Neural Networks 77 (2016) 95–106

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Function approximation in inhibitory networks

Bryan Tripp [a,b,*], Chris Eliasmith [a,b,c]

[a] Department of Systems Design Engineering, University of Waterloo, Canada
[b] Centre for Theoretical Neuroscience, University of Waterloo, Canada
[c] Department of Philosophy, University of Waterloo, Canada

ARTICLE INFO

Article history:
Received 26 September 2015
Revised and accepted 27 January 2016
Available online 18 February 2016

Keywords:
Dale's principle
Inhibition
Function approximation
Recurrent dynamics
Basal ganglia

ABSTRACT

In performance-optimized artificial neural networks, such as convolutional networks, each neuron makes excitatory connections with some of its targets and inhibitory connections with others. In contrast, physiological neurons are typically either excitatory or inhibitory, not both. This is a puzzle, because it seems to constrain computation, and because there are several counter-examples that suggest that it may not be a physiological necessity. Parisien et al. (2008) showed that any mixture of excitatory and inhibitory *functional* connections could be realized by a purely excitatory projection in parallel with a two-synapse projection through an inhibitory population. They showed that this works well with ratios of excitatory and inhibitory neurons that are realistic for the neocortex, suggesting that perhaps the cortex efficiently works around this apparent computational constraint. Extending this work, we show here that mixed excitatory and inhibitory functional connections can also be realized in networks that are dominated by inhibition, such as those of the basal ganglia. Further, we show that the function-approximation capacity of such connections is comparable to that of idealized mixed-weight connections. We also study whether such connections are viable in recurrent networks, and find that such recurrent networks can flexibly exhibit a wide range of dynamics. These results offer a new perspective on computation in the basal ganglia, and also perhaps on inhibitory networks within the cortex.

1. Introduction

In performance-optimized artificial neural networks, such as convolutional networks, individual neurons generally make excitatory connections with some of their targets and inhibitory connections with others. It would make little sense to constrain each neuron *a priori* to be either excitatory or inhibitory, as this would restrict the functions that could be approximated in each layer by a factor of 2^n.

But, strangely, the brain does so. There is a clear division between excitatory and inhibitory neurons. This is related to "Dale's Principle" (Eccles, 1976), which states that neurons typically release the same transmitters at all branches of their axons. There are exceptions, including excitatory/inhibitory cotransmission in the retina (Yoshida et al., 2001) and possibly in the mammalian uterus (Burnstock, 2004); the capacity for GABA to depolarize a cell, depending on resting membrane potential and internal Cl^- concentration (Chavas & Marty, 2003; Wagner, Castel, Gainer, & Yarom, 1997); and receptor-dependent mixed effects of glutamate (Katayama, Akaike, & Nabekura, 2003). Some synapses can also switch rapidly between excitatory and inhibitory transmission (Yang, Slonimsky, & Birren, 2002). Nonetheless, most neurons are exclusively excitatory or inhibitory most of the time, apparently constraining the computational power of physiological neural networks.

However, it has been shown (Parisien, Anderson, & Eliasmith, 2008) that any idealized projection model (in which each neuron may excite some targets and inhibit others) can be transformed into a more physiologically-realistic projection that is functionally nearly equivalent, solving what they call the "negative weights problem". This transformed projection is consistent with typical cortical anatomy (see e.g. Somogyi, Gamaás, Lujan, & Buhl, 1998), in that (1) the primary projection neurons are excitatory, (2) these neurons synapse onto both excitatory neurons and inhibitory interneurons in the target area, (3) the inhibitory interneurons in turn synapse onto local excitatory neurons, (4) there is substantial convergence and divergence in the synapses between each group of neurons, and (5) about 20% of the neurons are inhibitory.

Interestingly, many projections outside the cortex have essentially the same form. For example, this is true of descending

* Corresponding author at: Department of Systems Design Engineering, University of Waterloo, Canada.
E-mail address: bptripp@uwaterloo.ca (B. Tripp).

http://dx.doi.org/10.1016/j.neunet.2016.01.010
0893-6080/

projections onto thalamic nuclei (Jones, 1985). Projections from cortex onto the striatum (the main input nucleus of the basal ganglia) are also similar, in that excitatory projection neurons synapse onto the striatal projection neurons, and also onto fast-spiking interneurons that inhibit the projection neurons (Plenz, 2003).

The main counterexample in mammals is the networks within the basal ganglia. Most projection neurons in the basal ganglia (apart from those of the sub thalamic nucleus and substantia nigra pars compacta) are inhibitory. This suggests that computation in the basal ganglia may be very different from that in the cortex. For example, while function approximation is a useful analogy for understanding cortical processes such as visual feature extraction, coordinate transforms, and even processing of complex sentence-like concepts (Eliasmith, 2013), the preponderance of inhibition in the basal ganglia suggests that function approximation may not be a useful abstraction for understanding basal ganglia function.

Accordingly, many models of the basal ganglia consist of inhibition-like computations such as subtraction and competition. For example, the classic Albin/DeLong model describes basal ganglia function in terms of a balance between inhibitory and disinhibitory paths to the output nuclei (Albin, Young, & Penney, 1989; DeLong, 1990), and more elaborate extensions of this model (Gurney, Prescott, & Redgrave, 2001; Stewart, Bekolay, & Eliasmith, 2012) also treat computation in terms of inhibition and disinhibition of represented information. One notable counterexample is the dimension reduction model of Bar-Gad, Morris, and Bergman (2003), an abstract model that assumes mixed excitatory and inhibitory weights.

We wondered to what extent the range of computations in these models, i.e. largely subtraction and competition, is dictated by the dominance of inhibition. To explore this question, we extended the approach of Parisien et al. (2008) to networks that consist only of inhibitory neurons. The resulting models suggest that if the postsynaptic neurons are inhibitory, tonically active, and have local collaterals, then such networks can indeed perform function approximation much like networks with mixed excitatory and inhibitory synaptic weights. These conditions are met through most of the basal ganglia. The large projection from globus pallidus externus to the subthalamic nucleus is an exception (the target neurons are excitatory), but the same computations might be possible in this case due to collaterals in the globus pallidus.

Our results here cast little doubt on existing basal ganglia models, which are supported by a variety of evidence. However, ongoing refinement of basal ganglia models is likely, for example to match physiology in increasing detail (e.g. Humphries, Stewart, & Gurney, 2006; Wei, Rubin, & Wang, 2015), and these results may help to expose subtle mechanisms that have been less obvious before.

To reach this conclusion, we briefly review the Parisien et al. (2008) analysis for transforming mixed-weight feedforward circuits into biologically plausible circuits. We then extend this approach to feedforward inhibitory circuits in Section 4. We then explore the range of functions that can be approximated by a single projection from a presynaptic to a postsynaptic population in an inhibitory network (Section 5). We then turn to the more challenging problem of constructing recurrent circuits, in which we consider possible modes of instability that may be introduced by the transform (Section 6). Here we show that the inhibitory transform is stable when driving recurrent network dynamics with a wide range of time constants. We conclude by discussing consequences of the existence of this "inhibitory Parisien transform" for our understanding of the basal ganglia.

2. Methods

2.1. Neural Engineering Framework

Network models were constructed using the Neural Engineering Framework (NEF), which was detailed by Eliasmith and Anderson (2003). We briefly review the most relevant aspects of this approach below. We also note that while it is convenient to use the NEF, the Parisien transform applies to any network in which groups of neurons are driven by common inputs (Parisien et al., 2008).

In the NEF, populations of neurons are taken to be cosine-tuned to vector variables, and driven by a current

$$I_j = \alpha \mathbf{e_j}^T \mathbf{x} + I_j^{bias}, \tag{1}$$

where $\mathbf{e_j}$ is the jth neuron's encoding vector (or preferred direction vector), $\mathbf{x}$ is a vector that is represented by the population activity, and I_j^{bias} is a constant intrinsic bias current. Variations in the bias current across a population contribute to heterogeneity of responses to input. We say that $\mathbf{x}$ is represented by population activity in the sense that both contain the same information. In particular, the neurons' activity can be expressed as a function of $\mathbf{x}$, and $\mathbf{x}$ can be estimated from population activity. (For example, in a model of the primate middle temporal area, $\mathbf{x}$ might consist of visual velocities, stereoscopic disparities, etc.)

Populations are usually taken to have a dimensionless operating range of $\|x\|_2 < 1$. In this study we focus on representation of one-dimensional variables. In this case scalar e are typically drawn at random from $\{-1, 1\}$. In higher dimensions, vector $\mathbf{e}$ are usually drawn from the surface of a hypersphere.

Various functions $\mathbf{f}(\mathbf{x})$ can be approximately decoded as linear combinations of the output of each neuron. Specifically,

$$\hat{\mathbf{f}}(\mathbf{x}) = \sum_i \mathbf{d}_i a_i(\mathbf{x}), \tag{2}$$

where $\hat{\mathbf{f}}(\mathbf{x})$ is an approximation of $\mathbf{f}(\mathbf{x})$, $\mathbf{d}_i$ is the ith neuron's decoding vector, and a_i is the ith neuron's filtered spike train. The filter is a first-order low-pass filter that models post-synaptic current dynamics. In a simulation, neuron output is modelled as a sequence of impulse functions that correspond to spikes. Filtering these sequences with a model of post-synaptic current dynamics yields an online approximation of the spike rate. This means that function approximations that correspond to linear combinations of tuning curves can be read out while the population spikes, although they are somewhat corrupted by spike-related fluctuations.

