Chapter 14. Self-consistent Models of Cortical Dynamics

Anthony Zador
Cold Spring Harbor Labs, Long Island, USA

Abstract. The structural homogeneity of the neocortex suggests that different regions may be executing similar computations, and motivates the search for a general theory of cortical operation. A basic prerequisite of any such theory is that it be self-consistent. Because the majority of cortical input is intracortical, self-consistency requires that, if a cortical neuron is selected at random, its firing rate should reflect that of its neighbors. One apparent violation of self-consistency arises in the high-gain formulation (Shadlen and Newsome 1998) of a simple model of cortical operation. In this formulation, a typical cortical neuron integrates synaptic inputs from dozens or hundreds of other independently active neurons, but only a few dozen active excitatory inputs are sufficient to saturate the firing rate; the proposed resolution is that inhibitory input counterbalances the excitation. We argue instead that the experimental evidence favors a low-gain model of firing in cortical neurons. That is, the number of presynaptic firing events required to sustain a high firing rate is much larger—so large, in fact, that activity can only be sustained when synaptic input is organized into synchronous volleys. This alternative model implies that the basic unit of cortical operation may be a synchronous volley, and that these volleys may represent the target for the modulatory effects of attention and other cognitive processes.

The structural homogeneity of the neocortex is striking. While the retina is highly specialized for the transduction of light energy into neural impulses, and the cochlea for sound energy, the corresponding visual and auditory cortices are remarkably similar. It is hard to distinguish a slab of visual cortex from a slice of auditory cortex based on the morphological or pharmacological characteristics of the constituent neurons. This homogeneity suggests that different cortical regions may be executing similar computations, and motivates the search for a general theory of cortical operation [7, 13, 19]. Such theories make explicit the often implicit assumptions that guide research. A basic prerequisite of any theory is, of course, that it be self-consistent; inconsistencies underscore gaps in our understanding.

The majority of cortical input is intracortical, so that the input to one cortical neuron consists largely of the output from the others. According to the simplest "standard" model of cortical circuitry, during each spike a typical cortical neuron integrates synaptic inputs from dozens or hundreds of other independently active neurons. Surprisingly, this simple model is hard to reconcile with the irregular firing of cortical neurons [17]. The conundrum is as follows. Responses by cortical neurons are noisy. The spike train elicited by a sensory stimulus is highly irregular, more like the ticks of a Geiger counter than of a clock. Each of these inputs is itself an irregular spike train, generated by other typical neurons. Thus in the standard model, the input to a neuron is a constant shower of excitatory and inhibitory postsynaptic potentials (EPSPs and IPSPs, respectively). However, because the spike-generating mechanism performs a temporal
integration, it tends to average out the inputs, so the output spike train generated by each neuron should, by the central limit theorem, be much more regular than its input. This standard model of cortical operation is, then, internally inconsistent, since the outputs cannot be explained by the inputs. The vigorous debate [14, 15, 17, 20, 21, 23] generated in the attempts to resolve this apparent conundrum has led to the formulation of competing theories which make different predictions about the neural signals out of which cortical representations are constructed.

This simple model of cortical activity appears plagued by a second internal inconsistency, sometimes referred to as the dynamic range problem [18]. Dynamic range is a problem if the input-output function of a typical cortical neuron is assumed to have high gain. The problem arising from the high-gain model of cortical firing is illustrated by the following simple calculation. Cortical neurons receive as many as 10,000 synaptic contacts, the majority of which are presumed excitatory [2]. Intracellular electrophysiological recordings suggest that the postsynaptic response due to firing a single presynaptic neuron can, in some cases, depolarize the postsynaptic membrane by as much as 3-10% of the difference between the rest potential and the spike threshold, implying that as few as 10-40 EPSPs are sufficient to cause a spike. If such connections were typical, then a very small number of active afferent connections could saturate the neuron's output — could drive it to fire at its maximum possible rate.

