 1998). **E.** SC RFs. RFs of converged units from a Sparse coding (SC) model(Olshausen & Field, 1996, 1997).

We also trained models based on two widely-employed normative theories of early visual computation – Independent Component Analysis (Hoyer & Hyvärinen, 2000; van Hateren & van der Schaaf, 1998) (ICA; Figure 1D), and Sparse coding (Olshausen & Field, 1996, 1997) (SC; Figure 1E). While the RFs from ICA and SC are elongated and highly directional, STDP RFs are more compact and less sharply tuned. This is in agreement with simple-cell recordings in the macaque and cat, where studies have reported that not all RFs are optimally tuned for edge-detection (Jones & Palmer, 1987; Ringach, 2002). A quantitative measure of this phenomenon can be obtained by fitting Gabor functions to the RFs and considering the frequency-normalised spread vectors or FSVs (Ringach, 2002) of the fit. The first component (n_x) of an FSV characterises the number of cycles in the RF and the second component (n_y) is a measure of the elongation of the RF perpendicular to carrier propagation. Most simple-cell RFs measured in macaque and cat tend to fall within the square: $\{|n_x| < 0.5, |n_y| < 0.5\}$. Since they are not elongated along either the direction of sinusoid propagation (they show few, typically 2-3 on/off regions) or perpendicular to it, they tend to be compact – making them less effective as edge-detectors compared to more crisply tuned elongated RFs. In Figure 2 we see that RFs of a considerable number of STDP neurones (33.8%) tend to fall in this realistic regime, while the ICA (10.9%) and SC (0%) neurones show a distinctive shift upwards and to the right (Ringach, 2002). Moreover, we find that this trend also holds for a purely additive

STDP network ($\mu^{\pm} = 0$, see Methods), and is not dependent on the nonlinearity of the plasticity rule.