A presynaptic population a can be connected to a postsynaptic population b, so that population b becomes tuned to a function that is approximated from population a's activity. That is, we can connect the populations so that $\mathbf{x}_b = \hat{\mathbf{f}}(\mathbf{x}_a)$, where $\mathbf{x}_a$ is the vector encoded by the presynaptic population and $\mathbf{x}_b$ is the vector encoded by the postsynaptic population. This implies that the current flowing into the post-synaptic neurons is

$$\begin{aligned} I_j &= \alpha \mathbf{e_j}^T \mathbf{x}_b + I_j^{bias} = \alpha \mathbf{e_j}^T \hat{\mathbf{f}}(\mathbf{x_a}) + I_j^{bias} \\ &= \alpha \mathbf{e_j}^T \sum_i \mathbf{d}_i a_i(\mathbf{x}_a) + I_j^{bias}. \end{aligned} \tag{3}$$

Note that the expression on the right is in terms of the presynaptic activities a_i. This dependence can be rewritten in terms of synaptic weights w_{ji} that connect individual presynaptic and postsynaptic neurons, as

$$I_j = \alpha \sum_i w_{ji} a_i(\mathbf{x}_a) + b, \tag{4}$$

where

$$w_{ji} = \mathbf{e}_j^T \mathbf{d}_i. \tag{5}$$

In idealized mixed-weight projections, the decoders are regularized optimal linear decoders (Salinas & Abbott, 1994), which minimize the squared approximation error,

$$\int_{\mathbf{x}} \left(\sum_i \mathbf{d}_i a_i(\mathbf{x}) - \mathbf{f}(\mathbf{x}) \right)^2 d\mathbf{x} + \lambda \sum_i \|\mathbf{d}\|_2^2, \tag{6}$$

where λ is a regularization parameter that penalizes large decoders. This optimization leads to a mixture of excitatory and inhibitory synaptic weights (according to Eq. (5)), because a given neuron may synapse onto other neurons with encoders that are either parallel to its decoder, antiparallel, or anything in between.

The transformed projections in both this study and in Parisien et al. (2008) involve inhibitory synapses from interneurons onto postsynaptic neurons. As described below (Section 3), the associated encoders are all negative, so the synapses are made inhibitory by constraining the decoders to be positive. We used Matlab's lsqlin function for constrained optimization of these decoders.

Post-synaptic currents following a spike are modelled with single-time-constant exponential dynamics with time constant τ^{psc}. Network dynamics of the form

$$\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u} \tag{7}$$

can be approximated by connecting a population that represents $\mathbf{u}$ to a population that represents $\mathbf{x}$, and recurrently connecting the $\mathbf{x}$ population to itself, with appropriate function approximation in each connection. Specifically, the feedforward projection should approximate the function

$$B'\mathbf{u} = \tau^{psc} B\mathbf{u}, \tag{8}$$

and the recurrent projection should approximate

$$A'\mathbf{x} = \tau^{psc} A\mathbf{x} + \mathbf{x} \tag{9}$$

(Eliasmith & Anderson, 2003). Nonlinear network dynamics can also be obtained through nonlinear feedback functions (Eliasmith, 2005).

2.2. Point-Neuron model

Numerical simulations were performed with leaky-integrate-and-fire (LIF) neurons (Knight, 1972). In the subthreshold regime of this model,

$$\tau_m \dot{v} = -v + \alpha \mathbf{e}^T \mathbf{x} + I^{bias}, \tag{10}$$

where τ_m is the membrane time constant, v is the membrane potential, $\dot{v}$ is its derivative, α is a scaling factor, $\mathbf{e}$ is the neuron's preferred direction, $\mathbf{x}$ is the vector of variables to which the neuron is tuned, and I^{bias} is a constant bias. v is normalized to between 0 and 1. When v crosses a spike threshold of 1, a spike occurs, v is reset to 0, and subthreshold integration is paused for a post-spike refractory time τ_{ref}.

2.3. Population parameters

To ensure that our main results were not dependent on a specific set of neuron model parameters, we repeated key simulations with multiple random populations, with parameters drawn from six different distributions. These parameter distributions are described in Tables 1 and 2. The parameters included τ_m (the membrane time constant), τ_{ref} (spike refractory time), α (input gain), and constant bias current I^{bias} (see Eq. (10)).

Each neuron's bias I^{bias} was derived from another parameter, which we call the "intercept". This is the value of $\|\mathbf{x}\|_2$ at which a neuron transitions from zero to non-zero spike rate, when $\mathbf{x}$ is parallel to the neuron's preferred direction. In one group of distributions (Table 1), the intercepts were gaussian-distributed with mean zero and standard deviation σ. In this case, the scale factor α was chosen first, from a gamma distribution with shape k and scale θ, and I^{bias} was then set to produce the chosen intercept. In the second group of distributions (Table 2), intercepts were uniformly distributed between T_{min} and T_{max}, and spike rates at $\mathbf{e}^T\mathbf{x} = 1$ were uniformly distributed between R_{min} and R_{max}.

The fourth distribution (see Table 2) is the default in the Nengo NEF simulator (Bekolay et al., 2014; Stewart, Tripp, & Eliasmith, 2009). This simulator was used to develop Spaun (Eliasmith et al., 2012) and many other models.

Table 1
Parameters of population distributions with gaussian-distributed intercepts. The parameters are absolute spike refractory time (τ_{ref}); membrane time constant (τ_m); standard deviation of gaussian intercept distribution (σ); and shape (k) and scale (θ) of the gamma distribution of the scale factor α.

	τ_{ref}	τ_m	σ	k	θ
1	.005	.04	2/3	2	2
2	.003	.03	1.5	3	.2
3	.004	.01	1	3	.2

Table 2
Parameters of population distributions with uniformly-distributed intercepts. The parameters are absolute spike refractory time (τ_{ref}); membrane time constant (τ_m); minimum and maximum intercepts (T_{min} and T_{max}, respectively); and minimum and maximum spike rates at $\mathbf{x} = \mathbf{e}$ (R_{min} and R_{max}, respectively). Note that this is the value of $\mathbf{x}$ within the population's nominal operating range (i.e. $\|\mathbf{x}\|_2 < 1$) at which the neuron's rate is highest.

	τ_{ref}	τ_m	T_{min}	T_{max}	R_{min}	R_{max}
4	.002	.02	−1	1	200	400
5	.005	.02	−1	1	30	80
6	.002	.1	−.95	.95	50	100

3. Feedforward excitatory projections

In this section we review the methods for transformation of a mixed-weight projection model to a model with excitatory projection neurons, which is more realistic for the cortex. The material in this section was introduced by Parisien et al. (2008), elaborating on suggestions by Eliasmith and Anderson (2003).

Beginning with an idealized projection, in which each presynaptic neuron can make both excitatory and inhibitory synapses, Parisien et al. (2008) define a transform to a realistic model in which each neuron is either excitatory or inhibitory. The transform consists of two steps. The first is to offset all of the original (mixed-sign) synaptic weights so that they become excitatory. This eliminates the mixed weights, but introduces extraneous excitatory current into the post-synaptic neurons. The second step is to cancel out this extraneous excitatory current by introducing inhibitory neurons. This could be done by introducing one inhibitory interneuron for each (now-excitatory) projection neuron (e.g. Churchland, 1996). However, this would require as many inhibitory as excitatory neurons. In the method of Parisien et al., the inhibitory neurons instead encode all of the necessary bias as a population. The size of the inhibitory population is set to 1/4 the size of the projection population, reflecting the ratio of excitatory to inhibitory neurons in the neocortex. This transform is illustrated in Fig. 1.

We now describe the transform in more detail. We begin with a connection from a presynaptic population a to a postsynaptic population b, in which a function is approximated so that $\mathbf{x}_b = \mathbf{f}^o(\mathbf{x}_a)$. Here the "o" superscript stands for "original" (pre-transform). The "original" synaptic weight between the ith presynaptic neuron and the jth post-synaptic neuron is w^o_{ji}. Each of these weights can be either positive or negative.

