Biophysical Models of Neural Computation: Max and Tuning Circuits

Ulf Knoblich, Jake Bouvrie and Tomaso Poggio

CBCL Paper

April 20, 2007

Abstract

Pooling under a softmax operation and Gaussian-like tuning in the form of a normalized dot-product were proposed as the key operations in a recent model of object recognition in the ventral stream of visual cortex. We investigate how these two operations might be implemented by plausible circuits of a few hundred neurons in cortex. We consider two different sets of circuits whose different properties may correspond to the conditions in visual and barrel cortices, respectively. They constitute a plausibility proof that stringent timing and accuracy constraints imposed by the neuroscience of object recognition can be satisfied with standard spiking and synaptic mechanisms. We provide simulations illustrating the performance of the circuits, and discuss the relevance of our work to neurophysiology as well as what bearing it may have on the search for maximum and tuning circuits in cortex.

1 Introduction

A recent theory of object recognition in the feed-forward pathway of the ventral stream in primate visual cortex [28, 30] is based on a hierarchical model with two main operations iterated several times throughout the hierarchy. The two basic operations are a bell-shaped tuning mechanism to provide selectivity of units – learned in an unsupervised way – along the pathway and a maximum-like operation to achieve invariance for position, scale and clutter while maintaining selectivity. To support the biological plausibility of the theory it is important to show that these two basic operations can be implemented using well-established, plausible properties of neurons and synapses. The primary goal here is to describe realistic circuits for the maximum and tuning operations utilized by this model. There are several possible circuit designs that can theoretically provide the level of robustness required by model simulations, and yet, little is known about how such circuits might be implemented in the brain. The circuits described herein therefore represent a plausibility proof and may also be seen as specific hypotheses that neurophysiologists may try to prove or disprove using, for example, extracellular and intracellular recordings.

Several groups have reported neurons in visual cortex that respond rapidly and sub-linearly to the combined presentation of two simple stimuli in their receptive field [10, 19], i.e. the response to the combined stimulus is significantly smaller than the sum of the responses to the single stimuli. It has been proposed that, instead of summing the inputs, these neurons compute either the maximum or the average of the inputs. Normalization circuits were suggested for gain control in [3] and for the biased competition model in [5] (see also [26, 27] and for older work on related dynamics of shunting inhibition [13]). Another possible mechanism for gain control relies on synaptic depression [1]. Several possible circuits for computing the maximum operation have been proposed on an abstract level [26, 38], but were not investigated with respect to their biophysical plausibility. A recent study presented a framework to unify both computations, softmax and tuning, providing a single equation that will produce either of the two behaviors depending on a small set of parameters [18]. The circuits presented here share some of the general architectural features presented in the above research. Additionally, we emphasize biological plausibility to a novel extent.

First, we introduce quite restrictive timing constraints on the circuits performing both computations imposed by physiological data and discuss how and where such constraints arise. We then present two different coding schemes in different sensory cortices and their implications, leading to the development of two sets of circuit models matched to the conditions prevalent in the respective areas. In section 2 we describe the architecture, dynamics, parameters and results of the Population Packet Code models, followed by the Poisson Code models in section 3 and a discussion of the results and their relevance.

1.1 Time, computational, and bandwidth constraints

A plausible circuit must satisfy several constraints imposed by the physiological mechanisms underlying synaptic transmission: we must perform the above computations assuming that information leaves and arrives in discrete spikes. Furthermore, object recognition simulations with a computational model [30] suggest that the connections between visual areas must be able to relay at least 2 bits of information (within about 10 ms– to be discussed later), *i. e.,* represent approximately 4 different levels of spike activity at the population level during time bins of as little as around 10-20 ms. It has been found that the model is robust to perturbations of the max and tuning operations. In particular, the input to the highest level can be binarized without significantly affecting performance in multi-class recognition tasks. The entire visual recognition model, encompassing many maximum and tuning sub-circuits, must be able to perform the basic computation within an amount of time that is faithful to the known physiology. Recordings from inferotemporal cortex (IT) [16] reveal that information containing sufficient accuracy for good categorization arrives in anterior IT about 100-120 ms after stimulus onset and about 20 ms after onset of activity in AIT (which is about 80 ms after stimulus onset). In addition, [16] found that neural activity in a population of 256 neurons during a time bin of just 12.5 ms at around 20 ms after onset of the AIT response contain significant information for good categorization accuracy. This implies that each stage in the recognition model – roughly corresponding to a visual area (V1, V2, V4, PIT and AIT) – has about 10 to 20 ms for processing and representing the information. In the mammalian brain, the firing rates of excitatory cells in visual cortex rarely exceed 100Hz, and we can expect at most 1-2 spikes per neuron within the prescribed time window. Thus, a single neuron cannot possibly provide enough dynamic range given the timing constraints imposed by available physiological data.

To overcome this insufficiency, we replace each unit with an ensemble of n cells. All cells in an ensemble have similar, but not identical, parameters and receive highly correlated inputs from the other units in the circuit. Instead of 0 or 1 spike, the postsynaptic cell can now receive up to $2n$ spikes, *i. e.,* the dynamic range of the input is multiplied by n, as shown in Fig. 1[1]. In early stages of the visual pathway, a large dynamic range of the inputs is needed, whereas at the other extreme in IT, only the binary presence or absence of each feature has to be conveyed.[2] In the spiking circuits presented below, we therefore consider ensembles of spiking neurons for each unit in the computational model presented in [30]. In particular, we will also consider redundant parallel copies of circuit inputs, outputs, and networks themselves in order to overcome the above constraints placed on the timing and resolution of the two computations of interest. This is the essence of our contribution: we use ensembles of redundant spiking cells to achieve fast, "non-binary" computation.

1.2 Different Coding Regimes in Sensory Cortices

Neurons in the visual cortex of cats and primates are thought to have spontaneous and evoked firing rates on the order of tens of Hertz. At these rates, Poisson spike trains are a fair approximation for the observed firing patterns. Each presynaptic neuron can contribute up to 3-4 spikes within the first 20-30ms of processing in the postsynaptic cell and dynamic circuit models can assume that the underlying rate of the Poisson spike train inputs is constant.

However, there are other sensory cortices in which the coding is very different. Rodent somatosensory "barrel" cortex is an example for extremely sparse representation of sensory stimuli. Both the spontaneous

[1] It is thought that the number of cells per ensemble n decreases along the visual hierarchy from V1 to IT.

[2] Contrast invariance data provide some indirect support for the idea that the *cables get thinner along the hierarchy.* [29] showed that the steepness of the contrast-response functions of neurons increases from LGN through V1, V2 to MT and that "cells become, in the contrast domain, progressively more like switches, being either on or off" [20].

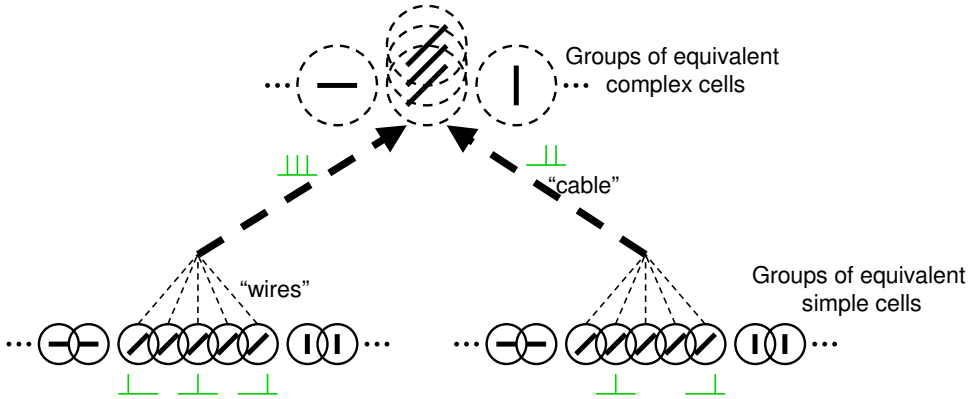

Figure 1: Signal propagation between two somata with spiking neurons and "cables". Each of the ensemble cells on the bottom, which receive highly correlated input, spikes independently. Each spike is propagated independently in its own "wire" in the "cable" and then triggers neurotransmitter release at one synapse for each of the ensemble postsynaptic neurons. This process is used for each synapse in the model.

and evoked firing rates in barrel cortex have been estimated to be below 1Hz by several groups. In other words, a single neuron responds to a given stimulus with on average less than one spike per trial, *i. e.,* either zero or one spike. These extremely low rates in conjunction with precise spike timing call for circuits that perform computations based on a single volley or "packet" of incoming spikes from an ensemble of presynaptic neurons. We call the corresponding coding scheme *Population Packet Code*.