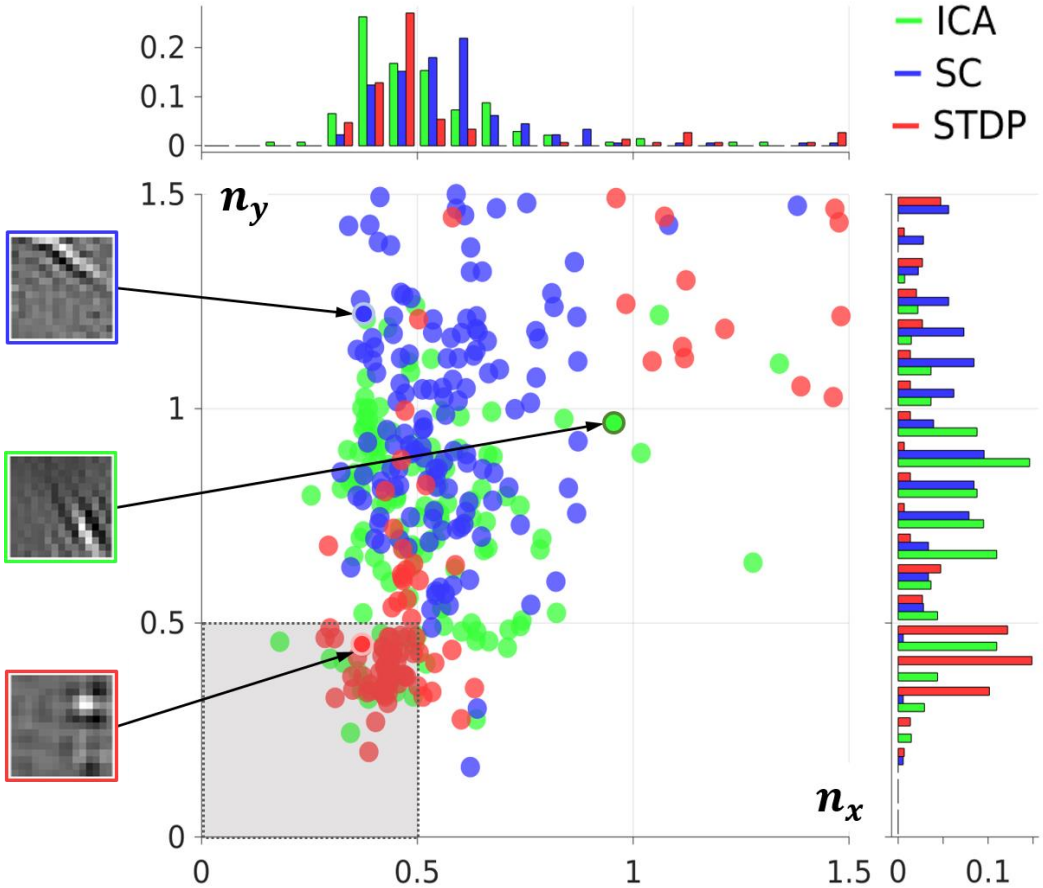

Figure 2: Frequency-normalised spread vectors (FSVs) of the converged RFs ($R^2_{Gabor\ fit} \geq 0.6$). The abscissa ($n_x$) is proportional to the number of cycles along the propagation of the sinusoid, while the ordinate (n_y) is a measure of the elongation of the RF perpendicular to the direction of the sinusoid. Measurements in macaque and cat primary visual areas indicate that simple-cells in these species tend to fall within the square bound by 0.5 along both axes (shown in the figure with a dotted outline and grey shading). The left margin shows three representative neurones chosen to illustrate elongated, multi-lobed and bloblike RFs. Both axes have been cut-off at 1.5 to clearly show the shaded region, and make comparison with biological data easier.

Given this sub-optimal nature of STDP RFs, we further investigated its implications on early visual processing using measures which span two important criteria employed by normative schemes – efficiency in terms of information transfer, and efficiency in terms of resources (spikes, energy etc.). In a typical generative formulation (such as ICA), information efficiency is implemented in terms of an error/objective function which, in some space, measures the information that is lost during encoding. The resource efficiency, on the other hand, is embedded in models (such as SC) using regularisation constraints which limit the parameter space explored by the optimisation process. In a first set of simulations, we presented the converged models with $3° \times 3°$ (visual angle) patches randomly sampled from the training dataset of natural images, and estimated the sparsity of the resulting activations using the Gini coefficient (Hurley & Rickard, 2009). The sparsity was examined in two contexts (Barth & Poulet, 2012) as shown in Figure 3A. First, sparsity of the entire ensemble was estimated for each presented stimulus – this is a measure of how many neurones, on average, are employed by the ensemble to encode a given stimulus. Second, the sparsity of individual neurones over the entire sequence of stimuli was estimated, allowing us to infer how frequently the features selected/encoded by the converged models occur in the sequence. We denote the former as spatial sparsity (Λ_s), and the latter as temporal sparsity (Λ_t). For STDP, the indices were calculated both for the membrane potential (which facilitates comparison with ICA and SC activations) and the thresholded spiking activity. STDP membrane potential (red, Figure 3B) shows high variability in Λ_s, whereas ICA (green) and SC (blue) show much lower variance in comparison. STDP spiking activity (pink), as expected, shows very high sparsity due to the thresholding. This suggests that while ICA and SC choose their features such that each image activates approximately equal number of units, the sparsity of the STDP neurones is stimulus-dependent, and likely driven by the relative probability of occurrence of specific features in the dataset.