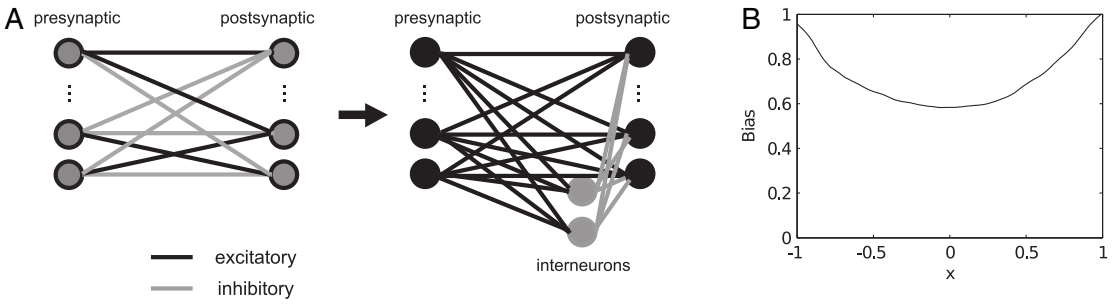

Fig. 1. The excitatory Parisien transform. A, An idealized projection in which each pre-synaptic neuron can act as a source of both excitation and inhibition is mapped to a physiologically realistic projection that performs the same computation. In the transformed projection, excitatory pre-synaptic neurons synapse both directly onto post-synaptic targets, and also indirectly through a small population of inhibitory interneurons. B, Shifting the synaptic weights in the main projection (so that they are all excitatory) introduces an excitatory bias current into the post-synaptic neurons. This bias current is a function of the variable **x** that is represented by the pre-synaptic population. This same bias function is projected to the interneurons, which in turn offset the excitatory current by inducing an approximately-equal inhibitory current into the post-synaptic neurons.
Source: Reproduced with permission from Tripp (2008).

The first step is to add a positive bias w^b_{ji} to each weight, so that transformed weights,

$$w_{ji} = w^o_{ji} + w^b_{ji},$$

are all excitatory. The only trick in defining the bias weight is that it must allow compensation by a correlated ensemble of interneurons. This correlation is what decouples the number of interneurons from the number of projection neurons.

Analogous to the NEF approach to defining synaptic weights (i.e. as products of encoders and decoders), Parisien et al. define the bias weight in terms of bias encoders and decoders, as

$$w^b_{ji} = e^b_j d^b_i,$$

where e^b_j is the bias encoder of the jth post-synaptic neuron, and d^b_i is the bias decoder of the ith pre-synaptic neuron.

These bias decoders d^b_i can be viewed as decoding a "bias function" $f^b(\mathbf{x}) = \sum_i d^b_i a_i(\mathbf{x})$ from the pre-synaptic neurons. The outcome is not highly sensitive to either the shape of the bias function or the values of the bias decoders, except that large differences between the magnitudes of different bias decoders are problematic. So these decoders are chosen to be uniform, i.e. $d^b_i = d^b$. With uniform d^b the form of this bias function is determined by the pre-synaptic neurons' tuning curves. For example, for cosine-tuned LIF neurons with a uniform distribution of intercepts, the bias function resembles a parabola that is lowest in the centre and highest at the extremes of the represented range (Fig. 1B). Typically, the uniform d^b are scaled so that this bias function has a maximum of one.

Given this definition of the bias decoder d^b, the bias encoder e^b_j of the jth post-synaptic neuron is chosen to be as small as possible, such that $w_{ij} \geq 0$ for all i. This achieved when

$$e^b_j = max_i \left(\frac{-w^o_{ji}}{d^b} \right).$$

To reiterate, the process to this point makes all the weights positive, i.e.

$$w_{ji} = w^o_{ji} + w^b_{ji} = w^o_{ji} + e^b_j d^b_i \geq 0. \tag{11}$$

However, this change in the synaptic weights also changes whatever function $f^o(\mathbf{x}_a)$ had been approximated by the original synaptic weights w^o_{ji}. Specifically, it adds the bias function $f^b(\mathbf{x}_a)$, so that the connection now approximates the function $f(\mathbf{x}_a) = f^o(\mathbf{x}_a) + f^b(\mathbf{x}_a)$. The next step is to introduce a population of inhibitory interneurons to cancel out the bias function, and thus recover the transform $f^o(\mathbf{x}_a)$ associated with the original mixed-sign weights w^o_{ji}.

To achieve this, we first add a connection that projects the bias function $f^b(\mathbf{x}_a)$ from the presynaptic neurons to the interneurons. The interneurons have uniform encoders $\mathbf{e}_k = 1$. Their decoders optimally approximate $-f^b(\mathbf{x}_a)$, within the constraint that these decoders must all be negative. The interneurons then project the decoded output $-\hat{f^b}(\mathbf{x}_a)$ to the post-synaptic neurons, which scale it with the bias encoders e^b_j. Each post-synaptic neuron therefore receives the following map of the pre-synaptic represented variable:

$$f(\mathbf{x}_a) = f^o(\mathbf{x}_a) + e^b_j f^b(\mathbf{x}_a) - e^b_j \hat{f^b}(\mathbf{x}_a) \approx f^o(\mathbf{x}_a).$$

In other words, the shift in the synaptic weights of the main projection adds excitatory bias current, and the interneurons add approximately equal inhibitory current, so that the elaborated projection model has effectively the same synaptic weights as the original idealized projection.

As discussed in the introduction, the general structure of the resulting projection (*e.g.* excitatory neurons projecting onto both excitatory neurons and a smaller number of locally-connected inhibitory neurons, *etc.*) is very common.

4. Feedforward inhibitory projections

This section shows that the method described above extends to inhibitory projection neurons. This case is less common generally, but it dominates the basal ganglia—projection neurons of the striatum, globus pallidus, and substantia nigra pars reticulata are all inhibitory. Purkinje cells, the projection neurons of the cerebellar cortex, are also inhibitory.

To transform an idealized mixed-sign projection into an inhibitory one, a *negative* bias is added to each of the original synaptic weights. In this case the bias decoders are uniform and negative, and we say that they decode the negative of the bias function $f^b(\mathbf{x}_a)$. The equation for the bias encoders actually remains the same, despite the fact that the largest-amplitude positive weight must be corrected in this case (rather than the largest-amplitude negative weight as before), because the bias decoders are negative. So again,

$$e^b_j = max_i \left(\frac{-w^o_{ji}}{d^b} \right).$$

The bias weight w^b_{ji} in this case introduces excessive inhibitory currents into the post-synaptic neurons. As before, the bias function is also projected to tonically-active inhibitory interneurons, which fire more slowly given the bias, and inhibit the post-synaptic neurons less. This input balances the increase in direct inhibition

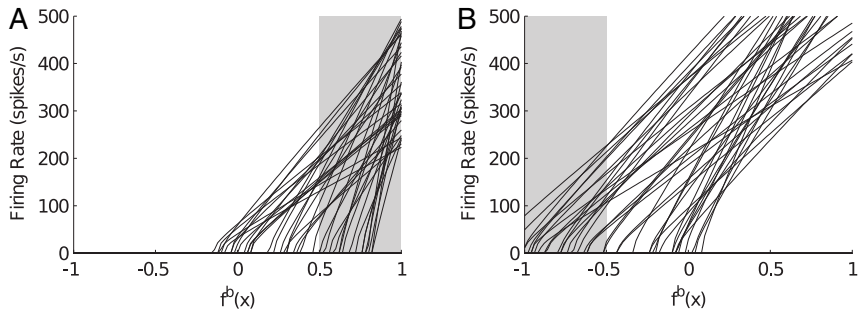

Fig. 2. Interneuron tuning in the excitatory and inhibitory transforms. The shaded area indicates the normal operating range, and the lines show tuning curves of example neurons from interneuron ensembles. A, In the excitatory transform, the excitatory pre-synaptic neurons *increase* interneuron firing from *low* intrinsic rates (*i.e.* at $f^b = 0$). B, In the inhibitory transform, the inhibitory pre-synaptic neurons *reduce* firing activity from *high* intrinsic rates.
Source: Reproduced with permission from Tripp (2008).

from the pre-synaptic neurons. The tonic activity of the interneurons is critical, because reduction in this activity is needed to disinhibit the post-synaptic neurons (see Fig. 2).

In addition, the post-synaptic neurons must also be tonically active. Specifically, tonic input from the inhibitory interneurons must be offset either by intrinsic currents or separate excitation, in order to allow the inhibition by both the pre-synaptic neurons and interneurons to result in changes in neural activity. Our models assume intrinsic currents, which we model in terms of a common factor b that is shared across the postsynaptic population, such that the intrinsic current of the jth postsynaptic neuron is $e^b_j b$. The output of the interneuron population approximates $-(x_{int} + b)$. (Thus, for example, if the bias $x_{int} = -f^b(\mathbf{x}_a)$ were zero, the interneuron output would cancel the intrinsic pacemaking currents of the postsynaptic neurons.)

Fig. 3 compares an idealized mixed-sign circuit with its transformation into both excitatory and inhibitory circuits as per the two transforms described above. In this example, the projection calculates a nonlinear and non-monotonic function of the pre-synaptically represented variable x_a, illustrating that this type of computational flexibility is retained in both the excitatory and inhibitory cases. The plotted value is an estimate of x_b (the value represented by the postsynaptic population) from the postsynaptic population's spikes. This estimate is based on unconstrained optimal linear decoders of postsynaptic population activity, which were found independently from the synaptic weights between the presynaptic and postsynaptic population. (Note that these postsynaptic decoding weights have mixed signs, and do not correspond to realistic synaptic weights; they just allow us to show that the desired signal is present within the postsynaptic spikes.)

Parisien et al. demonstrated that the transform works equally well for vectors and other nonlinear transformations, which we have confirmed holds for the inhibitory transform (results not shown).

Building on this example of approximation of a nonlinear, non-monotonic function (Fig. 3), the next section describes the range of such functions that can be approximated by the inhibitory transform.