Rat primary somatosensory cortex (SI) is an ideal model system to address questions about circuitry and coding, not only because of its sparse coding but also due to its regular columnar architecture, the barrel columns [35] and its externalized receptors, the vibrissae. These properties allow exquisite control over the stimulation, recording and analysis of discrete channels of information processing which are impossible with visual stimulation, since every visual stimulus will excite a large number of photoreceptors while the deflection of a single whisker only activates a single follicle. In addition, SI shares several basic properties of its circuitry and cells with cat or primate primary visual cortex (V1) and the two systems exhibit similar generalization and learning capabilities, likely because both are high-resolution sensory systems.

2 Population Packet Code Models

Transferring analog graded information with digital spikes poses the problem of how to encode analog values with spikes. Three common proposals to overcome this problem are to either use the population firing rate, a purely temporal code such as time-to-spike or representations based on synchrony. Our proposed coding scheme lies somewhat in between continuous population rate coding and synfire chains, two more widely studied coding schemes [32]. Given the constraints outlined above, we assume here that information is coded by the number of spikes in a brief packet of spikes from a presynaptic population, i.e., by a short-time "packet" population rate code. In this coding scheme, when a single cell integrates spikes from several converging inputs, the dynamic range of the signal is strongly compressed. Either the combination of inputs from the ensemble of presynaptic neurons provides enough excitatory conductance to drive the postsynaptic membrane potential above spike threshold or not. Considering an ensemble of identical postsynaptic neurons all receiving identical input from the presynaptic ensemble, the responses of all cells are identical. Thus, there will be no postsynaptic spikes for any combination of inputs below a certain threshold. Above this threshold, all postsynaptic cells will spike together. The dynamic range of the signal is compressed into a binary signal. However, neurons in cortex show diversity in their morphology and physiological parameters and receive ongoing "background", noise-like input from many different presynaptic cells that are overlapping but not identical for a given set of postsynaptic neurons.

It has been shown that this synaptic background input can decrease the slope of the spike probability function dependent on input strength [6]. In an ensemble of neurons, the synaptic activity together with variable intrinsic properties of the cells effectively linearizes the transfer function of the ensemble under the proposed coding scheme (see also [25] for a general argument).

2.1 Network architecture for softmax and Gaussian-like tuning

The population packet code model is a simplified canonical cortical microcircuit [8] using ensembles of integrate-and-fire neurons. Importantly, *the same canonical circuit perform either a max or a tuning operation depending on just different synaptic conductance values.* In the proposed coding scheme, computation has to be quasi-instantaneous, *i. e.,* carried out over in a short period of time during which each neuron can only spike at most once or twice. Thus the notion of a steady state firing rate is not appropriate for this case. Unlike the circuit presented in [21], which operates on a timescale of hundreds of milliseconds, our model is entirely feed-forward since the need for very fast and inherently transient computation implies that recurrent connections would most likely be too slow to contribute.

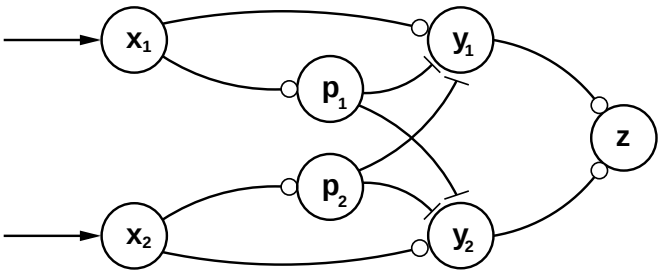

Figure 2: Population packet code network architecture. Small circles indicate excitatory synapses, bars represent inhibitory connections. Note that the *same circuit is used to approximate a max as well as a tuning operation with different conductances for some of the synapses.*

The most salient feature of our model architecture is the combination of monosynaptic excitation ($x \rightarrow y$) with disynaptic inhibition ($x \rightarrow p \rightarrow y$), which is strongly supported by the observation that strong excitation in form of an excitatory postsynaptic potential (EPSP) is generally followed by an inhibitory postsynaptic potential (IPSP) after about 4ms, which has been made during intracellular recordings by several labs [22, 24, 33, 34]. Following the analogy with barrel cortex, the x units correspond to thalamic cells while the p, y and z units correspond to cortical cells. The inhibitory interneurons p synapse onto y units of the same channel as well as the other channel. The operating regime of these interneurons will be the determining factor for which computation (max or tuning) is performed by the microcircuit. Both y units synapse onto the output z unit, which relays the output. Each unit depicted in Fig. 2 is implemented as an ensemble of 100 cells modeled as a variant of integrate-and-fire neurons as described below.

Sparse connectivity Although commonly used in models and simulations, there has been no anatomical evidence for ensembles of hundreds of neurons exhibiting all-to-all connectivity. Instead, connectivity patterns of cortical neurons depend on spatial distance. Although more careful analysis reveals a dependence on specific cell types [36, 37], nearby neurons are generally more interconnected and show a high correlation in their membrane potential [22]. Because not all cells in an ensemble are driven by exactly the same primary input, their output is more variable, increasing the dynamic range of the ensemble. We model this kind of connectivity by using a noisy Gaussian synaptic strength pattern between a given presynaptic cell and all its potential targets in the postsynaptic ensemble. For a presynaptic cell with index m and a postsynaptic cell with index n, the synaptic scaling factor is

$$\tilde{g}^{mn} = exp(\frac{-(m-n)^2}{2\sigma_d^2}) \cdot \mathcal{N}(1, \sigma_g) \tag{1}$$

Consistent with experimental observations [31], inhibitory interneurons receive less specific inputs, *i. e.,* σ_d is larger for interneurons. The parameters used for these simulations are $\sigma_{de} = 50$ and $\sigma_{di} = 100$.

For the max computation, the synaptic conductances were $\hat{g}_{xy} = 1.2nS$, $\hat{g}_{xp} = 0.16nS$, $\hat{g}_{py} = 0$, $\hat{g}_{pyl} = 0.8nS$ and $\hat{g}_{yz} = 0.5nS$. The input conductance to the inhibitory interneurons p is smaller than the inputs to the excitatory neurons because of the different morphology, physiology and connectivity of these cells. The, in comparison to the tuning, relatively high $\hat{g}_{xp}$ causes the p units to be fairly active even for moderate input levels, causing strong cross-inhibition via the cross-channel $p \to y$ synapse.

For the tuning computation, the synaptic conductances were $\hat{g}_{xy} = 1.2nS$, $\hat{g}_{xp} = 0.09nS$, $\hat{g}_{py} = 4nS$, $\hat{g}_{pyl} = 4nS$ and $\hat{g}_{yz} = 0.72nS$. The tuning point, *i. e.,* the input eliciting the maximal output, is governed by the balance of excitatory and inhibitory input to the y unit in combination with the transfer function of the p unit. The lower $\hat{g}_{xp}$ in this configuration causes the p units to be activated only for stronger inputs, causing the decreasing output for those high input values because the inhibitory transfer function is steeper than the excitatory one, compatible with high-threshold, high-gain inhibitory interneurons found in cortex such as Chandelier cells [39].

2.2 Neural dynamics

Each cell is modeled as a conductance-based single compartment leaky integrate-and-fire neuron with a simple kinetic model of neurotransmitter receptors. Variables and constants that are unique to each of the n_w instances in a group are marked with upper indices such as V_i^m and similarly for terms specific to an instance of a synapse such as g_{ij}^{mn}, the conductance of the synapse between cell m of group i and cell n of group j. If the upper indices are missing, the value is the same for all instances.

2.2.1 Integrate-and-Fire dynamics

The membrane potential of a postsynaptic neuron is calculated according to the membrane equation

$$C_j^n \frac{dV_j^n}{dt} = \sum_{i,m} g_{ij}^{mn}(E_{ij} - V_j^n) + g_{Lj}(E_{Lj} - V_j^n) \tag{2}$$

The capacitive current charging the cell membrane is equal to the sum of the synaptic and leak currents. In order to better capture the firing properties of cortical neurons, we do not reset the membrane potential once it reaches the spike threshold V_θ but instead activate a strong hyper-polarizing current with a time constant of a few milliseconds which will bring the cell's potential back to its "resting" state.