Self-consistency requires that, if a cortical neuron is selected at random, its firing rate should reflect that of its neighbors. For example, if a neuron fires at 20 Hz, then on average so too should the neurons that provide its input. However, the high-gain model predicts that as few as 100 active afferents (just a few percent of the 3000-10,000 available), each firing at only 20 Hz, would cause an output firing rate of 50-200 Hz. Higher input firing rates, or a higher fraction of active afferents, would therefore completely saturate the neuron (see Fig. 1a). Self-consistency is violated.

How do cortical neurons avoid this saturation? What assumptions must be modified in this simple model to achieve self-consistency? One widely held view is that the inhibitory component of synaptic activity, neglected in the simple model discussed above, is comparable in size to the excitatory component and just sufficient to offset the excitation [14, 15, 21]. This raises the new question of how the excitation and inhibition are achieved; but recent theoretical proposals have addressed this issue [18].

Part of the appeal of the balanced excitation-inhibition model is that it also provides a resolution of the conundrum of irregular firing. The balance of excitation and inhibition that drives cortical neurons consists of large fluctuations, causing the neuron to fire irregularly. In this view, the irregular firing of cortical neurons is a consequence of the very large amounts of inhibition they receive. It appears, then, that the balanced-inhibition model can resolve two seemingly unrelated cortical conundras.

Appealing though the balanced-inhibition model is, we argue here that closer scrutiny of the experimental data leads to an alternative view of cortical dynamics. We argue that the experimental evidence favors a low-gain model of firing in cortical neurons. That is, the number of presynaptic firing events required to sustain a high firing rate is large — more than 10-fold higher than assumed in the high gain model.
in generated by each neuron input. This standard model of cannot be explained by the the attempts to resolve this theories which make different ons are constructed.

ed by a second internal n [18]. Dynamic range is a assumed to have high gain. illustrated by the following aptic contacts, the majority logical recordings suggest neuron can, in some cases, difference between the rest 8Ps are sufficient to cause number of active afferent it to fire at its maximum

d at random, its firing rate 20 Hz, then on average so gain model predicts that as 0 available), each firing at her input firing rates, or a turate the neuron (see Fig.

usions must be modified in view is that the inhibitory issed above, is comparable excitation [14,15,21]. This are achieved; but recent
del is that it also provides a citation and inhibition that zuron to fire irregularly. In f the very large amounts of on model can resolve two here that closer scrutiny of namics. We argue that the neurons. That is, the number large — more than 10-fold

Fig 1. The dynamic range problem. (A) Output firing rate is plotted as a function of the average firing rate of the inputs for a simple model of cortical dynamics in which the output firing rate is linearly related to the input firing rate with a slope given by $m = n / p$ (see Results and Table 1). The output firing rate is much higher than the input rate, even when only a moderate number of input afferents is active. The reference trace indicates the input-output slope $m = 1$ required by a perfectly self-consistent model of cortical activity; dashed line: 1% of input afferents active; dotted line: 5% input afferents active. Parameters: 10 EPSPs/spike; 10,000 inputs; see Table 1, “High-gain model” for details.

1. Simple input-output model of a cortical circuit

A self-consistent model of cortical activity requires models of both the neuronal input-output function and of the functional characteristics of cortical circuits. In the simplest model of neuronal spiking, the output firing rate $F_o$ depends linearly on the input activity $F_i$ as

$$F_o = m F_i$$

(1)

where $m$, the input-output slope, depends in turn on parameters related to the cortical circuit (see Table 1 and below). In the simplest cortical model, each neuron is exactly like every other neuron, and so each output neuron is just like every input neuron. In such a model, then, self-consistency requires that the output firing rate $F_o$ equal the input firing rates $F_i$.

$$F_o = F_i$$

(2)

Eq. 2, which specifies the self-consistency condition, is satisfied only when the input-output slope is exactly unity, $m = 1$.

This “spherical cow” model is of course highly oversimplified—the cortex is neither perfectly homogenous nor linear—so small deviations of the input-output slope $m$ from unity are to be expected. However, a dramatic deviation from unity suggests deeper problems. In

---

1 The reference here is to the old joke about a theoretical physicist who was charged with figuring out how to increase the milk production of cows. Although many farmers, biologists, and psychologists had tried and failed to solve the problem before him, the physicist had no trouble coming up with a solution on the spot. “First,” he began, “we assume a spherical cow…”
particular, if the slope is much greater than unity, \( m >> 1 \), then the output neuron's firing rate \( F_o \) saturates for even low levels of input activity \( F_i \).