5. Supported computations

The goal of this section is to characterize the range of functions that can be computed by Parisien projections, relative to idealized mixed-weight projections.

In an idealized mixed-weight projection, with no constraints on synaptic weights, the space of computable functions can be understood in terms of the principal components of the presynaptic tuning curves, where the tuning curves are spike rates as functions of $\mathbf{x}_a$ (equivalently, the time-averaged spike trains $T^{-1}\int_T a_i(\mathbf{x}_a)dt$ as $T \to \infty$). On a long enough time scale, a linear combination of neuron outputs can closely approximate any function in the span of the tuning curves. However, on behaviourally relevant timescales, noise due to spike timing dominates the approximation of some of these functions (Eliasmith & Anderson, 2003).

The functions that can be approximated with a high signal-to-noise ratio correspond to linear combinations of the first few principal components of the set of presynaptic tuning curves. This is because the principal components are ordered by the amount of variance they explain. As the explained variance decreases, sensitivity to noise increases because noise is isomorphic along all components. Thus the signal (explained variance) to noise ratio goes down, until the noise dominates the component. The dimension of the computable space therefore corresponds roughly to the number of singular values that are larger than the noise power.

The interneurons in a Parisien projection are an additional source of noise. The Parisien transform might therefore be expected to further restrict the space of computable functions. This can be seen qualitatively in Fig. 3, in the larger errors in panels D and E (transformed projections) than C (idealized projection).

To understand the effects of the excitatory and inhibitory transforms on function approximation more generally, we explore how the additional noise they introduce varies with principal component rank, e.g. whether they add a small constant amount of noise that becomes irrelevant for smaller principal components, or whether the noise increases so rapidly with smaller principal components that it severely restricts function approximation. Some growth in interneuron-related error might be expected, because larger synaptic weights are needed to approximate smaller principal components, and this would require larger bias signals. Thus the bias encoders must be larger, amplifying noise from the interneurons.

Fig. 4 shows how RMS error increases when the synaptic weights approximate principal components (i.e. of presynaptic tuning curves) of increasing rank. Results for idealized, excitatory, and inhibitory projections are shown. These results are averages over five randomly generated populations for each of six sets of population parameters (see Section 2.3). Errors are lowest for idealized projections, higher for excitatory projections, and highest for inhibitory projections. Specifically, the RMS errors of the excitatory and inhibitory transforms, respectively, were 9.2% and 14.7% higher (on average over the different principal components) than that of the idealized transform. These differences are more subtle than differences across principal component rank. For example, in the idealized projections, error in decoding the fourth principal component was 141% higher than error in decoding the second principal component. Simulations with larger populations (not shown) showed similar trends to those of Fig. 4, but with lower noise throughout.

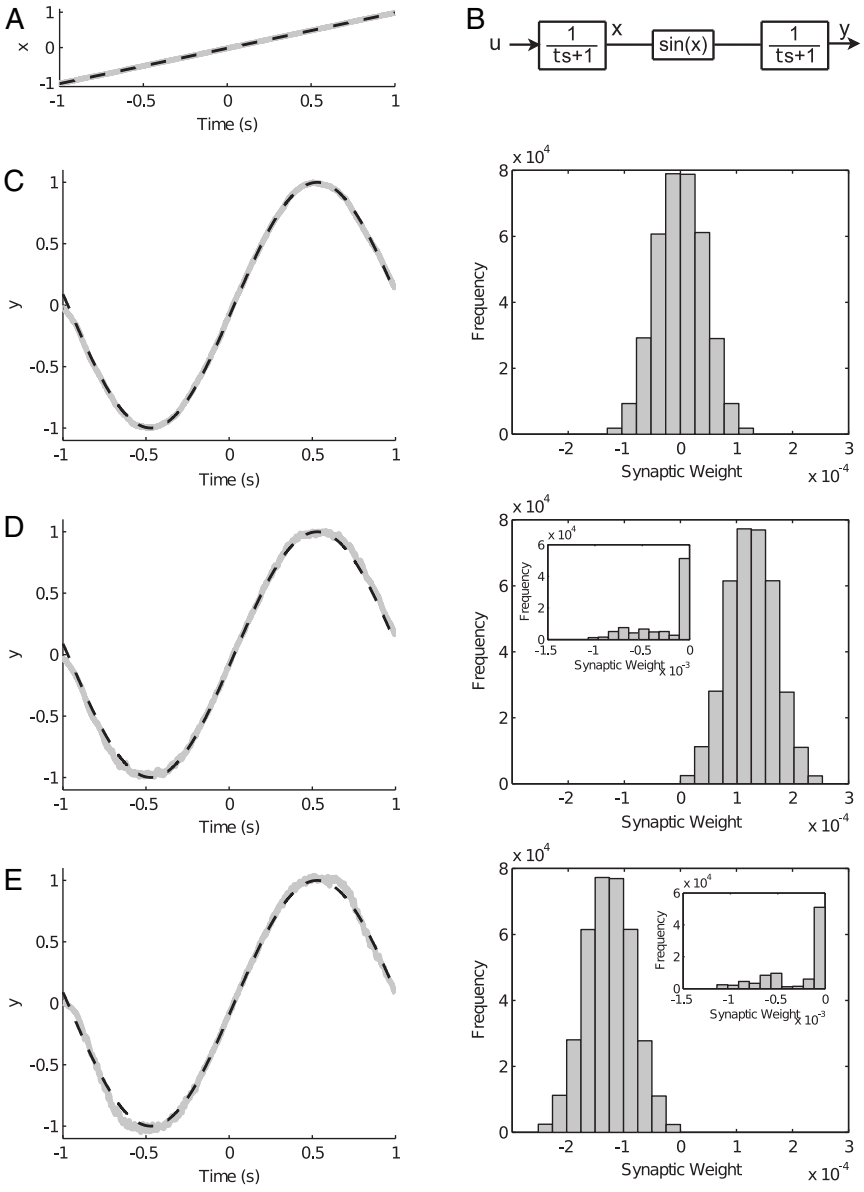

Fig. 3. Example simulations illustrating that both the excitatory and inhibitory transforms can calculate non-monotonic functions. In each of the left panels, the black dashed line indicates the ideal value of the represented variable in the postsynaptic population, and the grey line indicates its estimate, decoded from spiking activity in this population. Each of these simulations was performed with ensembles of 600 pre-synaptic neurons, 600 post-synaptic neurons, and (in the transformed projections) 150 interneurons. A, The pre-synaptic ensemble represents an input variable that increases linearly with time. B, Diagram of the network structure, consisting of a single projection from a one-dimensional ensemble to another, in which the synaptic weights approximate the map $y = sin(x)$. C, Optimal linear decoding of y from spiking neural activity in the post-synaptic ensemble, with an idealized mixed-weight projection. The right panel shows a histogram of the synaptic weights in this projection. D, Optimal linear decoding of y from activity in the post-synaptic ensemble, after the excitatory Parisien transform. The right panel shows the shifted distribution of synaptic weights in the main projection (all above zero). The inset shows the distribution of synaptic weights in the projection from the inhibitory neurons to the post-synaptic neurons. E, As (D), but with the inhibitory Parisien transform.
Source: Reproduced with permission from Tripp (2008).

Error also arises from lag in the path through the interneurons (not shown). These errors are small when $\mathbf{x}_a$ changes slowly (as in Fig. 3), and also when a large principal component is approximated (because the bias encoders are small). However, they become prominent when $\mathbf{x}_a$ changes rapidly and smaller principal components are approximated (e.g. as Fig. 3 but with a 10x faster ramp).

In summary, both the excitatory and inhibitory projections can perform somewhat restricted computations relative to an idealized mixed-weight projection, but the restrictions are subtle. Qualitatively, Parisien projections exhibit gradually increasing errors when approximating smaller principal components, much like idealized mixed-weight projections. These results reinforce the conclusions of Parisien et al., and also suggest that networks of inhibitory neurons may have a capacity for function approximation that is comparable to that of excitatory/inhibitory networks.