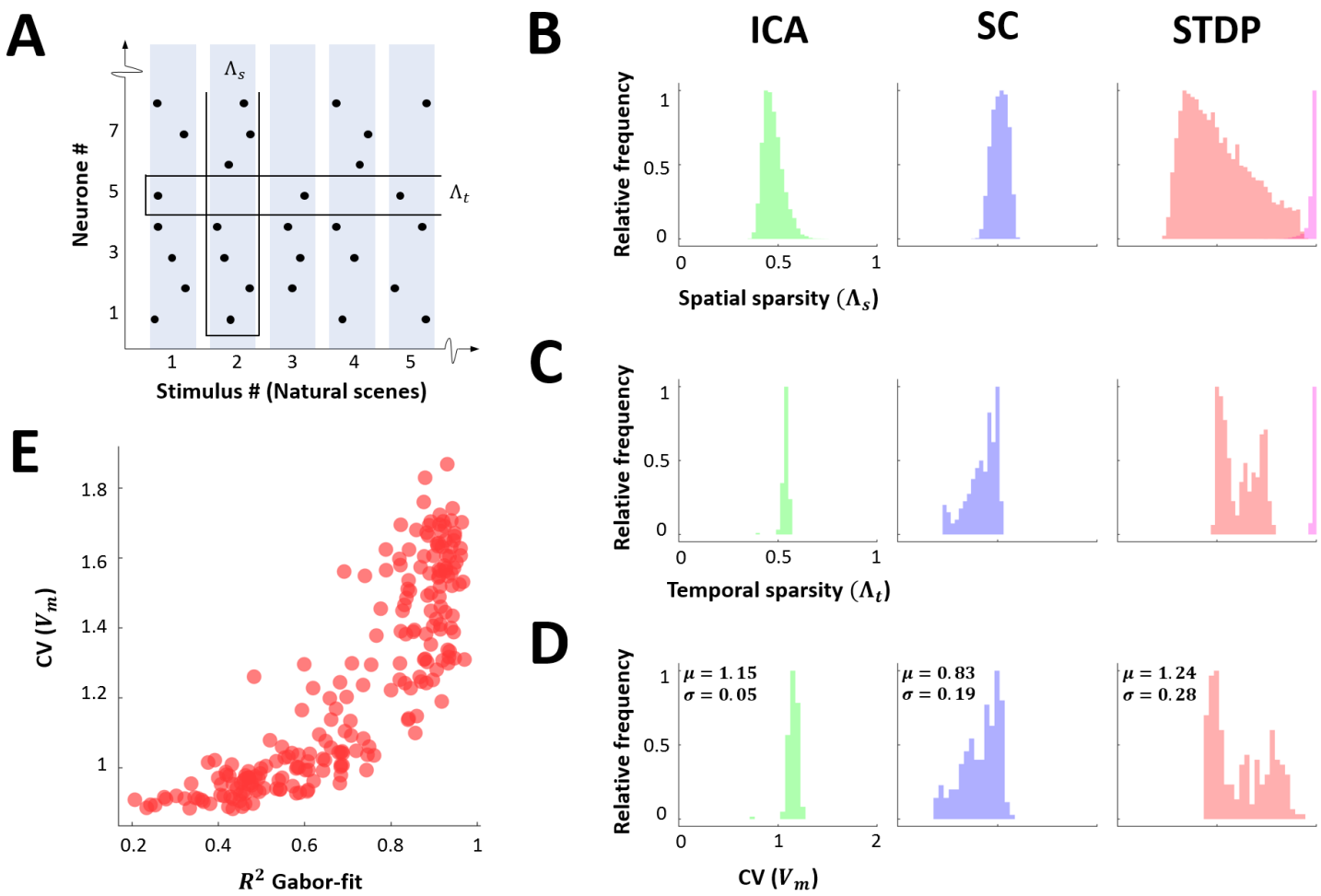

Figure 3: Sparsity indices. **A.** Spatial and Temporal Sparsity. To estimate measures of sparsity, 10^4 patches of size $3° \times 3°$ randomly sampled from natural scenes were presented to the converged models. Spatial sparsity Index (Λ_s) was defined as the average sparsity of the activity of the entire neuronal ensemble, while Temporal sparsity Index (Λ_t) was defined as the average sparsity of the activity of single neurones to the entire input sequence. **B.** Spatial sparsity Index. Histograms of Λ_s for the 10^4 input patches for converged units from ICA (left), Sparse coding (centre) and STDP (right). The same colour scheme as Figure 2 is used. For STDP, calculations using both membrane potentials (red) and spiking activity (pink) are shown. To show the distributions clearly, the histograms have been scaled relative to the maximum bin-count. Λ_s varied between 0 (all units activate with equal intensity) and 1 (very few neurones activate for any given input) by definition. **C.** Temporal sparsity index. Histograms of Λ_t, in a manner analogous to **B**. Λ_t also varies between 0 (homogeneous activity for the entire input sequence) and 1 (activity only for few inputs in the sequence). **D.** Coefficient of Variation (CV) of activations. The CV of

the activations of each neurone in response to the entire set of inputs (10^4 patches) was calculated. Each panel shows the histogram of the CV across the entire ICA, SC and STDP ensembles. The top-left corner shows the mean and standard deviation of the estimates. **E.** Variation of activation CV with Gabor goodness-of-fit in the STDP population. 2-D Gabor functions were fit to the RFs of STDP neurones. The R^2 goodness-of-fit parameter for each neurone (abscissa) is plotted against the CV of its activation (ordinate) in response to 10^4 input patches randomly sampled from natural images.

This high variability in STDP membrane potentials is also observed for temporal sparsity Λ_t (Figure 3C). Here, our model predicts two distinct groups of neurones corresponding to the two observed Λ_t modes. ICA and SC, as before, show much lower variability – indicating that they are optimised for features both within, and across patches. To further investigate the bimodality of Λ_t, we examined the coefficient of variation of the activations (membrane potential without thresholding for STDP neurones). We found that the activations for STDP neurones are bimodal and show much higher dispersion compared to ICA and SC neurones (Figure 3D, with mean and standard deviations shown on the upper left). Interestingly, this bimodality is correlated with the Gabor goodness-of-fit to the RFs (Figure 3E), suggesting that the sharper the RF tuning of an STDP neurone (to edges), the higher the dispersion of its subthreshold responses to natural stimuli.

