Intrinsic Plasticity in a Generative Model of V1

Cornelius Weber and Jochen Triesch
Frankfurt Institute for Advanced Studies
Johann Wolfgang Goethe University
Frankfurt am Main, Germany

Abstract  Current models for the learning of feature detectors work on two time scales: on a fast time scale the internal neurons’ activations adapt to the current stimulus; on a slow time scale the weights adapt to the statistics of the entire set of stimuli. Here we explore the role of medium time scale adaptation of a neuron’s transfer function parameters to fluctuating statistics of the stimuli. By such an intrinsic plasticity mechanism, a neuron maintains homeostasis of its firing rate in a dynamic environment and compensates for possible rate changes during weight adaptation.

Two parameters of a neuron’s transfer function, threshold and gain, are adapted to maintain a sparse exponential firing rate distribution. Building on previous results [6, 8], the neurons become localized edge detectors by training on natural images, hence they are models of V1 simple cells. The medium time scale involved allows to explore physiological visual effects such as the tilt-aftereffect. In this effect, after a viewer adapts to a grid of a certain orientation, grids of a nearby orientation will be perceived as tilted away from the adapted orientation. Our results show that adapting the neurons’ gain but not their thresholds leads to quantitatively similar aftereffects as found psychophysically. This effect is mainly due to gain decrease at the adapting orientation. A sole gain increase at orientations away from the adapting orientation would yield a similar, smaller aftereffect. The model does not have additional horizontal connectivity as previous accounts of the tilt-aftereffect [1].

Contrasting with the weakening of adapted stimulus representations in the tilt-aftereffect, long-term exposure to gratings of a certain orientation, e.g. 45°, yields enhanced representations of that orientation [5]. The model also reproduces these results. When trained with natural images it reproduces the experimentally observed enhanced representations of vertical and horizontal edges.

The intrinsic plasticity mechanism achieves robust neuronal behavior, because it automatically adjusts the two transfer function parameters to maintain firing rate homeostasis. Despite its simplicity it accounts for more experimental details than models without this mechanism.

Model  Figure 1 a) shows the model architecture [8]. The input vector \( \mathbf{x} \) at the bottom is conveyed via the bottom-up weights to the “net input” \( \mathbf{y} = \mathbf{W}^{bu} \mathbf{x} \) of the hidden neurons. The transfer function \( g \), shown at the top of Fig. 1 a), computes the neuronal output \( \mathbf{z} \). It has two modifiable parameters: \( a \), which scales the input and resembles a gain, or slope, and \( b \), which shifts the curve along the input axis. The weights are normalized so that \( \sum_j (w_{ij})^2 = 1 \) for each hidden neuron \( i \). The hidden variables \( \mathbf{z} \) are subject to an exponential prior density function \( f_{z}(z) \) on \( z_i \) in the range \([0, 1]\), drawn right of Fig. 1 a). Its sparseness is parameterized
a) \[ z = g(y) = \frac{1}{1 + \exp(-(ay + b))} \]

\[ y = W^{bu} x \]

\[ x \quad \hat{x} = W^{td} \tilde{z} \]

b) Figure 1: a) Model architecture and variables. b) Some sample receptive fields (RFs) trained from images. c) Some RFs (same as the two bottom rows in b)) after additional training with 1,000,000 samples of 45° gratings, d) after additional 3,000,000 samples of gratings.

by its mean \( \mu \). The reconstruction \( \tilde{x} \) of the input is computed from the hidden variables using the top-down weights: \( \tilde{x} = W^{td} \tilde{z} \).

**Maintaining Sparseness** Given the data distribution over \( x \) and slow changing weights \( W^{bu} \), the transfer function parameters \( a_i \) and \( b_i \) of each neuron \( i \) individually are adjusted to make \( z_i \) distributed as an exponential \( f(z_i) \approx f_{\exp}(z_i) \), as indicated in the right of Fig. 1 a). This is done by adjusting the parameters \( a_i \) and \( b_i \) so to minimize the Kullback Leibler divergence \( d(f_{z_i} || f_{\exp}) \) between the desired distribution \( f_{\exp}(z_i) \) and the actual distribution \( f_{z_i}(z_i) \). Gradient descent yields incremental local update rules [6]:

\[ \Delta a_i = \eta_a \frac{\partial}{\partial a_i} d(f_{z_i} || f_{\exp}(z_i)) = \eta_a \left( \frac{1}{\mu} \right)^2 \frac{\partial}{\partial a_i} \left( \frac{1}{\mu} \right)^2 \]

\[ \Delta b_i = \eta_b \frac{\partial}{\partial b_i} d(f_{z_i} || f_{\exp}(z_i)) = \eta_b \left( 1 - 2 \mu \right)^2 \frac{\partial}{\partial b_i} \left( 1 - 2 \mu \right)^2 \]

\( f_z \) is made sparse by setting the average small: \( \mu = 0.01 \). Update rates are \( \eta_a = \eta_b = 0.001 \).