Intrinsic cell parameters Neurons in cortex have very different morphologies and channel distributions. While we cannot model these directly in our single compartment models, we impose some variability on the resulting physiological parameters such as membrane capacitance, input resistance and spike threshold. In accordance with several physiological studies, interneurons are assumed to be smaller, *i. e.,* having smaller membrane capacitance. The interneurons in this model are of the high-threshold, high-gain class discussed in numerous studies, *i. e.,* their spike threshold is higher than for the excitatory neurons ($V_{\theta e} = -50mV$, $V_{\theta i} = -45mV$). The membrane capacitance of a cell is a truncated normally distributed random but fixed variable of the form: $C_j^n = \hat{C}_{[e,i]} \cdot \tilde{C}_j^n$ with $\tilde{C}_j^n = \mathcal{N}(1, \sigma_C) \in [0.5, 1.5]$. The reversal potential of the leak current is set to $E_L = -70mV$ and its conductance is $g_L = 10nS$.

2.2.2 Synaptic input

Release of neurotransmitter, the first step of synaptic transmission, is modeled as an all-or-none release of one "vesicle" of $1mM$ transmitter into the synaptic cleft which is present for $1ms$.

To model ionotropic receptors such as AMPA/Kainate and $GABA_A$ receptors, we assume simple two-state kinetics where the receptor is either open or closed. In the closed state it can bind neurotransmitter which results in an transition to the open state. In the open state, the channel will close as a result of dissociation.

$$C + T \rightleftharpoons O, \tag{3}$$

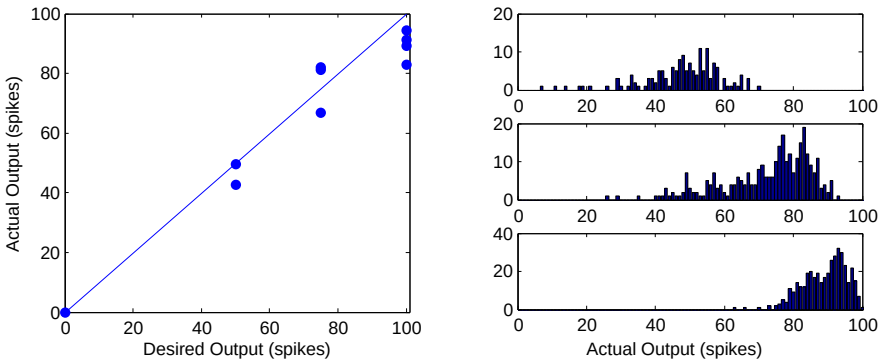

Figure 3: Mean response of max circuit depicted in Fig. 2 over 50 runs for all possible combinations of 0, 50, 75 and 100 spikes per input packet, plotted against the desired (true) maximum of the inputs (left). Histogram of all outputs (spike count in output packet) for three cases (right). The true maximum of the inputs is 50, 75 and 100 spikes, respectively (top to bottom).

These simple kinetics can be described by the following first-order kinetic equation:

$$\frac{dr_{ij}^{mn}}{dt} = \alpha_{ij} T_{ij}^{m} (1 - r_{ij}^{mn}) - \beta_{ij} r_{ij}^{mn} \tag{4}$$

The rate constants α and β, are set to the values reported in [7] for either AMPA and GABA$_A$, respectively. The synaptic input conductance is calculated from the fraction of open channels calculated in (4), adding a normally distributed random background component (see below):

$$g_{ij}^{mn} = \left[r_{ij}^{mn} \hat{g}_{ij} \tilde{g}_{ij}^{mn} + \mathcal{N}(\bar{g}_{ij}, \sigma_{g_{ij}}) \right]^{+} \tag{5}$$

The current results are based on AMPA as excitatory and GABA$_A$ as inhibitory neurotransmitters and their respective receptors. The parameters are $E_{AMPA} = 0mV$, $E_{GABA_A} = -80mV$, $\alpha_{AMPA} = 1.1 \cdot 10^6 M^{-1}s^{-1}$, $\alpha_{GABA_A} = 5 \cdot 10^6 M^{-1}s^{-1}$, $\beta_{AMPA} = 190s^{-1}$, $\beta_{GABA_A} = 180s^{-1}$.

Noise-like background synaptic input Background synaptic noise, which is likely to represent unknown inputs, helps to recover the dynamic range of a population of neurons by linearizing the spike probability function [6]. If different neurons in an ensemble have slightly different synaptic inputs, their membrane potential will slightly vary. Thus, the same incoming postsynaptic potential (PSP) might drive some neurons with higher membrane potentials above the spike threshold, initiating a spike, while others with lower potential will stay below threshold and not spike. Adding synaptic background inputs also introduces a variability in the output spike timing. Compatible with data from *in vivo* and *in vitro* recordings, spike time jitter decreases with increasing strength of the input. The parameters used for the conductances (5) are $\bar{g}_e = 2nS$, $\bar{g}_i = 1nS$, $\sigma_{g_e} = 1nS$, $\sigma_{g_i} = 1nS$.

2.2.3 External input

As input to the circuit, the x units receive a variable number of spikes proportional to the input strength. Each spike is assumed to originate at one out of 100 presynaptic cells. The connectivity pattern is the same Gaussian shaped synaptic strength profile as for the other excitatory units in the model and the spike arrival times are normally distributed as $\mathcal{N}(\bar{t}_x, \sigma_{t_x})$ with $\bar{t}_x = 30ms$ and $\sigma_{t_x} = 5ms$.

2.3 Population Packet Code Simulations

2.3.1 Maximum

Figure 3 (left) shows the mean output over 50 runs of the circuit depicted in Fig. 2 in maximum configuration for all possible combinations of four levels of inputs (0, 50, 75, 100 spikes in a packet). For

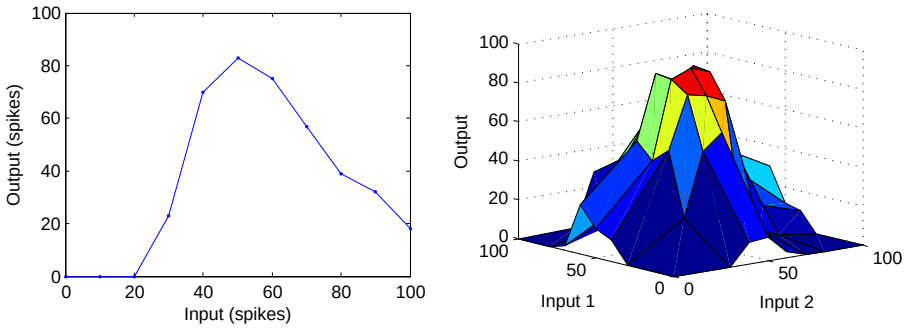

Figure 4: Output (spike count in output packet) of a one-dimensional Gaussian-like tuning circuit tuned to 50 a spike packet input (left). Output (spike count in output packet) of the two-dimensional tuning circuit depicted in Fig. 2 tuned to the combination of two 50 spike packet inputs (right).

input combinations including the maximum (100 spikes) input, the circuit output is actually below the desired maximum value. Interestingly, similar effects can be seen *in vivo*, *e. g.,* in rat barrel cortex , where subthreshold neural responses tend to be more suppressive for stronger inputs. The histograms on the right show the distribution of outputs for the three possible desired output values (50, 75 and 100 spikes from top to bottom). As mentioned above, the circuit underestimates the desired output for inputs that include full activity (100 spikes) for at least one of the two channels. For all three cases, about 20-25% of the runs result in an output that does not match the desired output value. While this behavior is consistent with physiological data, it has to be investigated how well a large-scale model of object recognition [30] will perform under these conditions.

2.3.2 Tuning

In the tuning configuration, the circuit shown in Fig. 2 effectively performs a one-dimensional tuning for each of the input channels, *i. e.,* each y unit will be maximally active for the input level it is tuned to and activity will decrease for lower or higher input values. The y unit activity is then combined to yield the output which is tuned in the multidimensional space. An example for the activity of a y unit can be seen in Fig. 4 (left). In this case, the circuit is tuned for an input of 50 spikes. Output activity is highest for a 50 spike packet and falls off on both sides. The tuning curve is not completely symmetric but approximates a Gaussian tuning curve well.

The output of a circuit tuned to the combination of two 50 spike packets as inputs is shown on the right of Fig. 4. The peak output activity is located at the (50,50) input pair and the response falls off in a roughly Gaussian (bell-shaped) fashion for any other input configuration.

2.4 Problems and future goals

- The main issue with the circuit of Fig. 2 is how well it can be extended to a significantly higher number of inputs and dimensions.
- The performance in the softmax configuration should be improved in order to meet the likely requirements of the recognition model (which are not stringent however).
- Learning the parameters that determine the Gaussian-like "centers" – that is the optimal stimulus for the tuning circuit – is an open problem.
- More careful measurements of the time required by the computation and of the bandwidth of the output activity are needed.