Next, we investigated the direct consequences of STDP RF sub-optimality on the efficiency of information-encoding. To do this, we presented the converged network (trained on natural images) with a classical stimulus manipulated by electrophysiologists to characterise RFs in the early visual system – oriented SWGs (Bradley, Skottun, Ohzawa, Sclar, & Freeman, 1987; De Valois, Albrecht, & Thorell, 1982; Ohzawa & Freeman, 1986). To quantify the information content in neuronal responses, we estimated the Fisher information (FI) of the system. Using

the Crámer-Rao bound, this also allowed us to draw inferences about the theoretical limits of the optimal orientation discrimination performance (Wei & Stocker, 2017). Orientation discrimination tasks are especially suited to such an analysis as V1 activity has been shown to correlate with orientation discrimination thresholds (Vogels & Orban, 1990), something that is not true for other visual properties such as binocular disparity (Nienborg & Cumming, 2006) and motion (Grunewald, Bradley, & Andersen, 2002).

Figure 4A shows the inverse square-root of the FI plotted against the ground-truth value of the orientation parameter for each encoding scheme. In an encoding/decoding paradigm, this can be interpreted as maximal encoding efficiency of the model, assuming an optimal decoder. From a neuroscientific perspective, this local precision is often interpreted as the theoretically minimal just-noticeable-difference (JND) in a discrimination task (Wei & Stocker, 2017). We find that both ICA (green) and SC (blue) are much better at discriminating the orientations of the SWGs compared to STDP latencies (pink). When one only considers the activations of the STDP population (removing thresholding), the STDP neurones (red) carry much more information. Thus, in STDP neurones, spike-latencies are less informative than membrane-potentials – an expected result considering thresholding discards all sub-threshold information. But interestingly, this removal of information also leads to a strong modulation of the ordinate with orientation, with the cardinal directions producing the most informative/discriminable responses. This bias for cardinal directions is also reported in humans (see Discussion).