Dynamic range thus depends on the slope \( m \), which relates input to output firing rates. This slope depends on four parameters: the number of synaptic inputs \( N \), the fraction of neurons \( f \) activated by a sensory stimulus, the average number \( Z \) of excitatory postsynaptic currents (EPSCs) per action potential, and the synaptic release probability \( p \):

\[
m = N f p / Z.
\]

The internal consistency of the model thus depends on the value of these parameters. We therefore consider below what reasonable experimental estimates for these values might be.

2. Parameters of low-gain vs. high-gain firing

We consider first the parameter \( f \), the fraction of inputs activated during a typical stimulus. This fraction \( f \) is very hard to measure directly, but it has been estimated to be as large as 5-10% or larger [16]. The difficulty arises in part because conventional extracellular recording methods are limited to neurons that show stimulus-locked spiking, and therefore cannot be used to estimate the denominator—the number of neurons not activated by a stimulus; but a recent study using intracellular methods reported that about 75% of neurons in the cat primary visual cortex responded with a firing rate of 1 Hz or more [3].

Next, we consider the release probability, \( p \), which is the probability that a synapse will release neurotransmitter given that an action potential has invaded the nerve terminal. The high-gain model implicitly assumes perfect synaptic reliability (\( p=1 \)). Release probability has, however, been shown to be appreciably less than unity at neocortical [4, 1] and hippocampal synapses [6, 9, 11] in vitro, although it has not yet been possible to measure in vivo. Release probability is different at different synapses and ranges from \( p<0.1 \) to \( p=1 \); for the low-gain model we choose \( p=0.5 \), although the actual value is probably lower. Moreover, synaptic depression—the use-dependent decrease in release probability during periods of high presynaptic firing—reduces the release probability still lower; we defer discussion of this effect (see Synaptic depression below).

The third parameter we consider is the number of excitatory synaptic contacts, \( N \). Anatomical studies place this number between 3000 and 10,000 [2]. When interpreting this number, it should be remembered that a presynaptic neuron in the neocortex often makes several contacts onto its postsynaptic target (Fig. 2). For example, in one study of nearby pairs in layer 5, the average connectivity was 5.5 synapses within each pair, with a maximum of eight [10]. This means that estimates of the physiological connection strength between neuron pairs are based on the action of multiple boutons, rather than on a unitary synaptic input.

The final parameter, \( Z \), is the average number of EPSCs required to produce a spike. The parameter \( Z \) depends on both the size of an EPSC and the properties of the spike generating mechanism. One way to estimate \( Z \) is to assume that the neuron has a voltage threshold, and to ask how many EPSPs must be summed in order to reach threshold. Extrapolating from experimental evidence obtained from cortical and hippocampal [12] slices, Shadlen and Newsome [16] estimate that "the excitatory drive from only 10-40 inputs, discharging at an average rate of 100 spikes/second, should cause the postsynaptic neuron to discharge near 100 spikes/second." Based on this estimate, the number of (temporally coincident) EPSPs required to achieve threshold is 10-40 (see high-gain model, Table 1).
inputs activated during a typical period has been estimated to be as cause conventional extracellular stimulus-locked spiking, and therefore of neurons not activated by a delay that about 75% of neurons in Hz or more [3].

is the probability that a synapse has invaded the nerve terminal. 
probability ($p=1$). Release probability unity at neocortical [4, 4] and 
yet been possible to measure in all ranges from $p<0.1$ to $p=1$; for 
the is probably lower. Moreover, probability during periods of higher 
probability decreases during periods of low probability. We defer discussion of this
excitatory synaptic contacts, $N$. 0,000 [2]. When interpreting this 
information in the neocortex often makes example, in one study of nearby 
fields in each pair, with a maximum 
connection of strengths between 
rather than a unitary synaptic 
SCs required to produce a spike. 
and the properties of the spike 
that the neuron has a 
voltag 
ed in order to reach threshold. 
ical and hippocampal [12] slices, 
drive from only 10-40 inputs, 
cause the postsynaptic neuron to 
tte, the number of (temporarily 
high-gain model, Table 1).