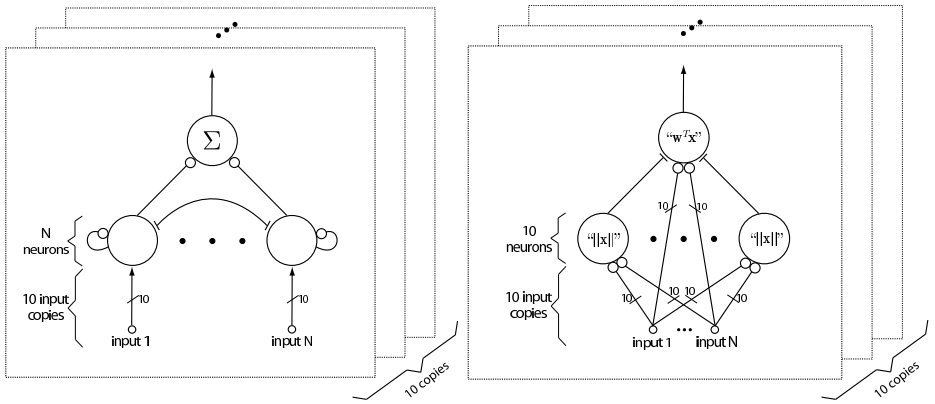

Figure 5: Maximum circuit architecture (left), and tuning circuit architecture (right), assuming Poisson distributed inputs. Small circles and bars denote excitatory and inhibitory connections respectively.

3 Poisson Spike Code Models

In this section we present independent spiking models for the maximum operation performed by complex V1 cells, and for normalized tuning observed in simple V1 cells. In both cases, and in the simulations which follow, we assume that the networks are driven by Poisson distributed spike inputs. For the maximum circuit, our goal is to output a sequence of spikes that encodes the maximal level of activity present at the circuit's inputs. In the case of the tuning operation, we wish to output a sequence of spikes which encodes some notion of the similarity between the input activity pattern and a preferred stimulus. We describe a circuit for which the similarity peaks when the input activity vector is collinear with the preferred stimulus, and falls off to zero along orthogonal directions in a Gaussian-like fashion. In the sections that follow, we will make these ideas more precise.

3.1 Max Circuit Architecture

3.1.1 Neural Dynamics

The maximum circuit design incorporates leaky integrate-and-fire models of neurons augmented with plausible synapse dynamics and an absolute refractory period. The choice of a first order linear model for the membrane dynamics was made in order to make large scale simulations involving many circuits computationally tractable. The maximum operation itself can be seen as a particular instance of K-winner-take-all behavior, and has been explored in networks of continuous nonlinear dynamical elements [2, 14, 15], and to a lesser extent, in networks of "spiking" elements [17, 23]. The use of dynamics which include reset rules, such as integrate-and-fire neurons, makes detailed analytical explorations of the behavior of interconnected networks of such elements difficult, and we do not carry out such an analysis here. Instead, we provide a description of the membrane potential of a single unit as it evolves in time, explain how individual elements should be connected together and tuned in order to compute the maximum over the input activities, and finally, show how multiple copies of circuits can be arranged to match the timing observed in cortex.

The particular winner-take-all design at the core of the maximum implementation exploits a balance between self-excitation and, in this case, all-to-all inhibitory dynamics. The connections are illustrated graphically in Figure 5 (left), where we denote excitatory feedback connections with small circles and inhibitory connections with bars. For additional computational simplicity, we do not include sign-changing interneurons in the inhibitory feedback paths.

We denote by V_L the neuron's resting potential, g_L the leakage conductance, and by g_j the synaptic conductances. With these definitions, the dynamics of each cell's subthreshold membrane potential $V_i(t)$

can be described by:

$$C_m \frac{dV_i}{dt} = g_L(V_L - V_i) + \sum_{j=1}^{N} g_j(V_j^r - V_i) \quad i = 1, ..., N. \tag{6}$$

The $g_j = g_j(t)$ are time-varying conductances that depend on activity at the incoming synapses, while C_m, g_L and V_L, are constants. The synaptic reversal potentials V_j^r take on one of two constant values $V_j^r \in \{V_{ex}, V_{inh}\}$ depending on whether the synapse is excitatory ($V_{ex} > V_\theta$) or inhibitory ($V_{inh} < V_\theta, V_{inh} \neq V_L$). The subthreshold dynamics (6) are combined with a nonlinear reset rule which stipulates that when the membrane potential $V_i(t)$ reaches the threshold V_θ, an action potential is fired and the potential is then immediately reset to the reset voltage $V_0 < V_\theta$. Following reset, an absolute refractory period is simulated by further suspending the subthreshold dynamics (6) for τ_{abs} seconds. Integration is then resumed with the initial condition $V_i(0) = V_0$.

We can rearrange the subthreshold dynamics (6) into the form

$$\tau_i \frac{dV_i}{dt} = V_i^\infty - V_i \tag{7}$$

where

$$\tau_i = \frac{C_m}{g_L + \sum_j g_j}, \quad \text{and} \quad V_i^\infty = \frac{g_L V_L + \sum_j g_j V_j^r}{g_L + \sum_j g_j}. \tag{8}$$

When viewed in this way, it is clear that the presence of time-varying synaptic inputs, in contrast to electrical feedback connections, implies that the total synaptic conductance as well as the membrane time constants τ_i, must depend on time via the input activity.

3.1.2 Synapse Dynamics

As an added degree of realism, the synapse conductances $\{g_i\}$ corresponding to excitatory feed-forward inputs are not scaled versions of the input spike trains, but are instead modeled as filtered (averaged) approximations. Given a discrete spike train $p_i(t)$, a post-synaptic current profile $\alpha(t)$, and a constant input conductance multiplier $\bar{g}_{in}$, synapse integration dynamics can be modeled as

$$g_i(t) = \bar{g}_{in}(\alpha * p_i)(t). \tag{9}$$

The change in conductance $\alpha(t)$ in response to an incoming spike is modeled as a decaying exponential with finite initial rise time. While this choice of post-synaptic conductance response is indeed more realistic than a simple decaying exponential of the form $e^{-t/\tau_{syn}}$, the finite rise time exponential was ultimately chosen because it closely approximates the detailed dynamics (4). In the simulations presented in section 3.3.2, we simply pre-compute the response and perform a table lookup, rather than simulate additional dynamics online. The particular parametrization we have chosen models the time course of the conductance as a difference of two exponentials [11]

$$\alpha(t) = \frac{1}{\tau_d - \tau_r} \left(e^{-t/\tau_d} - e^{-t/\tau_r} \right) \tag{10}$$

where the rise and decay time constants are denoted τ_r and τ_d respectively. Given this description of the conductance, the free parameters were tuned so that the shape of (10) matched as closely as possible the dynamics (4) given the parameters introduced in section 2. We find that $\tau_r = 0.4$ms and $\tau_d = 4$ms fits the dynamics for the AMPA-based excitatory response, and $\tau_r = 0.2$ms with $\tau_d = 5.4$ms fits the dynamics for the $GABA_A$-driven inhibitory response. In the absence of multiple spikes arriving in a short time interval, the channel dynamics (4) are well approximated by this more familiar "alpha function"[3]. Given the firing rate constraints described above, it is unlikely that multiple spikes will arrive within a short (e.g. 12.5ms) window.

In the interest of computational expediency, conductance changes for the excitatory inputs can be computed by convolving the kernel (10) with the input spike trains prior to simulation. Convolution

[3]We will refer Equation (10) as an alpha-function, even though it is not, strictly speaking, of the form $te^{-t/\tau}$.

amounts to summation of responses in the case where a spike arrives before the previous spike's alpha-function has decayed to zero. In the case of inhibitory synapses in the feedback paths of the circuit, however, it is often more convenient to follow a slightly different convention: If a new spike arrives before the previous spike's alpha function has decayed to zero, the two need not be added. Instead, the alpha function is "reset" to its initial value for that synapse. Resetting can be justified by making the assumption that an incoming spike causes most of the channels in the vicinity of the synapse to open, and that if another spike arrives, more than the maximum number of channels cannot open. Once again, because firing rates are limited to approximately 100Hz, we can assume that both choices will produce roughly equivalent behaviors.

With the input conductances computed beforehand, numerical integration of the system (7) is accomplished by discretizing time into finite steps δt, and applying an Euler update rule. The conductance response $\alpha(t)$ for inhibitory spikes that occur during the simulation can be incorporated into the integration by simply retrieving from a table the particular conductance along the curve (10) corresponding to the amount of time since the last spike arrived for the synapse of interest.