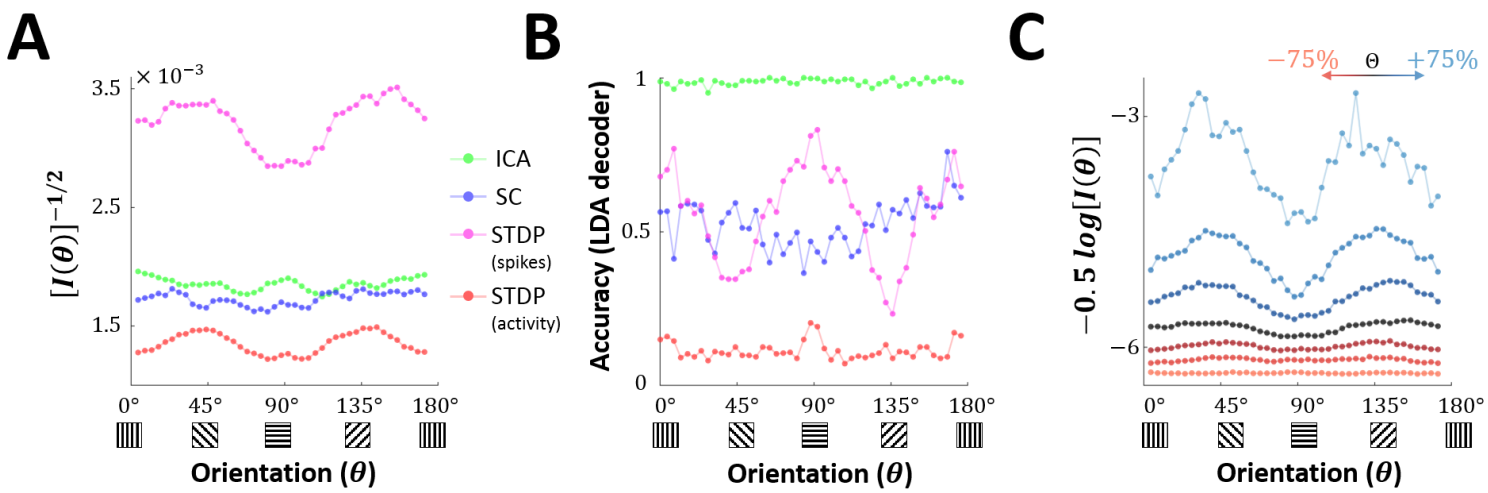

Figure 4: Fisher Information (FI) and Decoding accuracy. **A.** FI. Responses to sine wave gratings (SWGs) propagating at an angle θ to the horizontal were calculated (see Methods for details). These responses were then used to estimate the observed FI. The abscissa shows the ground-truth value of the parameter θ, and the ordinate shows the inverse square-root of the estimated FI. Due to the Cramér-Rao bound, the ordinate is proportional to the theoretical lower limit of the standard deviation of unbiased estimates of the parameter θ. Consequently, low values on the ordinate denote more precise encoding of the orientation, while higher values denote lower precision. Results for ICA are shown in green, SC in blue, and STDP in pink (spikes) and red (unthresholded activation) . **B.** Decoding accuracy. The stimuli and responses from **A** were used to train linear-discriminant classifiers. The ordinate shows the accuracy (probability of correct classification) for each ground-truth value of the orientation (abscissa). The same symbols as **A** are used. **C.** Threshold variation after convergence in the STDP population. The STDP population was trained on natural images using a threshold (say Θ), and subsequently, simulations were carried out by increasing (blue colours) or decreasing (red colours) the threshold of the converged population and testing the network using oriented SWGs (in a manner analogous to **A**). The abscissa of the graph shows the orientation of the SWGs, and the ordinate shows half the negative-log of the estimated FI (equivalent to the log of the ordinate in **A**, chosen here for clarity of presentation). The threshold was varied linearly in steps of 25% with respect to Θ ($\pm 25\%$, $\pm 50\%$, and $\pm 75\%$ were tested). The magnitude of change in the threshold is denoted by a corresponding change in the lightness of the colour from black (Θ) to clear blue (increase) or red (decrease).

Although FI gives us a mean of quantifying the information content in the neuronal activity, it does not automatically imply that this information is available for downstream processing (Quian Quiroga & Panzeri, 2009). To better quantify what this modulation of FI entails, we used linear-discriminant classifiers to evaluate how, in the simplest case, neural activity can be used to decode the orientation of SWGs. In Figure 4B we present the results of the decoding in terms of classification accuracy (ordinate) for each orientation label (abscissa). STDP latencies show an almost two-fold increase in decoding performance for cardinal directions. As one would expect from the FI curves in Figure 4A, this is not the case for ICA (green) and SC (blue) which allow an almost uniform decoding of SWG orientations. Surprisingly, despite a high information-throughput, STDP membrane potentials show very poor decoding accuracy (Figure 4A and B, red) – indicating that sub-threshold membrane potentials are noisy, and do not support stimulus decoding through simple linear mechanisms.

Discussion

Traditionally, process-based descriptions (often studied through mean-field statistics) have been used to model fine-grained neuronal dynamics (Harnack, Pelko, Chaillet, Chitour, & Rossum, 2015; Kang & Sompolinsky, 2001; Moreno-Bote et al., 2014), while more global, normative schemes are employed to predict population-level characteristics (Hoyer & Hyvärinen, 2000; Lee & Seung, 1999; Olshausen & Field, 1997; van Hateren & van der Schaaf, 1998). Detailed process-based models suffer from constraints imposed by computational complexity, prohibitively long execution times (for large networks), hardware that is geared towards synchronous processing, and non-availability of measured ground-truth values for key model parameters. On the other hand, most one-shot models can leverage faster computational libraries and architectures developed over decades, thereby leading to more efficient computation. Through this work, we argue that more abstract forms of process-based models, when used to answer specific questions, can still give a closer approximation of biological processes than normative schemes. Our model, despite important limitations such as the use of rank-based first-spike coding, lack of inhibitory cortico-cortical connections, and feedback from higher visual areas, is still able to describe a number of phenomena which are not predictable through normative schemes.

The converged RFs of our model are closer to those reported in electrophysiological measurements. A number of them display a characteristic departure from the optimal, sharply tuned edge-detectors predicted by ICA and SC (Figure 2). Moreover, we show that the STDP population is also suboptimal in terms of orientation encoding (Figure 4A). While ICA and SC are equally performant at discriminating all orientations, the STDP population shows a distinct bias for the cardinal directions with a two-fold increase in discrimination performance for horizontal and vertical SWGs. This bias for cardinal directions has also been observed in human

subjects (Li, Peterson, & Freeman, 2003; Orban, Vandenbussche, & Vogels, 1984), who also show an almost two-fold reduction in just-noticeable-differences for horizontal and vertical stimuli (Heeley & Timney, 1988). Our results suggest that such biases for cardinal directions (the so-called oblique effect) are supported by the cortical architecture as early as the thalamocortical interface – a claim that is reinforced by the findings of Furmanski & Engel (2000) who showed that neural activity in the primary visual cortex is directly correlated with the perceptual oblique effect. Thus, while being suboptimal in a generative sense, the early visual architecture is tuned for use in downstream cognitive processes and decision making.