Fig 2. Neurons in cortex can be multiply connected. On the left, a cortical neuron makes 
only a single synapse onto a neighboring neuron. On the right, a cortical neuron makes 
multiple (in this case, four) synapses onto its target. Connections mediated by multiple 
synapses are common in the cortex, and probably account for the large EPSPs occasionally 
recorded between neuronal pairs in cortical slices.

Most of the experimental evidence on which this estimate is based involves paired 
recordings between connected neurons in the cortex. However, as emphasized by Fig. 2, such 
connections are often mediated by multiple synapses. Since $N$ is the total number of synapses 
on the postsynaptic neuron, rather than the number of presynaptic neurons that make 
connections with the postsynaptic neuron, $Z$ must be expressed in units of the unitary synaptic 
response, i.e. the response at a single synapse; paired recordings cannot be used in 
conjunction with $N$ to estimate $Z$ unless the typical connection multiplicity is also known. 
Moreover, such estimates assume a very specific model of spike generation—i.e. integration 
followed by a voltage threshold—in the postsynaptic neuron.

Synthetic synaptic current injection provides an alternative experimental approach to 
estimating the gain of the postsynaptic spike generator. In such experiments, synthetic 
synaptic currents are generated online and injected through a somatic patch electrode [20]. 
Chemical synapses are pharmacologically silenced, so that all the synaptic drive is provided 
through the electrode. This approach has the advantage that the transfer function between the 
input synaptic current and the output spike train can be assessed directly, for any model of 
the postsynaptic input.

A critical parameter in this approach is the size of the unitary synaptic response. One way 
to estimate this unitary response makes use of spontaneous miniature EPSCs (mEPSCs). 
Miniature EPSCs arise from the spontaneous fusion of a vesicle of neurotransmitter with the 
presynaptic terminal. The mean amplitude of spontaneous mEPSCs in layer 2/3 cortical 
neurons recorded in the sodium channel blocker tetrodotoxin (TTX) has been reported to be 
6.4 pA ([20]; see also [8]). Using synthetic synaptic current injection at the soma, Stevens and 
Zador [20] found that 115 EPSCs/second could drive cortical neurons in slice at 21 Hz, 
assuming an EPSC size of 30 pA. For the present analysis we assume the actual measured
EPSC size of 6.4 pA, yielding $Z=540$ EPSCs/action potential, the value we use in the low-gain model (Table 1). Note that recording spontaneous mEPSCs in TTX tends to undercount the number of very small mEPSCs (which fall below the detection threshold and into the noise), suggesting that the actual value of $Z$ may be higher. Moreover, the lower input resistance of neurons recorded in vivo might be expected to raise the estimate further. The actual value of $Z$ may thus be even higher.

The more than ten-fold difference in the estimate of $Z$ between the low- and high-gain models (10-40 vs. 540 EPSC/action potential) arises from three key differences in how they were measured. First, the low-gain model estimates the size of unitary EPSCs from the distribution of spontaneous mEPSCs, rather than from recordings between synaptically coupled neuronal pairs. Second, the low-gain model uses experimentally measured firing rates, rather than a simple voltage threshold, to determine the sensitivity of the spike generating mechanism. We believe that both of these choices give rise to a better estimate of $Z$ in the low-gain model.

The third difference is that the low-gain model assumes that synaptic input arrives with a constant Poisson rate, while the high-gain model assumes that inputs arrive synchronously. Because the neuronal spike generating mechanism is more sensitive to synchronized than uniformly spaced inputs, the high-gain model tends again to underestimate $Z$. However, in light of in vivo whole cell recordings suggesting that input in the intact animal does indeed tend to occur in synchronous packets [5], in this respect the high-gain model seems to offer the better estimate. We return to the role of synchronous input below.

We can now compute the input-output slope predicted by the two models, neglecting the role of synaptic depression. In the high-gain model, we have $m = N f / s = (10.000)(0.1)(1/10=10)$; in the low-gain model, we have $m = N f / s = (10.000)(0.1)(0.5)/540=1$. Thus, the slope predicted by the low-gain model is self-consistent, while that predicted by the high-gain model is two orders of magnitude too high.