6. Recurrent projections

In an excitatory Parisien projection (i.e. a combined direct/indirect structure that can be obtained by applying the transform), interneuron currents are slightly lagged in time behind the direct bias currents, because of the extra synapse in the pathway through the interneurons. This lag introduces an error, the magnitude of which varies with the first time derivative of the bias function, df^b/dt. In a feedforward network this error tends to be small, in part because excitatory synapses onto inhibitory neurons

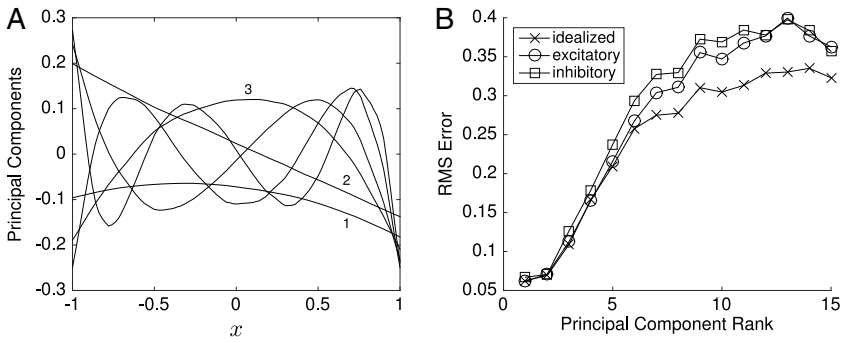

Fig. 4. A, The first few principal components of diverse responses of an example LIF neuron population. The first, second, and third principal components resemble (relatively) constant, linear, and parabolic functions, respectively, and later principal components have higher frequencies. B, Root-mean-squared error in the approximation of the principal components, vs. principal component rank, for idealized, excitatory-Parisien, and inhibitory-Parisien types of projections. For each function, the error is calculated over a one-second simulation in which a presynaptic population is driven with a ramping input which goes from −1 to 1, and this population drives a postsynaptic population with a decoded function of its input that corresponds to one of the principal components of the tuning curves. The error is the difference between the network output (decoded from the spikes of the postsynaptic neurons) and a filtered version of the input (filtered to match the feedforward synaptic dynamics of the network). The results are an average over five networks with parameters drawn from each of the six parameter distributions of Section 2.3 (i.e. a total of 30 networks).

tend to have fast dynamics (Carter & Regehr, 2002; Geiger, Lübke, Roth, Frotscher, & Jonas, 1997; Walker, Lawrence, & McBain, 2002). However, as Parisien et al. (2008) pointed out, the associated delay raises the possibility of instability in a recurrent network. They investigated this possibility using an integrator network as an example, and did not discover a stability problem. The integrator example is a reasonable choice, because by definition it operates on the border of instability. However, it remains possible that instability might arise in other types of recurrent networks.

In contrast with the excitatory transform, the interneurons of the inhibitory transform are of the same type as the post-synaptic neurons, and (we assume) have the same synaptic dynamics. Thus lag in the disynaptic branch is double that in the monosynaptic branch. On the other hand, inhibition arrives first in this case, suggesting that a relatively larger delay (in disinhibition) may not threaten stability in the same way.

This section reconsiders the stability issue in light of these differences. We examine a wider variety of networks than Parisien et al., and show that a Parisien network can be unstable even if the corresponding idealized network is stable. We also show that the stability limits of the excitatory transform are narrowed when the post-synaptic current time constant of the interneurons is as large as that of the others, but that this is not the case in the inhibitory transform.

A brief terminological note: in a recurrent network, the pre-synaptic and post-synaptic ensembles are the same, and are referred to below as the primary ensemble (as opposed to the interneuron ensemble).

6.1. New mode of instability

Instability that arises from the transform can be illustrated with a recurrent network that approximates the linear dynamical system,

$$\dot{\mathbf{x}} = A\mathbf{x}. \tag{12}$$

This is a special case of the dynamical system, $\dot{\mathbf{x}} = A\mathbf{x} + B\mathbf{u}$ (Eq. (7)). As discussed in the Methods, such idealized dynamics can be approximated by a spiking network in which a function $A'\mathbf{x} = \tau A\mathbf{x} + \mathbf{x}$ is decoded from an ensemble of neurons and fed back to the same neurons (Eliasmith & Anderson, 2003), where A' is the neural feedback matrix of Eq. (9) (a modification of A that accounts for the synaptic time constant). However, because there is some error in the spike-decoded approximation $\hat{\mathbf{x}}$ of $\mathbf{x}$ (i.e. $\hat{\mathbf{x}} \neq \mathbf{x}$), the dynamics of the neural network are more precisely,

$$\tau\dot{\mathbf{x}} = A'\hat{\mathbf{x}}(\mathbf{x}) - \mathbf{x}. \tag{13}$$

In the following we will assume that these dynamics are stable, and concentrate on the stability of new dynamics that the Parisien transform introduces.

In the transformed circuit, bias in the direct feedback projection is ideally cancelled out by feedback through the interneurons. However, the bias and interneuron feedback may be imbalanced due to imperfect decoding from the interneuron ensemble (i.e. interneuron output $\hat{x}_{int}(x_{int}) \approx f^b(\mathbf{x}_a)$), and also due to the additional lag in the path through the interneurons when the represented value is changing. This (usually small) difference,

$$\Delta^{di} = \pm f^b(\mathbf{x}_a) - \hat{x}_{int}, \tag{14}$$

between direct and indirect bias (where the sign is $+ve$ for the excitatory transform and $-ve$ for the inhibitory transform) is the key to understanding how the network can become unstable. Both the direct and indirect bias affect the primary neurons through synapses. So the effect of this difference on the primary neurons at any given instant in time can be modelled as $D(t) = h(t) * \Delta^{di}(t)$, where $h(t)$ is the impulse response function of the post-synaptic current dynamics, and $*$ denotes convolution. In other words, the spiking of the primary neurons depends in part on the difference between the direct and indirect bias, filtered by the post-synaptic current dynamics.

Recall that the bias encoders e^b_j all have the same sign, so the firing rates of all neurons in the primary ensemble rise and fall together with changing D. As described in Section 3, the bias function $f^b(\mathbf{x})$ is a sum of the activities of neurons in the primary ensemble (which are themselves a function of $\mathbf{x}$). Since these neurons are also affected by D, the bias is more accurately described as a function of both $\mathbf{x}$ and D. Consequently, D can be viewed as a state variable that forms part of an additional feedback loop through the network, as illustrated in Fig. A.6. Accounting for this new state variable D, the excitatory system has the following dynamics:

$$\tau\dot{\mathbf{x}} = A'\hat{\mathbf{x}}(\mathbf{x}, D) - \mathbf{x}, \tag{15}$$

$$\tau\dot{D} = f^b(\mathbf{x}, D) - \hat{x}_{int}(x_{int}) - D, \tag{16}$$

$$\tau_{int}\dot{x}_{int} = f^b(\mathbf{x}, D) - x_{int}, \tag{17}$$

where x_{int} is the variable represented by the interneuron ensemble, $\hat{x}_{int}(x_{int})$ is the decoded estimate of x_{int} from interneuron activity, and similarly $\hat{\mathbf{x}}(\mathbf{x}, D)$ is the decoded estimate of $\mathbf{x}$ from primary ensemble activity. Similarly, the inhibitory system has the dynamics,

$$\tau\dot{\mathbf{x}} = A'\hat{\mathbf{x}}(\mathbf{x}, D) - \mathbf{x}, \tag{18}$$

$$\tau \dot{D} = -f^b(\mathbf{x}, D) - \widehat{(x_{int} + b)}(x_{int}) - D + b, \quad (19)$$

$$\tau_{int}\dot{x}_{int} = -f^b(\mathbf{x}, D) - x_{int}, \quad (20)$$

where b is a bias that models intrinsic pacemaking currents, and the output of the interneurons approximates the function $-(x_{int} + b)$ (see Section 4).

Linearization of Eqs. (15)–(17) (see Appendix) suggests that the excitatory transform can become unstable when the idealized dynamics have large negative eigenvalues, and that the stability limits narrow if τ_{int} approaches τ. In contrast, linearization of the inhibitory transform (Eqs. (18)–(20)) suggests that it does not become unstable with large negative eigenvalues.

Fig. 5 shows simulation results that confirm the analytical results. In these simulations, recurrent networks were parameterized to act as low-pass filters that acted on the represented variable **x** with various time constants. We refer to these time constants as "network time constants" to distinguish them from synaptic time constants. In these networks **x** was one-dimensional, and the network time constant was simply the inverse of A (similarly, if **x** is multi-dimensional, the network time constants are the inverse eigenvalues of A).

The time constant of post-synaptic currents in the primary neurons was 10 ms. The time constant of the recurrent network dynamics depended on both the synaptic time constant and on feedback. In particular, the network time constant was greater than 10 ms when the feedback was positive (thus slowing the decay of x), and less than 10 ms when it was negative (hastening the decay of x).

The plots show fractions of networks (drawn from various parameter distributions) that were stable with various network time constants. All networks with time constants around 10 ms were stable, because 10 ms corresponds to zero feedback. The idealized (non-Parisien) networks (symbol x) were stable through the full range of time constants.

Consistent with Parisien et al. (2008), the excitatory Parisien networks were stable in networks with long time constants (which are similar to integrators). However, they were unstable with strong negative feedback associated with very short network time constants (consistent with our linearization analysis). We also simulated excitatory networks in which the time constants of synapses onto interneurons (which are normally fast) were the same as the other synaptic time constants. Consistent with our analysis, these networks had narrower stability margins. They became unstable with somewhat weaker negative feedback than the standard excitatory Parisien networks, and they were also unstable with strong positive feedback.

In contrast, the inhibitory Parisien networks were stable throughout the tested range of network time constants, with both positive and negative feedback. This is also consistent with the linearization results (see Appendix).

7. Discussion

The main implication of this study is that inhibitory networks can, in principle, support diverse computations including approximation of a variety of non-monotonic functions. In fact, our models suggest that the versatility of inhibitory networks is similar to that of excitatory–inhibitory networks, which in turn is comparable to that of the idealized mixed-sign projections of artificial neural networks.