We have also shown that the activations (unthresholded membrane potentials) of the STDP population show high dispersion in response to natural images (Figure 3D). An important consequence of this property is that even after the initial learning phase has ended, the network is still theoretically capable of shrinking or expanding its encoded dictionary without modifying its synaptic structure – simply by adapting its threshold (within the limits of the dispersion). This is not possible with the tested normative schemes since their activations show a very narrow range of dispersion. In the cortex, such changes are likely to be mediated by homeostatic processes (Harnack et al., 2015; Perrinet, 2019) and metaplastic synaptic regulation (Abraham, 2008; Cooke & Bear, 2010). In a set of simulations which investigate how post-convergence changes in activation thresholds affect the estimated orientation discrimination curves (Figure 4C), we found that the sensitivity to cardinal orientations can indeed be increased by increasing the post-convergence threshold (blue curves), and conversely, decreased by a decrease of the threshold (red curves). This suggests that purely bottom-up changes could indeed account for some perceptually relevant cortical changes in adults who are beyond the critical window of plasticity. This is particularly relevant for reports where low-level changes in the early visual

cortex have been shown to accompany perceptual learning (Bao, Yang, Rios, He, & Engel, 2010; C. S. Furmanski, Schluppeck, & Engel, 2004).

Finally, it must be noted that in this study we have used *classical* forms of the ICA and SC algorithms. More generally, these models are part of a class of data-adaptive encoding schemes which also comprises other algorithms such as PCA and nonnegative matrix factorisation. These algorithms rely on the efficiency of the encoding manifold at representing the training data, and their use in neuroscience is motivated by the proven (energetic and information-theoretic) efficiency of neural processes at representing natural statistics. A growing number of insightful studies now employ hybrid encoding schemes which address multiple optimisation criteria (Beyeler et al., 2019; Martinez-Garcia, Martinez, & Malo, 2017; Perrinet & Bednar, 2015), often through local process-based computation (Isomura & Toyoizumi, 2018; Savin, Joshi, & Triesch, 2010; Zylberberg, Murphy, & DeWeese, 2011).

By definition, process-based models offer a much deeper insight into the organisation and functioning of biological systems. With the common availability of increasingly faster and more adaptable computing solutions, we hope process-based modelling will be adopted more widely by cognitive and computational neuroscientists.

Bibliography

Abbott, L., Song, S., & Miller, K. (2000). Competitive Hebbian learning through spike-timing-dependent synaptic plasticity. *Nature Neuroscience*, *3*(9), 919–926. https://doi.org/10.1038/78829

Abraham, W. C. (2008). Metaplasticity: Tuning synapses and networks for plasticity. *Nature Reviews Neuroscience*, *9*(5), 387. https://doi.org/10.1038/nrn2356

Anderson, C. H., Van Essen, D. C., & Olshausen, B. A. (2005). CHAPTER 3—Directed Visual Attention and the Dynamic Control of Information Flow. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 11–17). https://doi.org/10.1016/B978-012375731-9/50007-0

Bao, M., Yang, L., Rios, C., He, B., & Engel, S. A. (2010). Perceptual learning increases the strength of the earliest signals in visual cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(45), 15080–15084. https://doi.org/10.1523/JNEUROSCI.5703-09.2010

Barth, A. L., & Poulet, J. F. A. (2012). Experimental evidence for sparse firing in the neocortex. *Trends in Neurosciences*, *35*(6), 345–355. https://doi.org/10.1016/j.tins.2012.03.008

Bell, A. J., & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338. https://doi.org/10.1016/S0042-6989(97)00121-1

Beyeler, M., Rounds, E. L., Carlson, K. D., Dutt, N., & Krichmar, J. L. (2019). Neural correlates of sparse coding and dimensionality reduction. *PLOS Computational Biology*, *15*(6), e1006908. https://doi.org/10.1371/journal.pcbi.1006908

Bi, G. Q., & Poo, M. M. (1998). Synaptic modifications in cultured hippocampal neurons: Dependence on spike timing, synaptic strength, and postsynaptic cell type. *The Journal of Neuroscience*, *18*(24), 10464–10472.

Bradley, A., Skottun, B. C., Ohzawa, I., Sclar, G., & Freeman, R. D. (1987). Visual orientation and spatial frequency discrimination: A comparison of single neurons and behavior. *Journal of Neurophysiology*, *57*(3), 755–772. https://doi.org/10.1152/jn.1987.57.3.755

Bruce, N., Rahman, S., & Carrier, D. (2016). Sparse coding in early visual representation: From specific properties to general principles. *Neurocomputing*, *171*, 1085–1098. https://doi.org/10.1016/j.neucom.2015.07.070

Caporale, N., & Dan, Y. (2008). Spike Timing–Dependent Plasticity: A Hebbian Learning Rule. *Annual Review of Neuroscience*, *31*(1), 25–46. https://doi.org/10.1146/annurev.neuro.31.060407.125639