3. Release probability and synaptic depression

Previous studies in the neocortex [1, 22] and hippocampus [6] have demonstrated the importance of synaptic depression at high firing rates. This depression is due to a decrease in synaptic release probability. At high firing rates—above about 20 Hz [1]—the postsynaptic response saturates to a maximum, so that the amount of neurotransmitter released per unit time is constant. This constant rate probably reflects the maximum capacity of the vesicular refilling rate of the synapse [7, 24].

To incorporate this into the model, we need to modify Eq. 1 by introducing a new expression for the release probability, $P_{\text{depress}}$, that decreases as the input firing rate increases above about 20 Hz. A simple way to do this is to impose a threshold,

$$p = \begin{cases} 
20/F_1 & \text{for } F_1 < 20 \text{ Hz} \\
(20/F_1) \frac{p}{20} & \text{for } F_1 \geq 20 \text{ Hz}
\end{cases} \quad (4).$$

Below an input firing rate of 20 Hz, $F_1 < 20 \text{ Hz}$, this expression is identical to eq. 1; but above 20 Hz, $F_1 > 20 \text{ Hz}$, $P_{\text{depress}}$, and therefore $m_{\text{depress}}$, become independent of the firing rate. That is, for high input rates $F_1 > 20 \text{ Hz}$, $F_0$ is given by,

$$F_0 = (m_{\text{depress}}) F_1$$
$$= \frac{20}{(m/F_1)} F_1$$
$$= 20 m \quad \text{5.}$$

According to this expression, at input rates above 20 Hz, synaptic depression causes the output firing rate to saturate at $F_0=20m=20 \text{ Hz}$ for the low-gain model.
Synaptic depression thus replaces the original dynamic range problem—How can low firing rates be sustained?—with a new one: How can high firing rates be maintained? Indeed, the problem is even more severe than suggested by eq. 4: In reality, synapses make the transition to the fully-depressed state gradually with increasing input firing rate, rather than abruptly as in the simple threshold model considered here. Thus even at relatively low firing rates—a few Hertz or more—the input-output slope may well fall below unity.

One resolution to this low-gain problem may rest in the value of \( f \), the fraction of neurons activated by a stimulus. We have assumed for simplicity that \( f \) is independent of firing rate, and that it is the same for all neurons. This is unlikely to be true. Instead, it is likely that \( f \) itself varies with the firing rate. That is, a larger fraction of inputs are likely activated by those stimuli that elicit the highest firing rates, thereby increasing the input-output slope \( m \). This may in part counteract the decrease in \( m \) caused by synaptic depression.

4. Synchronous firing and the low-gain model

While an increase in the fraction \( f \) of active neurons with increasing firing rate may contribute to keeping \( m \) near unity, synchronous firing probably plays a greater role. As noted above, neurons are preferentially sensitive to synaptic drive that is synchronous; we invoked this observation to explain part of the discrepancy between the low-gain and high-gain estimates of the input-output slope \( m \). Neuronal synchrony may thus serve to help overcome the reduction in gain expected when synaptic depression lowers \( m \) well below unity.

Synchrony may also explain how cortical dynamics can withstand the further reduction in cortical gain expected from inhibition. In the cortex, inhibition typically follows synchronous excitation in a characteristic sequence (Douglas and Martin 1991). By arriving immediately after the excitatory volley, the inhibitory volley reduces the gain by less than it would if it were superimposed upon the volley.

5. Summary and conclusions

Self-consistency is a powerful constraint on the formulation of theories of cortical function. Self-consistency requires that, if a cortical neuron is selected at random, its properties should reflect that of its neighbors. Two apparent violations have been considered. First, it has been suggested that neurons should, by the central limit theorem, fire regularity even when their inputs fire irregularly [17]. Second, it has been suggested that the cortical gain function is one or two orders of magnitude too high [16].

One popular resolution of these paradoxes relies on synaptic inhibition. Inhibition, balanced with uncorrelated excitation, has been proposed to induce large postsynaptic fluctuations in membrane potential and thereby lead to irregular firing [14, 15]. At the same time, inhibition could reduce the cortical gain function to near unity [16].