We also found that inhibitory networks could exhibit a wider range of dynamics than excitatory–inhibitory networks, without becoming unstable. This included very fast dynamics that arose form strong negative feedback. This suggests that one potential benefit of an all-inhibitory network is versatility as a dynamical system.

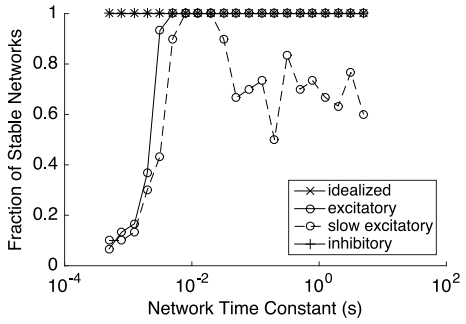

Fig. 5. Simulations of recurrent networks with the inhibitory transform. Simulations were performed with five networks with random properties drawn from each of the six parameter distributions (Section 2.3), for a total of 30 networks. Each network was repeatedly simulated with different feedback, in order to vary the network time constant from 0.0005 s to 5 s. The plots show the fraction of networks that were stable, as a function of network time constant. The populations were driven by external step input which was 0 for 0.2 s, stepping to 1 for a further 0.3 s. A network was considered stable if $\hat{x}$ was similar to the output of a perfect low-pass filter with the same time constant as the network (specifically, the mean output over 20 ms periods immediately before the step and immediately before the end of simulation had to match the filter within 0.15) and $\hat{x}_{int}$ was less than 1.5. Both the mixed-weight networks and the inhibitory Parisien networks were stable throughout the range of network time constants. The excitatory Parisien networks were unstable with strong negative feedback. Modified excitatory Parisien networks with uniform synaptic time constants (rather than the usually faster synapses onto interneurons) were unstable with relatively weaker negative feedback, and also with positive feedback.

7.1. Interaction between function approximation and dynamics

The interneurons introduce lag in one of the signal paths from presynaptic to postsynaptic neurons. Related to the present stability results, we previously found that interneuron lag in the excitatory transform had little effect on network dynamics, in a variety of networks designed to act as differentiators, with inputs over a wide range of frequencies (Tripp & Eliasmith, 2010, particularly figure 6).

However, as mentioned in Section 5, the effect of this lag is more pronounced when the approximated function is nonlinear. This is because nonlinear functions require larger decoders, which require larger bias encoders, and these amplify the lagged signal more strongly. Thus in general the error introduced by the transform depends on both the decoded function (larger for more strongly nonlinear functions, as shown in Fig. 4) and the frequency content of **x**.

7.2. Relevance to the basal ganglia

The inhibitory transform makes two key assumptions about circuit properties. The first (shared with the original transform of Parisien et al.) is that target neurons are locally connected through an inhibitory network. This assumption is satisfied by both cortical and basal ganglia networks. In addition, for the inhibitory transform, it is essential that many target and interneurons be tonically active. This is necessary so that inhibition modulates the activity of these populations. Significantly, the basal ganglia are rife with tonically active neurons (Surmeier, Mercer, & Chan, 2005), including the principal neurons of the globus pallidus, subthalamic nucleus, and substantia nigra pars reticulata, striatal cholinergic interneurons, and dopaminergic neurons. The key properties necessary for performing a rich set of computations are therefore available in the basal ganglia.

The inhibitory Parisien transform results in a network in which all neurons are inhibitory. This suggests that a recurrently-connected inhibitory population (e.g. within the globus pallidus) could potentially project a wide variety of functions recurrently, and consequently exhibit a wide variety of dynamics.

Notably, the Parisien transform is not affected by the synaptic action of the post-synaptic neurons, i.e. whether they are excitatory, inhibitory, or modulatory. As a result, either the excitatory or inhibitory transform is relevant to each of the excitatory and inhibitory projections within the basal ganglia, and in fact throughout the cortico-basal loops. For example, the excitatory transform is consistent with the cortico-striatal projection, despite the fact that the target neurons (the medium spiny neurons) are inhibitory. Similarly, the inhibitory transform is consistent with the pallido-thalamic projection – in which most pallidal neurons terminate both directly onto thalamic projections neurons and indirectly through local circuit interneurons (Ilinsky, Yi, & Kultas-Ilinsky, 1997) – despite the fact that the projection neurons of the thalamus are excitatory.

However, importantly, we have not modelled specific basal ganglia circuits here in detail. For example, we have not explored how rebound conductances in the subthalamic nucleus interact with synaptic dynamics, we have not considered implications of perisomatic varicosities in the projection from the external to internal segment of the globus pallidus, etc.

7.3. Future work

An important future direction is optimization of the performance of both the excitatory and inhibitory transforms. We have taken some preliminary steps in this direction. For example, we found that we could improve performance by using non-uniform bias decoders that were optimized to produce a flatter bias function, with the constraint that the bias encoders were not allowed to grow. Another potential approach would be to use the backpropagation algorithm to optimize all three parts of the projection together. Finally, because regularization affects weight magnitudes in the idealized projection, greater regularization would reduce the role of the interneurons, perhaps reducing noise overall. It may be possible to parameterize regularization of the idealized weights in a way that further optimizes the full Parisien structure.

Another useful step forward would be to optimize feedback networks to avoid the new instability that we have described. A relevant degree of freedom is the choice of state–space realization, given desired input–output dynamics. In preliminary work in this direction (Tripp, 2008), we used a change of basis of the state variables to stabilize a network without changing the nominal input–output dynamics. However, the stable realization was relatively noisy. It may be possible to define a "canonical neural realization" that considers both stability and noise propagation, analogous to the various canonical realizations of linear systems theory. This would require a suitable model of noise propagation through spiking networks, such as in our previous work (Tripp & Eliasmith, 2010).

It may also be possible to decode alternative nonlinear functions from the interneurons of excitatory Parisien networks, which improve the stability of these networks.

Importantly, because the networks in the present study are not highly optimized, the results presented here can be seen as a lower bound on achievable performance. With further optimization, the ranges of stable dynamics and computable functions in Parisien models may more closely approximate those of idealized models.

Finally, we have not yet shown that either of the Parisien projections can be learned. Differences in cortical and basal ganglia learning could well be a source of differences in their function approximation and dynamic capacities.

7.4. Experimental validation

The transform suggested here would ideally be subject to experimental validation. That is, if we observe anatomy of essentially the right form, ideally we should be able to tell whether a Parisien projection would result in the connectivity observed in a given circuit. However, this type of validation presents a difficult problem, because the transform is robust to a variety of changes that result in different predictions about connectivity and firing patterns.

In a specific Parisien-transformed model, the firing patterns of the interneurons and post-synaptic neurons are different. This suggests that one could develop a Parisien-transformed model of a specific system, and then check experimentally whether a minority of neurons exhibits firing patterns that resemble the model interneurons' firing patterns. The first problem is that it is not clear how many interneurons to expect. Parisien et al. (2008) assume a 1:4 ratio of interneurons to excitatory cells (to match the proportion of inhibitory neurons in the cortex), but the proportion is less critical for performance than the absolute number. Furthermore, several projections might share the same interneuron ensemble. With more projections sharing the same interneurons, the performance would degrade gradually, because finer differences in the value of the bias function would become significant. Eventually the physical limit of convergence onto the interneurons would be reached, but this limit could be surpassed if the bias function were coded by only a subset of the correlated pre-synaptic neurons. In summary, the proportion of interneurons required for the Parisien transform is not well defined.

A second issue is that the distinct firing pattern of the interneurons is not well defined. Parisien et al. (2008) assume for convenience that the bias decoders are uniform, but, as they discuss, this assumption is not critical. Different bias decoders would result in a different bias function, and consequently a different pattern of interneuron activity. Careful selection of bias decoders might allow matching of a variety of experimental observations. Failure to do so would cast doubt on the transform, but success would provide only minor support.

Finally, specifically for the case of the inhibitory transform, there is no reason that the interneurons and post-synaptic neurons have to be distinct groups. Instead, all the network states including $\mathbf{x}$ and x_{int} could be encoded as a single vector by a single group of neurons with multidimensional tuning. As a result, a single recurrently-connected, multi-dimensional ensemble could operate in the same manner as two separate inhibitory groups. This further confounds expectations about classes of firing patterns in the network.

In sum, empirical support for the transform is likely very difficult to come by. This, of course, is not an unfamiliar position for theoretical analyses to be in. We believe that the generality of the transform, its applicability to both excitatory and inhibitory circuits, and the insights it offers regarding biologically constrained implementations of dynamical systems make it a very useful tool for comparing models to experimentally observable networks. At the very least, it provides a systematic way to determine if a hypothesized function for a neural system can be plausibly ascribed to that system given known anatomical and physiological constraints. In the specific case of the basal ganglia, the existence of this transform significantly broadens the set of possible functions compared to what is typically assumed.

8. Conclusion

We have shown a way in which networks of inhibitory neurons can perform function approximation and exhibit flexible recurrent dynamics. Their performance is comparable to idealized network

models in which neurons have mixed excitatory and inhibitory effects. The range of stable dynamics in these inhibitory models is broad, in fact somewhat broader than in excitatory–inhibitory networks modelled on the cortex. These results suggest that the predominance of inhibition in the basal ganglia does not, independent of other factors, prevent very flexible function approximation and dynamics.