Chauhan, T., Masquelier, T., Montlibert, A., & Cottereau, B. (2018). Emergence of binocular disparity selectivity through Hebbian learning. *The Journal of Neuroscience*, *38*(44), 9563–9578. https://doi.org/10.1523/JNEUROSCI.1259-18.2018

Cooke, S., & Bear, M. (2010). Visual Experience Induces Long-Term Potentiation in the Primary Visual Cortex. *Journal of Neuroscience*, *30*(48), 16304–16313. https://doi.org/10.1523/JNEUROSCI.4333-10.2010

De Valois, R., Albrecht, D., & Thorell, L. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research*, *22*(5), 545–559. https://doi.org/10.1016/0042-6989(82)90113-4

Delorme, A., Perrinet, L., & Thorpe, S. J. (2001). Networks of integrate-and-fire neurons using Rank Order Coding B: Spike timing dependent plasticity and emergence of orientation selectivity. *Neurocomputing*, *38–40*, 539–545. https://doi.org/10.1016/S0925-2312(01)00403-9

Furmanski, C., & Engel, S. (2000). An oblique effect in human primary visual cortex. *Nature Neuroscience*, *3*(6), 535–536. https://doi.org/10.1038/75702

Furmanski, C. S., Schluppeck, D., & Engel, S. A. (2004). Learning Strengthens the Response of Primary Visual Cortex to Simple Patterns. *Current Biology*, *14*(7), 573–578. https://doi.org/10.1016/j.cub.2004.03.032

Geisler, W. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Gollisch, T., & Meister, M. (2008). Rapid Neural Coding in the Retina with Relative Spike Latencies. *Science*, *319*(5866). Retrieved from http://science.sciencemag.org/content/319/5866/1108

Grunewald, A., Bradley, D. C., & Andersen, R. A. (2002). Neural Correlates of Structure-from-Motion Perception in Macaque V1 and MT. *Journal of Neuroscience*, *22*(14), 6195–6207. https://doi.org/10.1523/JNEUROSCI.22-14-06195.2002

Gütig, R., Aharonov, R., Rotter, S., & Sompolinsky, H. (2003). Learning Input Correlations through Nonlinear Temporally Asymmetric Hebbian Plasticity. *Journal of Neuroscience*, *23*(9), 3697–3714.

Harnack, D., Pelko, M., Chaillet, A., Chitour, Y., & Rossum, M. C. W. van. (2015). Stability of Neuronal Networks with Homeostatic Regulation. *PLOS Computational Biology*, *11*(7), e1004357. https://doi.org/10.1371/journal.pcbi.1004357

Heeley, D. W., & Timney, B. (1988). Meridional anisotropies of orientation discrimination for sine wave gratings. *Vision Research*, *28*(2), 337–344. https://doi.org/10.1016/0042-6989(88)90162-9

Hoyer, P., & Hyvärinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, *11*(3), 191–210. https://doi.org/10.1088/0954-898X_11_3_302

Hübener, M., & Bonhoeffer, T. (2014). Neuronal Plasticity: Beyond the Critical Period. *Cell*, *159*(4), 727–737. https://doi.org/10.1016/j.cell.2014.10.035

Hunter, D., & Hibbard, P. (2015). Distribution of independent components of binocular natural images. *Journal of Vision*, *15*(13), 6. https://doi.org/10.1167/15.13.6

Hurley, N., & Rickard, S. (2009). Comparing Measures of Sparsity. *IEEE Transactions on Information Theory*, *55*(10), 4723–4741. https://doi.org/10.1109/TIT.2009.2027527

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications. *Neural Networks*, *13*(4–5), 411–430. https://doi.org/10.1016/S0893-6080(00)00026-5

Isomura, T., & Toyoizumi, T. (2018). Error-Gated Hebbian Rule: A Local Learning Rule for Principal and Independent Component Analysis. *Scientific Reports*, *8*(1), 1835. https://doi.org/10.1038/s41598-018-20082-0

Jones, J., & Palmer, L. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1233–1258.