3.1.3 Circuit Organization & Simulation

As the stacked "planes" in Figure 5 (left) suggest, we combine multiple redundant copies of the circuit in order to (1) reliably estimate input activity and (2) compute the maximum within a short time window. Each individual "copy" of the circuit we have described is defined to have identical integrate-and-fire parameters, conductances, and architectures. The external inputs applied to each copy, however, will differ on a spike-by-spike basis and are not assumed to be synchronized, but will still have identical mean activities. Thus the outputs of the circuits will also not be synchronized, but will have identical average activities and mean times to the first spike. The collection of outputs taken from a group of circuits can then be fed into subsequent stages so as to enforce the required timing constraints. One can think of a group of such circuits as a larger meta-circuit capable of delivering an answer in a short amount of time. In this case, the larger circuit requires $K(2N+1)$ neurons, for K circuit copies and N inputs. One complex unit in the model [30] thus corresponds to this same quantity of integrate-and-fire neurons.

The physical cell parameters for the max circuit described in this section, and simulated in section 3.3.2, were chosen as follows: leakage conductance $g_L = 25$nS, feedback (self-)excitatory and inhibitory conductances $g_{ex} = 10g_L$ and $g_{inh} = 15g_L$, external (excitatory) input conductance $g_{in} = 4g_L$, refractory delay $\tau_{ref} = 8.5$ms, inhibitory and excitatory reversal potentials $V_{inh} = -80$mV and $V_{inh} = 0$mV, leakage potential $V_L = -70$mV, reset potential $V_0 = -70.25$mV, spike threshold potential $V_\theta = -54$mV, membrane capacitance $C_m = 0.5$nF, and maximum firing rate $F_{max} = 100$Hz.

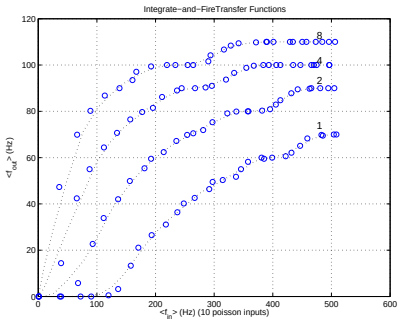

Figure 6: Integrate-and-fire spike-rate transfer functions for a variety of input conductances. Each trace represents the output of one neuron given 10 Poisson inputs. Conductance units shown on each curve are in multiples of a canonical conductance.

3.2 Circuit Architecture for Gaussian-like Tuning in the Direction of the Input Vector

3.2.1 Neural Dynamics

In this section we present a circuit which performs *normalized* tuning: given a preferred stimulus encoded by the strengths of the input synapses of the circuit, we would like the output activity of the network to peak when the input activity "vector" is collinear with the preferred stimulus, and fall off to zero, much like a Gaussian does, when the input and preferred stimulus vectors move towards orthogonality. In addition, the output activity should be normalized, in some way, by the total input activity. This normalization can take on a variety of possible forms, and, depending on its strength, leads to behaviors where the output activity of the circuit decreases or remains constant while the input activity increases, for a given fixed angle between the inputs and the preferred stimulus. While the tuning circuit presented in section 2 approximates a multidimensional Gaussian in the input space, the model presented in this section exhibits the latter aforementioned behavior. The tuning function is Gaussian-like for the direction of the normalized input vector with respect to a preferred stimulus, but is not shaped like a Gaussian in the input space because the output activity does not decrease when the total activity increases for a fixed angle between the input and the preferred stimulus vectors. Normalization is still, however, applied so that the total output activity of the circuit does not increase when input activity increases, but instead remains constant.

We approximate normalized tuning by delivering both divisive, normalizing inhibition and weighted excitatory signals to units operating in nonlinear regions of the spike-rate transfer function. We define the spike-rate "transfer function" to be the instantaneous output firing rate of a single integrate-and-fire neuron versus the sum of the instantaneous arrival rates of spikes at the unit's excitatory synapses, assuming no additional background input or current applied to the neuron. This transfer function is of course sensitive to the conductance of the input synapses, and we illustrate how different synaptic conductances can lead to different transfer function shapes in Figure 6.

One possible tuning circuit architecture is shown in Figure 5 (right), where units in the lower layer sum input activity and then inhibit a global pooling unit which also receives the original input. We take as our desired tuning model the normalized dot-product formulation described in [30]

$$y = g\left(\frac{\sum_j w_j x_j^p}{k + (\sum_j x_j^q)^r}\right) \tag{11}$$

where $g(\cdot)$ is a sigmoid nonlinearity, $\mathbf{w} = (w_1, ..., w_N)^T$ is the vector of synaptic strengths, and $\mathbf{x} = (x_1, ..., x_N)^T$ is the vector of inputs to the circuit. If the (integer-valued) exponents in (11) are chosen so that $p < qr$, then the output y will peak when the input $\mathbf{x}$ is "close" to the preferred stimulus encoded by $\mathbf{w}$, but will fall off as the total input activity increases while maintaining the same angle with the preferred stimulus vector. If $r = 1$ and $p \approx q$, then the output will peak when the input is close to the preferred stimulus, and will remain at the peak activity level if the total input activity increases (but will not increase any further as one would observe in the absence of any normalization).

In the spiking model described here, we attempt to roughly approximate exponentiation in the numerator and denominator terms in Equation (11) by choosing suitable operating points on the spike-rate transfer curves for units computing x^q, and separately, x^p. With r set to 1, $p \leq q$ means that neurons implementing the denominator in Equation (11) should exhibit locally steeper, more nonlinear transfer functions compared to numerator neurons. In the simulations that follow in section 3.3, the steepness of the initial rise in the spike-rate transfer function is controlled by adjusting the conductance of the inputs synapses only, however we found that even with a denominator conductance much larger than that of the numerator, the circuit behaves as if $p \approx q$ with $r = 1$. In this case, the normalization compensates for increasing input activity, but not to the extent that the total output activity begins to decrease, as discussed above: the "shape" of the tuning function in the input space is not a symmetric, multidimensional Gaussian as shown in section 2.

3.2.2 Circuit Organization & Simulation

After selecting parameters yielding suitable operating points and transfer functions, the circuit performs the tuning function in two feed-forward stages. First, the sum of the input activities is computed and represented by an output spike train 10 separate times. These 10 outputs are connected via shunting inhibitory synapses to a single pooling unit which also receives multiple copies of the input. The inputs at the pooling unit are each assigned specific conductance strengths which together encode a preferred stimulus. In the second stage of the circuit, the weighted sum of the input activity is combined with the effect of the inhibitory inputs in a divisive manner, giving normalized tuning at the output of the pooling unit. As before, multiple copies of the entire circuit are utilized in order to meet realistic timing requirements. One simple unit (S_n, for $n > 1$) in the model described by [30], would therefore correspond to the 10 integrate-and-fire output (pooling) neurons for each of the 10 circuit copies. In general, each tuning unit requires $K(P + 1)$ neurons, for K circuit copies and P inhibitory normalization neurons per circuit.

In the simulations presented in Section 3.3, we implement all units with integrate-and-fire neurons with refractory periods, synaptic conductance functions, and numerical simulation methods identical to those described in Section 3.1. We note, however, that unlike the maximum circuit, the tuning architecture is purely feed-forward and therefore allows for simplified computer simulation. In particular, the normalization and pooling components are decoupled and may be and computed independently, while the effect of the alpha-function on the synaptic conductances may be computed using convolution (as shown in Equation (9)) before numerical integration of the integrate-and-fire dynamics.

The physical cell parameters in the tuning circuit were chosen as follows: leakage conductance $g_L = 25{\times}10^{-9}S$, excitatory input conductance $g_{ex} = 8g_L$ for the lower row of normalization units and excitatory conductance $g_{ex} = 4g_L$ with inhibitory conductance $g_{inh} = 20g_L$ for the upper pooling unit, refractory delay $\tau_{ref} = 8.5$ms, and inhibitory reversal potential equal to the leak potential $V_{inh} = V_L = -70$mV. All other physical parameters were identical to those described in section 3.1.

3.3 Poisson Spike Code Simulations

3.3.1 Spike Process Model

In the simulations that follow, we use a Poisson distribution to model the spike arrival process. It is worth mentioning that the variance of a homogeneous Poisson spike process with rate parameter λ is also λ, and thus it is possible to occasionally generate biophysically unrealistic instantaneous firing rates in simulation. One advantage of this fact, however, is that the Poisson model leads to a conservative estimate as to the number of circuit copies necessary to achieve an accurate max computation within a small time window (by averaging over multiple circuits); restricting the variance of our spike processes can only improve simulation performance.

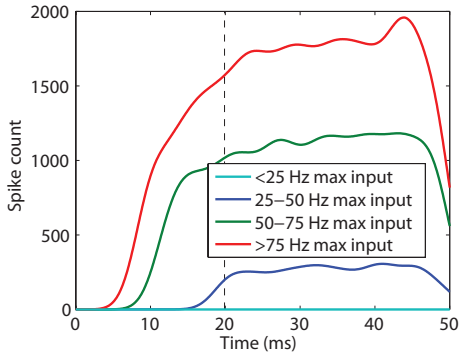

Figure 7: Max circuit time response simulation. (This figure courtesy Tim Masquelier.)