A direct experimental test of the balanced excitation-inhibition model concluded that it could not account for irregular firing [20]. Instead, it was proposed that synaptic drive was organized into large synchronous barrages, representing the correlated activity of dozens or hundreds of neurons. Only with such synchronous activity was it possible to reproduce in acute cortical slices the irregular firing observed in vivo.

In the present chapter, we have argued that the experimental data do not support the notion that there is a dynamic range problem at all. Using a very simple “spherical cow” model of cortical dynamics, we have shown that the cortical gain function is probably near unity, and therefore satisfies self-consistency. We have further suggested that, to the extent that the cortical gain does not lead to a self-consistent model, it is likely to be too small.
because of synaptic depression. However, refinements to the model, such as synchronous excitatory activity, can readily be included to reconcile the cortical gain function.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Symbol</th>
<th>High gain spike model default estimate (range)</th>
<th>Low gain spike model default estimate (range)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of excitatory inputs</td>
<td>$N$</td>
<td>10,000 (3000-10,000)</td>
<td>10,000 (3000-10,000)</td>
</tr>
<tr>
<td>Synaptic release probability</td>
<td>$p$</td>
<td>1</td>
<td>0.5* (0.01-1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>*at low firing rates. $p$ decreases at high firing rates due to synaptic depression</td>
</tr>
<tr>
<td>Fraction of inputs activated during a stimulus</td>
<td>$f$</td>
<td>0.1 (0.01-1)</td>
<td>0.1 (0.01-1)</td>
</tr>
<tr>
<td>Average number of EPSCs per action potential</td>
<td>$Z$</td>
<td>10 (10-40)</td>
<td>540 (100-1000)</td>
</tr>
<tr>
<td>Input-output slope</td>
<td>$m=Nfp/Z$</td>
<td>100 (3-1000)</td>
<td>1 (0.0006-20)</td>
</tr>
<tr>
<td>Input firing rate</td>
<td>$F_t$</td>
<td>1-100 Hz</td>
<td>1-100 Hz</td>
</tr>
</tbody>
</table>

Table 1. Comparison of the parameters for the low and high gain model. The models differ primarily in parameter $Z$ (the average number of EPSCs per action potential), which leads to an approximately 50-fold difference in $m$, the input-output slope. For reasons outlined in the text, we favor the value of $Z$ used in the low-gain model. Since self-consistency requires $m=1$, the low-gain model does not raise major self-consistency concerns while the high-gain model does.
inements to the model, such as synchronous 
concise the cortical gain function.

<table>
<thead>
<tr>
<th>High gain spike model</th>
<th>Low gain spike model</th>
</tr>
</thead>
<tbody>
<tr>
<td>default estimate</td>
<td>default estimate</td>
</tr>
<tr>
<td>(range)</td>
<td>(range)</td>
</tr>
<tr>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td>(3000-10,000)</td>
<td>(3000-10,000)</td>
</tr>
<tr>
<td>1</td>
<td>0.5*</td>
</tr>
<tr>
<td>(0.01-1)</td>
<td>(0.01-1)</td>
</tr>
<tr>
<td>*at low firing rates.</td>
<td></td>
</tr>
<tr>
<td>p decreases at high</td>
<td></td>
</tr>
<tr>
<td>firing rates due to</td>
<td></td>
</tr>
<tr>
<td>synaptic depression</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>(0.01-1)</td>
<td>(0.01-1)</td>
</tr>
<tr>
<td>10</td>
<td>540</td>
</tr>
<tr>
<td>(10-40)</td>
<td>(100-1000)</td>
</tr>
<tr>
<td>100</td>
<td>1</td>
</tr>
<tr>
<td>(3-1000)</td>
<td>(0.0006-20)</td>
</tr>
<tr>
<td>1-100 Hz</td>
<td>1-100 Hz</td>
</tr>
</tbody>
</table>

<low and high gain model. The models differ of EPSPs per action potential>, which leads to input-output slope. For reasons outlined in the v-gain model. Since self-consistency requires self-consistency concerns while the high-gain

References