Kang, K., & Sompolinsky, H. (2001). Mutual Information of Population Codes and Distance Measures in Probability Space. *Physical Review Letters*, *86*(21), 4958–4961. https://doi.org/10.1103/PhysRevLett.86.4958

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Li, B., Peterson, M. R., & Freeman, R. D. (2003). Oblique Effect: A Neural Basis in the Visual Cortex. *Journal of Neurophysiology*, *90*(1), 204–217. https://doi.org/10.1152/jn.00954.2002

Martinez-Garcia, M., Martinez, L. M., & Malo, J. (2017). Topographic Independent Component Analysis reveals random scrambling of orientation in visual space. *PLOS ONE*, *12*(6), e0178345. https://doi.org/10.1371/journal.pone.0178345

Masquelier, T. (2012). Relative spike time coding and STDP-based orientation selectivity in the early visual system in natural continuous and saccadic vision: A computational model. *Journal of Computational Neuroscience*, *32*(3), 425–441. https://doi.org/10.1007/s10827-011-0361-9

Masquelier, T., & Thorpe, S. J. (2007). Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology*, *3*(2), e31. https://doi.org/10.1371/journal.pcbi.0030031

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., & Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, *17*(10), 1410–1417. https://doi.org/10.1038/nn.3807

Nienborg, H., & Cumming, B. G. (2006). Macaque V2 Neurons, But Not V1 Neurons, Show Choice-Related Activity. *Journal of Neuroscience*, *26*(37), 9567–9578. https://doi.org/10.1523/JNEUROSCI.2256-06.2006

Ohzawa, I., & Freeman, R. (1986). The binocular organization of complex cells in the cat's visual cortex. *Journal of Neurophysiology*, *56*(1), 243–259.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. https://doi.org/10.1038/381607a0

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325. https://doi.org/10.1016/S0042-6989(97)00169-7

Orban, G. A., Vandenbussche, E., & Vogels, R. (1984). Human orientation discrimination tested with long stimuli. *Vision Research*, *24*(2), 121–128. https://doi.org/10.1016/0042-6989(84)90097-X

Perrinet, L. U. (2019). An Adaptive Homeostatic Algorithm for the Unsupervised Learning of Visual Features. *Vision*, *3*(3), 47. https://doi.org/10.3390/vision3030047

Perrinet, L. U., & Bednar, J. A. (2015). Edge co-occurrences can account for rapid categorization of natural versus animal images. *Scientific Reports*, *5*, 11400. https://doi.org/10.1038/srep11400

Quian Quiroga, R., & Panzeri, S. (2009). Extracting information from neuronal populations: Information theory and decoding approaches. *Nature Reviews Neuroscience*, *10*(3), 173–185. https://doi.org/10.1038/nrn2578

Raichle, M. E. (2010). Two views of brain function. *Trends in Cognitive Sciences*, *14*(4), 180–190. https://doi.org/10.1016/j.tics.2010.01.008

Ringach, D. (2002). Spatial Structure and Symmetry of Simple-Cell Receptive Fields in Macaque Primary Visual Cortex. *Journal of Neurophysiology*, *88*(1). Retrieved from http://jn.physiology.org/content/88/1/455.short

Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*(2), 147–166. https://doi.org/10.1207/s15516709cog2802_2

Rullen, R. V., & Thorpe, S. (2001). Rate Coding Versus Temporal Order Coding: What the Retinal Ganglion Cells Tell the Visual Cortex. *Neural Computation*, *13*(6), 1255–1283. https://doi.org/10.1162/08997660152002852

Savin, C., Joshi, P., & Triesch, J. (2010). Independent Component Analysis in Spiking Neurons. *PLoS Computational Biology*, *6*(4), e1000757. https://doi.org/10.1371/journal.pcbi.1000757

Shriki, O., Kohn, A., & Shamir, M. (2012). Fast Coding of Orientation in Primary Visual Cortex. *PLoS Computational Biology*, *8*(6), e1002536. https://doi.org/10.1371/journal.pcbi.1002536

van Hateren, J., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings. Biological Sciences / The Royal Society*, *265*(1394), 359–366. https://doi.org/10.1098/rspb.1998.0303

Vogels, R., & Orban, G. A. (1990). How well do response changes of striate neurons signal differences in orientation: A study in the discriminating monkey. *Journal of Neuroscience*, *10*(11), 3543–3558. https://doi.org/10.1523/JNEUROSCI.10-11-03543.1990

Wandell, B., & Smirnakis, S. (2010). Plasticity and stability of visual field maps in adult primary visual cortex. *Nature Reviews Neuroscience*, *10*(12), 873. https://doi.org/10.1038/nrn2741

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249. https://doi.org/10.1073/pnas.1619153114

Zhaoping, L. (2006). Theoretical understanding of the early visual processes by data compression and data selection. *Network: Computation in Neural Systems*, *17*(4), 301–334. https://doi.org/10.1080/09548980600931995

Zylberberg, J., Murphy, J. T., & DeWeese, M. R. (2011). A Sparse Coding Model with Synaptically Local Plasticity and Spiking Neurons Can Account for the Diverse Shapes of V1 Simple Cell Receptive Fields. *PLoS Computational Biology*, *7*(10), e1002250. https://doi.org/10.1371/journal.pcbi.1002250