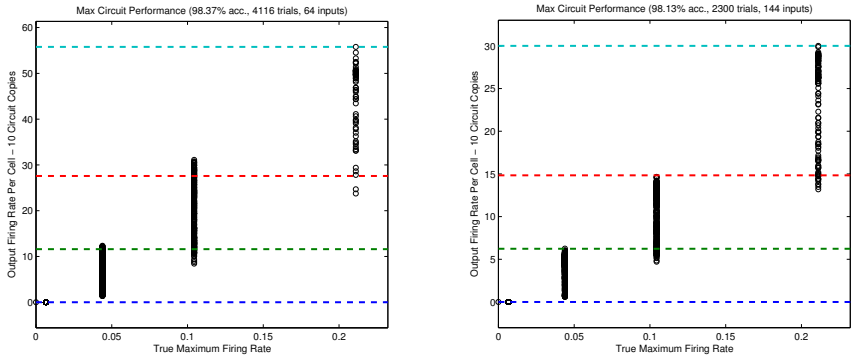

Figure 8: Max simulations with 64 (left) and 144 (right) inputs (over 2300 trials). The abscissa gives the true, desired activity level, while the ordinate gives the estimated instantaneous firing rate per output neuron at a point 20ms after stimulus onset. In all trials the circuit is simulated with Poisson spike inputs so that input activities are never exactly equal to the ideal discrete levels shown along the x-axis.

3.3.2 Maximum Circuit Simulations

To simulate the maximum circuit, we applied as inputs several sets of features returned by the $S1$ layer of the model described in [30]. These features were computed by filtering a set of natural images with oriented edge detectors at different scales, positions and orientations, and passing the result through a sigmoid nonlinearity. The feature sets ranged in size from 64 (8x8 patches) to 484 (22x22 patches), while the feature values were quantized into 4 levels, including "zero". Given the features, we generated 10 copies of each input by sampling 10 separate Poisson spike input trains with mean arrival rates proportional to the feature's value. Thus, for N inputs, for example, we generated $10N$ Poisson spike processes, and applied them to each input unit in each circuit copy. With 10 circuit copies, this corresponds to $100N$ separate input synapses. To evaluate the performance of the maximum circuit, we examine output spike rates as well as transient output characteristics (e.g. time to the first spike).

In Figure 7 we show the time course of the circuit's response to four 64-input stimuli in which the true maximum activity was one of the respective allowed input levels. Each trace represents the combined mean firing rate of 10 circuit copies estimated using a short Gaussian-shaped time window. The transient response of the circuit can be seen to carry a great deal of information, with larger input activities evoking earlier and larger responses. Figure 8 depicts the accuracy of the circuit given 64- and 144-input stimuli, over 2300 different stimulus instances. Each open circle marks the approximate instantaneous firing rate produced by each output neuron at a point 20ms following presentation of a distinct set of inputs. The instantaneous firing rate was estimated by computing the total output firing rate in the simulation time interval [15ms,25ms], and dividing by the total number of output neurons (64 or 144 of them, in this case) for all circuit copies (10 in these simulations). We then arrive at an instantaneous firing rate per output neuron, where there are $N \times K$ outputs if the dimensionality of the input is N and the number of circuit copies is K. Finally, because the circuit was presented with stochastic Poisson spike trains with mean spike rate equal to one of the levels of activity specified by the quantized image features, it should be noted that there is significant variance in the firing rates actually delivered to the circuit.

In order to score the accuracy of the maximum circuit, we divided up the range of the output firing rates into 4 bins that proportionally preserve the division of the input domain into its 4 levels. Using these bins, we then counted the number of instances where the output did not fall into the correct output bin, as determined by the bin of the maximum input. Thus if the output firing rate falls into the correct corresponding output bin, it is deemed a successful trial. Over all 2300 trials, only a small number did not meet this criteria: in 98.48% of the 64-input trials and 98.13% of the 144-input trials, the correct maximum bin was produced. We have also experimented with larger input sets, of size 256 and 484, and found that accuracy scales well with the number of inputs. The slight decrease in accuracy with input dimensionality seen in the above figures can be explained by the fact that more inputs offer more of an

opportunity for there to be repeated inputs which take on the maximum value. Because the winner-take-all dynamics involve noisy spike processes rather than continuous quantities, if there are several inputs close to the maximum value, not all of them will be suppressed all of the time. This situation can lead to exaggerated spike rates and results in a response that occasionally falls into a higher level bin than appropriate.

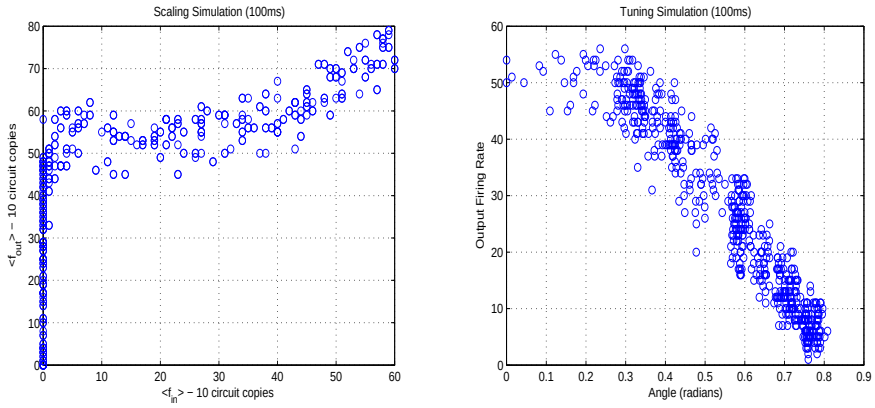

Figure 9: Simulations illustrating scaling (left) and tuning (right) properties of the tuning circuit shown in Fig. 5 (right), with 10 circuit copies and 64 inputs. Although we have plotted the circuit response for only positive angles in the right-hand panel, the tuning profile can be considered Gaussian-like because it is symmetric in the angle.

3.3.3 Tuning Circuit Simulations

In this section we describe experiments which illustrate the behavior of the tuning circuit shown in Figure 5 (right). In simulating this circuit, care must be taken to ensure that inputs and outputs maintain realistic firing rates. This constraint, however, complicates testing the tuning properties of the circuit because constraints on the activities of the inputs translate into constraints on the way in which we are able to test the circuit given specific desired angles between the input and the preferred stimulus. In particular, we would like to choose a set of input activities $\left\{\mathbf{x}^j = (x_i^j, \ldots, x_N^j)^T\right\}_{j=1}^J$ that sweep out a range of angles $\{\theta_1 = 0, \ldots, \theta_J = \pi/2\}$ while satisfying

$$0 \leq x_i^j \leq F_{max}, \quad \text{and} \quad \sum_{i=1}^{N} x_i^j = M \tag{12}$$

where N is the number of inputs to the circuit, M is a fixed, total level of activity, F_{max} is the maximum allowed firing rate for a single neuron, and J is the chosen number of input instances that we will apply to the circuit to test its behavior. The second (normalization) constraint in (12) is imposed when testing the tuning properties in order to fix the operating point along the scaling curve. [4] If the total activities changed over the input vectors, then the tuning properties would change, and responses for different angles would not be directly comparable. Conversely, when testing the scaling properties of the tuning circuit, the angle between the inputs and the preferred stimulus vector must be fixed while varying the total activity over a chosen range.