**Learning Weights** According to the wake-sleep algorithm [3], the generative weights \( W^{td} \) are trained from the data in a so-called wake phase. Each data point \( x \) was a patch randomly cut out of a set of natural grey scale images. These were pre-processed with a low-pass filter that is described in frequency space as \( f \cdot \exp(-f_0^4) \), with \( f_0 = 200 \), pixels in frequency space [4], and the mean of each image patch \( x \) was set to zero. Learning consists of an error-based Hebbian-like learning rule:

\[ \Delta w^{td}_{ji} = \eta_w \hat{x}_j z_i, \quad \text{where} \quad \hat{x} = x - W^{td} z, \quad \text{(reconstruction error)} \]

The recognition weights \( W^{bu} \) are trained to functionally invert the generative weights in the sleep phase in which no data are present. For this purpose, “fantasy” vectors \( \tilde{z} \) are presented on the hidden units. These are sparse, binary random activation patterns with activations set to either 0 or 0.5 and mean \( \mu \). They are projected as \( \tilde{x} = W^{td} \tilde{z} \) to the input units, and the recognition weights perform the functional inversion \( W^{bu} \tilde{x} \) by attempting to reconstruct the fantasy vectors. The error-based Hebbian-like learning rule is:

\[ \Delta w^{bu}_{ij} = \eta_w \tilde{z}_i \tilde{x}_j, \quad \text{where} \quad \tilde{z} = \tilde{z} - g(W^{bu} \tilde{x}) \]
which uses the transfer function $g$ of Fig. 1 a). Before computing the reconstruction error $\hat{z}$, the variance of the reconstruction $g(W^{bu}\hat{\mathbf{x}})$ was set to that of $\hat{z}$. Similar was done with $\mathbf{x}$ and $\hat{\mathbf{x}}$ in the wake phase but was later found unnecessary. \(\eta_w = 0.001\) is the step size.

The resulting weights are obtained after 5,000,000 learning steps. Each step consisted of a wake phase which involves Eqs. 1, 2 and 3, and one sleep phase using Eq. 4.

**Results** The receptive fields (RFs) of some units are shown in Fig. 1 b). They resemble localized edge detectors and their centers are approximately evenly distributed over the input plane as seen in Fig. 2 a). RF spatial frequencies and orientation vary a lot, as can be seen in Fig. 2 b). The $a$-parameter varies approximately two-fold, and neurons with a larger RF tend to have a smaller gain $a$, as shown in Fig. 2 c). This variety is despite the constant mean set for all neurons. In contrast, the parameter $b$ varies by less than $\pm 3\%$ (result not shown).

**Long-Term Adaptation** The orientation histogram of the trained map, Fig. 3 left, shows that the cardinal axes (0° and 90°) are overrepresented. Orientations \(\sim 10°\) away from these are underrepresented. We have further trained the final map with gratings of random phase and frequency, but same orientation of 45°, mimicking experiments [5]. Figs. 1 c,d) and 3 show that then many receptive fields decay and several neurons with grating-shaped RFs of 45° orientation emerge.

**Short-Term Adaptation** is performed by presenting a grating stimulus for several iterations. Fig. 4 a) shows that the gain of the neurons (parameter $a$) at the adapting orientation

![Figure 2](image1.png)

Figure 2: a) Receptive field (RF) centers (each cross denotes one unit). b) Distribution of the RF filter frequencies. c) The values of the neurons' parameter $a$ (y-axis) over the RF diameter. The RF properties were determined from gabor fits.
Figure 4: Tilt aftereffect. a) The change of $a$ and $b$ during adaptation, dependent on the orientation preference of the neurons. The x-axis denotes the difference between the neurons’ preferred orientation and the orientation of the adapting stimulus. b) and c) The x-axis denotes the difference between the orientations of test- and adapting stimulus. The y-axis plots the tilt of the perceived orientation of the test pattern, compared to pre-adaptation perception. It is perceived as tilted away from the orientation of the adaptation pattern. Every curve is an average over 180 experiments performed with different adaptation orientations. b) shows that if both intrinsic plasticity parameters, $a$ and $b$, are allowed to adapt, then the model tilt aftereffect is much wider than experimental data (small black circles) [2]; similarly, if only $b$ is varied. However, if $b$ is fixed during adaptation ($\eta_b = 0$), then the TAE curve matches experimental data well. 360 adaptation patterns were shown. In c) $\eta_b = 0$, and the number of adaptation patterns ($\sim$ time of adaptation) is varied, from 10 to 14,400.

decreases, but increases at other orientations. Similarly, the “threshold” parameter $b$ becomes more negative at the adapting orientation. These changes give rise to the tilt- aftereffect (TAE) at which a test pattern is perceived tilted away from the orientation of the adapting pattern, as shown in Fig. 4 b). However, the shape of the effect matches the physiological experiment only if $a$ is adapted. Adaptation of $b$ yields a too wide influence of the adapting pattern. As observed experimentally [2], the TAE rises approximately logarithmically with adaptation time, however, unlike in the experiment, it does not saturate.

Discussion  The presented model joins a neuron’s intrinsic plasticity which regulates it’s firing rate to be exponential with a generative model that exploits this sparse firing for the learning of localized edge detectors. Biological neurons adapt on multiple time scales, for example, firing rates are stabilized by homeostatic mechanisms over time scales of $\sim$ 24 hours [7]. Our proposed mechanism could as well be applied to this result, but here we have focused on the TAE which happens on a time scale of several seconds. This was explained by adapting only the gain parameter $a$. The “threshold” parameter $b$ might change on a longer time scale (by setting $\eta_b \ll \eta_a$). Together, our objective function joins TAE and RF development.

Acknowledgments  JT and CW are financed by the Hertie Foundation, and the EU project PLICON, Contract No. MEXT-CT-2006-042484.

References