Acknowledgements

This work was supported by CFI and OIT (223144) infrastructure funding and grants from Canada Research Chairs, NSERC CGS-D and Discovery grants 261453 and 296878, ONR grant N000141310419, AFOSR grant FA8655-13-1-3084.

Appendix. Feedback stability

This appendix analyses feedback dynamics of the excitatory and inhibitory transforms by linearizing Eqs. (15)–(20) and considering the eigenvalues of the resulting linear systems. An eigenvalue with a positive real component reflects unstable linear dynamics, i.e. unstable growth in the magnitude of the state vector.

We simplify the dynamic model by the approximation $\hat{\mathbf{x}} = \mathbf{x}$. This approximation is reasonable because there are many primary neurons, with effectively unconstrained synaptic weights, so the linear decoding of $\mathbf{x}$ is relatively accurate. Furthermore, moderate changes in D have little effect on $\mathbf{x}$, because for any change in D, neurons with opposite preferred directions change their firing rates either up or down together. This approximation allows us to focus on the stability of the new feedback loop.

A.1. Linearization of excitatory system

Assume the system in $\mathbf{x}$ is linear and stable.

$$\begin{aligned}\tau\dot{D} &= f^b(\mathbf{x}, D) - (\hat{x}_{int})(x_{int}) - D \\ &= h(\mathbf{x}, D, x_{int}),\end{aligned}$$

$$\begin{aligned}\tau_{int}\dot{x}_{int} &= f^b(\mathbf{x}, D) - x_{int} \\ &= g(\mathbf{x}, D, x_{int}).\end{aligned}$$

Linearizing around some operating points, $\mathbf{x}_0, D_0, x_{int0}$, gives

$$\begin{aligned}\tau\dot{D} &= h(\mathbf{x}_0, D_0, x_{int0}) + \frac{\partial h}{\partial x_{int}}(x_{int} - x_{int0}) \\ &\quad + \frac{\partial h}{\partial D}(D - D_0) + \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) \\ &= h(\mathbf{x}_0, D_0, x_{int0}) - \frac{\partial \hat{x}_{int}}{\partial x_{int}}(x_{int} - x_{int0}) \\ &\quad + \left(\frac{\partial f^b}{\partial D} - 1\right)(D - D_0) + \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0)\end{aligned}$$

$$\begin{aligned}\tau_{int}\dot{x}_{int} &= g(\mathbf{x}_0, D_0, x_{int0}) + \frac{\partial g}{\partial x_{int}}(x_{int} - x_{int0}) \\ &\quad + \frac{\partial g}{\partial D}(D - D_0) + \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) \\ &= g(\mathbf{x}_0, D_0, x_{int0}) - (x_{int} - x_{int0}) + \frac{\partial f^b}{\partial D}(D - D_0) \\ &\quad + \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0).\end{aligned}$$

Because we do not assume that we are at $g(\mathbf{x}_0, D_0, x_{int0}) = h(\mathbf{x}_0, D_0, x_{int0}) = 0$, we can write these as equations as being the difference between these dynamics, and those at the operating point. This results in the equations describing the dynamics of deviations around the operating point, that is,

$$\begin{aligned}\tau(\dot{D} - \dot{D}_0) &= -\frac{\partial \hat{x}_{int}}{\partial x_{int}}(x_{int} - x_{int0}) + \left(\frac{\partial f^b}{\partial D} - 1\right)(D - D_0) \\ &\quad + \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0)\end{aligned}$$

$$\tau\dot{\delta}^D = -\frac{\partial \hat{x}_{int}}{\partial x_{int}}\delta^{x_{int}} + \left(\frac{\partial f^b}{\partial D} - 1\right)\delta^D + \frac{\partial f^b}{\partial \mathbf{x}}\delta^{\mathbf{x}}$$

$$\tau_{int}(\dot{x}_{int} - \dot{x}_{int0}) = -(x_{int} - x_{int0}) + \frac{\partial f^b}{\partial D}(D - D_0) + \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0)$$

$$\tau_{int}\dot{\delta}^{x_{int}} = -\delta^{x_{int}} + \frac{\partial f^b}{\partial D}\delta^D + \frac{\partial f^b}{\partial \mathbf{x}}\delta^{\mathbf{x}}.$$

To simplify notation, we let $\alpha = \partial f^b/\partial D$ and let $\beta = \partial \hat{x}_{int}/\partial x_{int}$. Assuming the input from the nominally stable system is not part of the analysis, we treat the autonomous system in $\delta^{x_{int}}$ and δ^D where the linearized dynamics matrix can be written:

$$A^L = \begin{bmatrix}(\alpha - 1)/\tau & -\beta/\tau \\ \alpha/\tau_{int} & -1/\tau_{int}\end{bmatrix}.$$

The eigenvalues λ of the system are:

$$2\lambda = \frac{(\alpha - 1)}{\tau} - \frac{1}{\tau_{int}} \pm \sqrt{\left(\frac{1 - \alpha}{\tau} + \frac{1}{\tau_{int}}\right)^2 - \frac{4(\alpha(\beta - 1) + 1)}{\tau\tau_{int}}}.$$

An unstable eigenvalue ($\lambda > 0$) will exist if $\tau/\tau_{int} < \alpha - 1$. As anticipated, this corresponds to interneuron post-synaptic current (PSC) dynamics τ_{int} that are too slow, relative to PSC dynamics τ in the primary ensemble.

Another threat to stability arises if the second term under the square root is negative, i.e. if $\beta < 1 - 1/\alpha$. This suggests that a sufficiently negative slope in the decoding error of the interneuron ensemble (Fig. A.6B) would cause a self-perpetuating divergence between the direct and indirect feedback. In such a scenario, ultimately all the neurons in the network would saturate at their maximum firing rates.

In both cases, the magnitude of α is a critical parameter. It varies with x and D, but unfortunately its range is hard to define. This is because it is a function of the bias encoders, which depend on the synaptic weights, which in turn depend in complex ways on the tuning curves of the primary ensemble. However, α generally increases (endangering stability through both of the above mechanisms) with increases in the absolute values of the entries in A'. Thus, counter-intuitively, idealized network dynamics with *large negative eigenvalues* are at risk of becoming unstable in excitatory Parisien form.

A.2. Linearization of inhibitory system

Assume the system in $\mathbf{x}$ is linear and stable.

$$\begin{aligned}\tau\dot{D} &= -f^b(\mathbf{x}, D) - (\widehat{x_{int} + b})(x_{int}) - D + b \\ &= h(\mathbf{x}, D, x_{int}),\end{aligned}$$

$$\begin{aligned}\tau_{int}\dot{x}_{int} &= -f^b(\mathbf{x}, D) - x_{int} \\ &= g(\mathbf{x}, D, x_{int}).\end{aligned}$$

Linearizing around some operating points, $\mathbf{x}_0, D_0, x_{int0}$ gives

$$\begin{aligned}\tau\dot{D} &= h(\mathbf{x}_0, D_0, x_{int0}) + \frac{\partial h}{\partial x_{int}}(x_{int} - x_{int0}) \\ &\quad + \frac{\partial h}{\partial D}(D - D_0) + \frac{\partial h}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) \\ &= h(\mathbf{x}_0, D_0, x_{int0}) - \frac{\partial(\widehat{x_{int} + b})}{\partial x_{int}}(x_{int} - x_{int0}) \\ &\quad - \left(\frac{\partial f^b}{\partial D} + 1\right)(D - D_0) - \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0)\end{aligned}$$

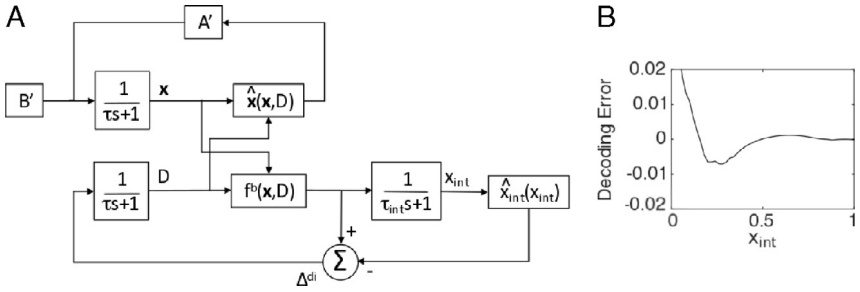

Fig. A.6. Sources of instability. A, Block diagram of a model of feedback dynamics including factors introduced by the Parisien transform (the excitatory transform is shown). In particular, neuron activity in the primary ensemble gives rise to the bias function $f^b(\mathbf{x}, D)$, which feeds back both directly through the main projection, and also indirectly through the interneurons. The indirect route introduces a lag, and an additional decoding error. Note that the two dynamic blocks $1/(\tau s+1)$ (where s is the Laplace variable) both correspond to synapses onto the primary neurons. These blocks are separated according to the logical distinction between the $\mathbf{x}$ and D state variables. Physically, the direct feedback, which includes both $\hat{\mathbf{x}}$ and $f^b(\mathbf{x}, D)$, corresponds to both of the feedback paths in the diagram that do not pass through x_{int}. B, Decoding error $\hat{x}_{int} - x_{int}$ in an example interneuron ensemble (150 neurons) from an excitatory transform. The constraint on the sign of the decoders makes the decoding error relatively large, particularly near zero.
Source: Reproduced with permission from Tripp (2008).