To select vectors for testing the tuning behavior of the circuit, we set the first input vector in a set of inputs equal to the preferred stimulus, $\mathbf{x}^0 = \mathbf{w}$, and then generate a sequence of J vectors that successively drift away from collinearity with the weight vector by applying the following sampling process: At each

[4] We stress however that normalized inputs are *not* required during ordinary operation of the circuit; we only normalize the total activity here to evaluate tuning vs. input angle while controlling for overall input activity.

iteration t, we begin by selecting two random components from $\mathbf{x}^{(t-1)}$, $x_u^{(t-1)}$ and $x_v^{(t-1)}$. Given a small but fixed adjustment amount Δ, we generate the new vector $\mathbf{x}^{(t)}$ by applying the following updates:

$$x_i^{(t)} = x_i^{(t-1)}, \;\; i \neq u, v \tag{13a}$$

$$x_u^{(t)} = x_u^{(t-1)} + \Delta \tag{13b}$$

$$x_v^{(t)} = x_v^{(t-1)} - \Delta \tag{13c}$$

If a component falls outside of the range $0 \leq x_i \leq F_{max}$, then the candidate vector is rejected and a new one is resampled. The preferred stimulus was chosen randomly, but fixed for the duration of the analysis. The above procedure works well for finding vectors that slowly travel from collinear to angles near $\pi/4$ with respect to the weight vector when using the convention that we always take the smallest positive solution θ to $\theta = \cos^{-1}(z)$ for $z \geq 0$. [5]. To generate inputs near orthogonality, we searched for a vector $\mathbf{v} \perp \mathbf{w}$ by solving the constrained optimization problem

$$\mathbf{v} = \arg\min_{\mathbf{x}} \left\{ \cos^{-1}\left(\frac{\mathbf{w}^T \mathbf{x}}{\|\mathbf{w}\| \|\mathbf{x}\|} \right) - \frac{\pi}{2} \right\}^2$$

subject to the constraints (12), and again using the convention that we select the smallest positive angle satisfying $\theta = \cos^{-1}(z)$. We then applied the sampling procedure (13) to $\mathbf{v}$ and generated a sequence of inputs whose angles with the preferred stimulus drifted from orthogonal to approximately $\pi/4$. Finally, the scalar components of the activity vectors found above were used as mean arrival rates to sample Poisson spike input trains that were then applied to the circuit during simulation. Vectors for the scaling simulation were generated by fixing the angle to be collinear with the preferred stimulus for all test inputs, and simply varying the length of the test vector over a range which ensured that the individual activities did not exceed the maximum firing rate. In order to evaluate the performance of the circuit over a wide range of conditions, we did not constrain the scalar input vector components or Poisson arrival rates to take on one of the 4 allowed values represented in the model [30].

In Figure 9 we show the behavior of the circuit with fixed input angle over a range of activities (left panel), and with fixed activity over a range of input angles (right panel). For fixed input activities, it is clear that the circuit exhibits an appropriate fall-off in output activity as the input moves away from the preferred stimulus in angle. We have shown only the positive angles corresponding to each response, however the tuning curve is symmetric for negative angles as well, and thus approximates a Gaussian.

If the circuit is performing a *normalized* dot-product, as opposed to the canonical dot-product, then regardless of the magnitude of the input, and for a fixed angle, the response should either decrease or stay roughly constant depending on the choice of the exponents p, q and r in Equation (11). The left-hand panel in Figure 9 verifies that this is approximately the case over a wide range of input activities: when the average input firing rate is in between 5 and 20Hz, the output firing rate decreases slightly. Between 20 and 40Hz the output remains constant on average. Beyond 40Hz, the operating points of the neurons in the circuit become shifted into a regime where the divisive normalization no longer over-compensates for the "length" of the input, and we see a proportional increase in the output firing rate with respect to the input firing rate. Whether tuning in the higher levels of visual cortex takes the shape of a multi-dimensional Gaussian or not is still, however, under discussion. Several researchers have identified and modeled cells exhibiting a wide range of contrast normalization and tuning characteristics, including units whose output activity decreases or remains constant as a function of input activity [4, 12].

In all simulations, we assumed 10 circuit copies, and 10 input copies, 64 inputs, and took as "output" the sum of the average spike rates of the 10 circuit outputs over the first 50ms of the simulation. The dynamics were integrated using Euler step sizes $\delta = 0.1$ms.

3.4 Problems and Future Goals

- The tuning circuit presented above may be made more flexible and possibly configured to behave like a true Gaussian function in the input space through a more careful evaluation of the trade-off

[5] Normalized random vectors with positive uniformly distributed components form an absolute angle (modulo 2π) near $\pi/4$ with the optimal stimulus on average, and will rarely form angles near $\pi/2$ or 0 radians. For this reason, we sample input vectors starting at 0 radians moving to $\pi/4$, and then from $\pi/2$ back to $\pi/4$.

between conductances in neurons implementing the numerator and denominator of Equation (11).

- The resolution of the output and the time delay from input presentation to stable outputs in both circuits may be possibly improved by adding additional circuit copies and adjusting the integrate-and-fire membrane time constants (subject to biophysical constraints).
- The potential role of feedback in the tuning circuit should be explored. It is likely that feedback connections can be used to modulate the shape of the Gaussian-like tuning function by sharpening neural responses, and (separately) might facilitate normalization if used in a gain-control configuration.
- Learning the parameters that determine the Gaussian-like "centers" – that is the optimal stimulus for the tuning circuit – is also an open problem in this case.
- More careful measurements of the time required by the computation and of the bandwidth of the output activity are needed.

4 Discussion

We presented two different sets of circuits which implement the maximum and tuning operations under two different coding schemes. The circuits serve as a proof of concept that these computations can be performed by spiking neural circuits in the cortex.

We introduced a coding scheme in section 2 that is compatible with the spontaneous and evoked firing rates observed in rat barrel cortex and presented a canonical microcircuit that is able to perform both the max and the tuning operations, depending on different values for its synaptic weights. This not only argues for the plausibility of the model of object recognition presented in [30], but also provides support for the idea of a canonical microcircuit present in multiple (at least sensory) cortical areas [8]. The main architecture of the circuit follows the observation of monosynaptic excitation combined with disynaptic inhibition found in layer IV of primary sensory areas. Although this connectivity has mostly been investigated in the thalamo-cortical circuit, it is conceivable that higher cortical areas might follow the same architecture, even though they receive their inputs from other cortical areas instead of thalamic nuclei. Both the max and tuning configurations of the presented microcircuit exhibit trial-to-trial variability. Whether this variability is feasible for high-level computational models such as [30] needs to be investigated, however, it appears to be compatible with the variability of neurons in cortex. More stringent characterization of the variability in both cases is needed in order to quantitatively compare models and experimental data.

Under the assumption that the peak bandwidth for communication between visual areas is limited to approximately 2 bits of information over time intervals of approximatively 10-20 ms, the maximum architecture described in section 3.1 is capable of accommodating large numbers of inputs while maintaining high accuracy. Because the circuit was designed to operate given Poisson distributed inputs, it is relatively robust to noise in the form of spike addition, deletion, and timing jitter. The use of multiple redundant circuit copies is a critical feature that facilitates computation which satisfies biophysical timing and resolution constraints. Thus, several circuits may be connected together in series, as would be required by the model in [30], while maintaining realistic timing properties. The tuning circuit presented in section 3.2 was shown to have the desired tuning and normalization characteristics, and can also scale to accommodate large numbers of inputs. As in the case of the max circuit, multiple tuning circuit copies were utilized in order to satisfy the stringent timing and resolution constraints imposed by the neuroscience of object recognition.

There are several important assumptions underlying the circuits presented in this paper that must be considered when attempting to connect spiking artificial circuits to circuits in cortex. The winner-take-all configuration of the max circuit in section 3.1 requires all-to-all inhibition between units within a circuit copy, as well as some mechanism for self-excitation. Chemical positive feedback loops are probably more likely to be found than autapses, if this circuit is implemented in cortex. All-to-all inhibitory connectivity is an idealization that simplifies numerical and theoretical analyses, but is less likely to be found in the brain than, for example, dense but random connectivity. The winner-take-all computation is, however,

thought to be ubiquitous in the brain, possibly underlying some aspects of attention and decision making. Indeed, there are several network designs that exhibit winner-take-all behavior, including networks with a single global inhibitory neuron instead of all-to-all inhibition [9]. It is therefore likely that the all-to-all requirement in this max circuit (section 3.1) can be relaxed, while maintaining the desired performance; the circuit we have presented relies on winner-take-all computations, but does not critically depend on the particular instantiation of winner-take-all shown in Figure 5.

In the tuning circuit discussed in section 3.2, divisive, "shunting" inhibition was chosen as the normalization mechanism, while the selection of an operating point in a nonlinear region of the units' spike-rate transfer function was used to approximate the effect of the exponents in the normalized dot-product (11). Although several alternatives exist for both division and multiplication in spiking neural networks, we believe the choices we have made are among the most plausible in light of the design constraints. If one adopts the reasonable position that evolution has led to solutions which minimize both energy consumption and real-estate in the brain, the architecture discussed in section 3.2 can be seen to minimize the number of units necessary to accomplish normalization and exponentiation. It could also be that single cells are capable of performing normalized tuning using intrinsic mechanisms, in which case the number of cells in a cortical implementation is likely to be smaller than the number used in our circuit of simplistic integrate-and-fire units.

In order to improve the match with physiological data, we plan to extend our models to be more faithful to the morphology, biophysical properties and connectivity of different specific subtypes of cortical neurons in layers IV and II/III of somatosensory and visual cortices, utilizing the limited quantitative data available from other researchers [10, 19] and our own recordings.

In order to decide which of the presented models and variations, if any, best describes biophysical reality, new experiments are needed to help distinguish between the alternatives. It would be interesting to change the timing of the presented stimuli to test the temporal precision that is needed for the maximum and tuning effects to occur, and what kind of behavior can be observed outside of that precision window. In addition, experiments should try to extend from two to three or even more inputs (stimuli). This is difficult for visual stimuli because of the small receptive field sizes of the cells under investigation but it would help to tease apart different possible mechanisms as their behavior for more than two inputs can be significantly different. The rat vibrissae and barrel cortex system provides a suitable preparation for this task since it is much easier to stimulate more than two vibrissae at the same time. Finally, it is quite possible, despite the general skepticism of cortical physiologists, that the two operations described here may be performed by *circuits of mostly non-spiking neurons*, as proposed in [30]. Such models may need to be reconsidered and improved.

Acknowledgements

The authors would like to thank Timothe Masquelier for help with the experiments in section 3.3.2, and Thomas Serre, Christof Koch, Michael Okun and Ilan Lampl for helpful discussions and suggestions.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, the Department of Brain & Cognitive Sciences, and the Computer Sciences & Artificial Intelligence Laboratory (CSAIL). This research was sponsored by grants from: the US Defense Advanced Research Projects Agency, US Office of Naval Research, US National Science Foundation, - National Institutes of Health (NIMH). Additional support was provided by: Daimler-Chrysler AG, Eastman Kodak Company, Honda Research Institute USA, Inc., Komatsu Ltd., Oxygen, Siemens Corporate Research, Inc., Sony, Sumitomo Metal Industries, Toyota Motor Corporation, and the Eugene McDermott Foundation.

References

[1] Abbott, L. F., Varela, J. A., Sen, K. & Nelson, S. B. Synaptic depression and cortical gain control. *Science* **275**, 220–224 (1997).

[2] Amari, S.-I. & Arbib, M.A. Competition and cooperation in neural nets. In Metzler, J. (ed.) *Systems Neuroscience*, 119–165 (Academic Press, 1977).

[3] Carandini, M. & Heeger, D. J. Summation and division by neurons in primate visual cortex. *Science* **264**, 1333–1336 (1994).

[4] Carandini, M., Heeger, D. J. & Movshon, J. A. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* **17**, 8621–8644 (1997).

[5] Chelazzi, L., Duncan, J., Miller, E. K. & Desimone, R. Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophys.* **80**, 2918–2940 (1998).

[6] Destexhe, A. & Contreras, D. Neuronal computations with stochastic network states. *Science* **314**, 85–90 (2006).

[7] Destexhe, A., Mainen, Z. F. & Sejnowski, T. J. Kinetic models of synaptic transmission. In Segev, I. & Koch, C. (eds.) *Methods in Neuronal Modeling: From Ions to Networks*, 1–26 (MIT Press, 1998).

[8] Douglas, R. J. & Martin, K. A. C. Neuronal circuits of the neocortex. *Annu. Rev. Neurosci.* **27**, 419–451 (2004).

[9] Ermentrout, B. Complex dynamics in winner-take-all neural nets with slow inhibition. *Neural Networks* **5**, 415–431 (1992).

[10] Gawne, T. J. & Martin, J. M. Responses of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophys.* **88**, 1128–1135 (2002).

[11] Gerstner, W. & Kistler, W.M. *Spiking Neuron Models* (Cambridge University Press, 2002).

[12] Giese, M. A. & Leopold, D. A. Physiologically inspired model for the ecoding of face spaces. *Neurocomputing* **65-66** (2005).

[13] Grossberg, S. Contour enhancement, short term memory, and constancies in reverbarating neural networks. *Studies in Applied Mathematics* **52**, 213–257 (1973).

[14] Hahnloser, R H R, Seung, H.S. & Slotine, J.-J. Permitted and forbidden sets in symmetric threshold-linear networks. *Neural Comp.* **15** (2003).

[15] Hartline, H K & Ratliff, F. Spatial summation of inhibitory influences in the eye of limulus, and the mutual interaction of receptor units. *Journal of General Physiology* **41**, 1049–1066 (1957).

[16] Hung, C., Kreiman, G., Poggio, T. & DiCarlo, J. Fast read-out of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).

[17] Jin, D.Z. & Seung, H.S. Fast computation with spikes in a recurrent neural network. *Physical Review E* **65** (2002).

[18] Kouh, M. & Poggio, T. A canonical cortical circuit for gaussian-like and max-like operations. *In submission* (2007).

[19] Lampl, I., Ferster, D., Poggio, T. & Riesenhuber, M. Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J. Neurophys.* **92**, 2704–2713 (2004).

[20] Lennie, P. Single units and visual cortical organization. *Perception* **27**, 889–935 (1998).

[21] Moldakarimov, Samat, Rollenhagen, Julianne E, Olson, Carl R & Chow, Carson C. Competitive dynamics in cortical responses to visual stimuli. *J. Neurophys.* **94**, 3388–3396 (2005).

[22] Okun, Michael & Lampl, Ilan. Synchronized excitation and inhibition during spontaneous and evoked response in the barrel cortex. In *Computational and Systems Neuroscience* (Salt Lake City, UT, 2007).

[23] Oster, M. & Liu, S.-C. Spiking inputs to a winner-take-all network. In Y. Weiss, B. Schlkopf & Platt, J. (eds.) *Advances in Neural Information Processing Systems*, vol. 18, 1051–58 (MIT Press, Cambridge, MA, 2005).

[24] Pinto, D J, Brumberg, J C & Simons, D J. Circuit dynamics and coding strategies in rodent somatosensory cortex. *J. Neurophys.* **83**, 1158–1166 (2000).

[25] Poggio, T. Stochastic linearization, central limit theorem and linearity in (nervous) "black-boxes". *In: Atti of III Congresso Nazionale di Cibernetica E Biofisica* 349–358 (1975).

[26] Poggio, T., Reichardt, W. & Hausen, W. A neuronal circuitry for relative movement discrimination by the visual system of the fly. *Network* **68**, 443–466 (1981).

[27] Reichardt, W., Poggio, T. & Hausen, K. Figure-ground discrimination by relative movement in the visual system of the fly – II: Towards the neural circuitry. *Biol. Cyb.* **46**, 1–30 (1983).

[28] Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nature Neurosci.* **2**, 1019–1025 (1999).

[29] Sclar, G., Maunsell, J. H. & Lennie, P. Coding of image contrast in central visual pathways of the macaque monkey. *Vision Res.* **30**, 1–10 (1990).

[30] Serre, T. *et al.* A theory of object recognition: computations and circuits in the feedforward path of the ventral stream in primate visual cortex. AI Memo 2005-036 / CBCL Memo 259, MIT CSAIL and CBCL, Cambridge, MA (2005).

[31] Swadlow, H A. Efferent neurons and suspected interneurons in S-1 vibrissa cortex of the awake rabbit: receptive fields and axonal properties. *J. Neurophys.* **62**, 288–308 (1989).

[32] Vogels, Tim P & Abbott, L F. Signal propagation and logic gating in networks of integrate-and-fire neurons. *J. Neurosci.* **25**, 10786–10795 (2005).

[33] Wehr, Michael & Zador, Anthony M. Balanced inhibition underlies tuning and sharpens spike timing in auditory cortex. *Nature* **426**, 442–446 (2003).

[34] Wilent, W Bryan & Contreras, Diego. Dynamics of excitation and inhibition underlying stimulus selectivity in rat somatosensory cortex. *Nature Neurosci.* **8**, 1364–1370 (2005).

[35] Woolsey, T A & der Loos, H Van. The structural organization of layer IV in the somatosensory region (SI) of mouse cerebral cortex. The description of a cortical field composed of discrete cytoarchitectonic units. *Brain Res.* **17**, 205–242 (1970).

[36] Yoshimura, Yumiko & Callaway, Edward M. Fine-scale specificity of cortical networks depends on inhibitory cell type and connectivity. *Nature Neurosci.* **8**, 1552–1559 (2005).

[37] Yoshimura, Yumiko, Dantzker, Jami L M & Callaway, Edward M. Excitatory cortical neurons form fine-scale functional networks. *Nature* **433**, 868–873 (2005).

[38] Yu, A. J., Giese, M. A. & Poggio, T. Biophysiologically plausible implementations of the maximum operation. *Neural Comp.* **14**, 2857–2881 (2002).

[39] Zhu, Yinghua, Stornetta, Ruth L & Zhu, J Julius. Chandelier cells control excessive cortical excitation: characteristics of whisker-evoked synaptic responses of layer 2/3 nonpyramidal and pyramidal neurons. *J. Neurosci.* **24**, 5101–5108 (2004).