$$\begin{aligned}\tau_{int}\dot{x}_{int} &= g(\mathbf{x}_0, D_0, x_{int0}) + \frac{\partial g}{\partial x_{int}}(x_{int} - x_{int0}) \\ &\quad + \frac{\partial g}{\partial D}(D - D_0) + \frac{\partial g}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) \\ &= g(\mathbf{x}_0, D_0, x_{int0}) - (x_{int} - x_{int0}) - \frac{\partial f^b}{\partial D}(D - D_0) \\ &\quad - \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0).\end{aligned}$$

Because we do not assume that we are at $g(\mathbf{x}_0, D_0, x_{int0}) = h(\mathbf{x}_0, D_0, x_{int0}) = 0$, we can write these as equations as being the difference between these dynamics, and those expressed at the shifted location. This results in the equations describing the dynamics of the system around the new point, that is,

$$\begin{aligned}\tau(\dot{D} - \dot{D}_0) &= -\frac{\partial(\widehat{x_{int} + b})}{\partial x_{int}}(x_{int} - x_{int0}) - \left(\frac{\partial f^b}{\partial D} + 1\right)(D - D_0) \\ &\quad - \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) \\ \tau\dot{\delta}^D &= -\frac{\partial(\widehat{x_{int} + b})}{\partial x_{int}}\delta^{x_{int}} - \left(\frac{\partial f^b}{\partial D} + 1\right)\delta^D - \frac{\partial f^b}{\partial \mathbf{x}}\delta^{\mathbf{x}} \\ \tau_{int}(\dot{x}_{int} - \dot{x}_{int0}) &= -(x_{int} - x_{int0}) - \frac{\partial f^b}{\partial D}(D - D_0) - \frac{\partial f^b}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) \\ \tau_{int}\dot{\delta}^{x_{int}} &= -\delta^{x_{int}} - \frac{\partial f^b}{\partial D}\delta^D - \frac{\partial f^b}{\partial \mathbf{x}}\delta^{\mathbf{x}}.\end{aligned}$$

To simplify notation, we let $\alpha = \partial f^b/\partial D$ and let $\beta = \partial(\widehat{x_{int} + b})/\partial x_{int}$. Assuming the input from the nominally stable system is not part of the analysis, we treat the autonomous system in $\delta^{x_{int}}$ and δ^D where the linearized dynamics matrix can be written:

$$A^L = \begin{bmatrix} -(\alpha+1)/\tau & -\beta/\tau \\ -\alpha/\tau_{int} & -1/\tau_{int} \end{bmatrix}.$$

The eigenvalues λ of the system are:

$$\begin{aligned}2\lambda &= \frac{-(\alpha+1)}{\tau} - \frac{1}{\tau_{int}} \\ &\quad \pm\sqrt{\left(\frac{\alpha+1}{\tau} + \frac{1}{\tau_{int}}\right)^2 - \frac{4(\alpha(1-\beta)+1)}{\tau\,\tau_{int}}}.\end{aligned}$$

Note that $\alpha = \partial f^b/\partial D > 0$, because D corresponds to over-excitation of the neurons. Therefore, the real part of the eigenvalues is always negative, indicating stability, in agreement with the spiking simulation results.

References

Albin, R. L., Young, a. B., & Penney, J. B. (1989). The functional anatomy of basal ganglia disorders. *Trends in Neurosciences*, *12*(10), 366–375.

Bar-Gad, I., Morris, G., & Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, *71*(6), 439–473.

Bekolay, T., Bergstra, J., Hunsberger, E., Dewolf, T., Stewart, T. C., Rasmussen, D., et al. (2014). Nengo: a Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, *7*(January), 48.

Burnstock, G. (2004). Cotransmission. *Current Opinion in Pharmacology*, *4*(1), 47–52.

Carter, A. G., & Regehr, W. G. (2002). Quantal events shape cerebellar interneuron firing. *Nature Neuroscience*, *5*(12), 1309–1318.

Chavas, J., & Marty, A. (2003). Coexistence of excitatory and inhibitory GABA synapses in the cerebellar interneuron network. *The Journal of Neuroscience*, *23*(6), 2019–2031.

Churchland, P. (1996). *The engine of reason, the seat of the soul: A philosophical journey into the brain*. MIT Press.

DeLong, M. R. (1990). Primate models of movement disorders of basal ganglia origin. *Trends in Neurosciences*, *13*(7), 281–285.

Eccles, J. (1976). From electrical to chemical transmission in the central nervous system. *Notes and Records of the Royal Society of London*, 219–230.

Eliasmith, C. (2005). A unified approach to building and controlling spiking attractor networks. *Neural Computation*, *17*(6), 1276–1314.

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. Oxford University Press.

Eliasmith, C., & Anderson, C. (2003). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. MIT Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–5.

Geiger, J. R. P., Lübke, J., Roth, A., Frotscher, M., & Jonas, P. (1997). Submillisecond AMPA receptor-mediated signaling at a principal neuron–interneuron synapse. *Neuron*, *18*(6), 1009–1023.

Gurney, K., Prescott, T. J., & Redgrave, P. (2001). A computational model of action selection in the basal ganglia. I. A new functional anatomy. *Biological Cybernetics*, *84*(6), 401–410.

Humphries, M. D., Stewart, R. D., & Gurney, K. N. (2006). A physiologically plausible model of action selection and oscillatory activity in the basal ganglia. *The Journal of Neuroscience*, *26*(50), 12921–12942.

Ilinsky, I. a., Yi, H., & Kultas-Ilinsky, K. (1997). Mode of termination of pallidal afferents to the thalamus: A light and electron microscopic study with anterograde tracers and immunocytochemistry in Macaca mulatta. *Journal of Comparative Neurology*, *386*(4), 601–612.

Jones, E. (1985). *The thalamus*. Springer.

Katayama, J., Akaike, N., & Nabekura, J. (2003). Characterization of pre- and postsynaptic metabotropic glutamate receptor-mediated inhibitory responses in substantia nigra dopamine neurons. *Neuroscience Research*, *45*(1), 101–115.

Knight, B. W. (1972). Dynamics of encoding in a population of neurons. *The Journal of General Physiology*, *59*(6), 734–766.

Parisien, C., Anderson, C. H., & Eliasmith, C. (2008). Solving the problem of negative synaptic weights in cortical models. *Neural Computation*, *20*(6), 1473–1494.

Plenz, D. (2003). When inhibition goes incognito: Feedback interaction between spiny projection neurons in striatal function. *Trends in Neurosciences*, *26*(8), 436–443.

Salinas, E., & Abbott, L. F. (1994). Vector reconstruction from firing rates. *Journal of Computational Neuroscience*, *1*, 89–107.

Somogyi, P., Gamaás, G., Lujan, R., & Buhl, E. H. (1998). Salient features of synaptic organization in the cerebral cortex. *Brain Research Reviews*, *26*(JUNE 1998), 113–135.

Stewart, T. C., Bekolay, T., & Eliasmith, C. (2012). Learning to select actions with spiking neurons in the basal ganglia. *Frontiers in Neuroscience*, *6*(JAN), 1–14.

Stewart, T. C., Tripp, B., & Eliasmith, C. (2009). Python scripting in the nengo simulator. *Frontiers in Neuroinformatics*, *3*(March), 7.

Surmeier, D. J., Mercer, J. N., & Chan, C. S. (2005). Autonomous pacemakers in the basal ganglia: who needs excitatory synapses anyway? *Current Opinion in Neurobiology*, *15*(3 SPEC. ISS.), 312–318.

Tripp, B. (2008). *A search for principles of basal ganglia function*. Ph.D. thesis, University of Waterloo.

Tripp, B. P., & Eliasmith, C. (2010). Population models of temporal differentiation. *Neural Computation*, *22*(3), 621–659.

Wagner, S., Castel, M., Gainer, H., & Yarom, Y. (1997). GABA in the mammalian suprachiasmatic nucleus and its role in diurnal rhythmicity. *Nature*, *387*(6633), 598–603.

Walker, H. C., Lawrence, J. J., & McBain, C. J. (2002). Activation of kinetically distinct synaptic conductances on inhibitory interneurons by electrotonically overlapping afferents. *Neuron*, *35*(1), 161–171.

Wei, W., Rubin, J. E., & Wang, X.-J. (2015). Role of the indirect pathway of the basal ganglia in perceptual decision making. *The Journal of Neuroscience*, *35*(9), 4052–4064.

Yang, B., Slonimsky, J. D., & Birren, S. J. (2002). A rapid switch in sympathetic neurotransmitter release properties mediated by the p75 receptor. *Nature Neuroscience*, *5*(6), 539–545.

Yoshida, K., Watanabe, D., Ishikane, H., Tachibana, M., Pastan, I., & Nakanishi, S. (2001). A key role of starburst amacrine cells in originating retinal directional selectivity and optokinetic eye movement. *Neuron*, *30*(3), 771–780.