Feedforward to the Past: The Relation between Neuronal Connectivity, Amplification, and Short-Term Memory

Surya Ganguli[1,*] and Peter Latham[2]
[1]Sloan-Swartz Center for Theoretical Neurobiology, University of California, San Francisco, San Francisco, CA 94143, USA
[2]Gatsby Computational Neuroscience Unit, UCL, London WC1N 3AR, UK
*Correspondence: surya@phy.ucsf.edu
DOI 10.1016/j.neuron.2009.02.006

Two studies in this issue of *Neuron* challenge widely held assumptions about the role of positive feedback in recurrent neuronal networks. Goldman shows that such feedback is not necessary for memory maintenance in a neural integrator, and Murphy and Miller show that it is not necessary for amplification of orientation patterns in V1. Both suggest that seemingly recurrent networks can be feedforward in disguise.

It's a poor sort of memory that only works backwards.
—The White Queen in Lewis Carroll's *Through the Looking Glass*

An enduring puzzle in the systems neuroscience of memory arises from the existence of a wide gap separating two distinct timescales: (1) the biophysical timescale of milliseconds, over which single neurons can remember their inputs, and (2) the cognitive timescale of seconds, over which our short-term memory operates. How can such rapidly forgetful neurons mediate short-term memory? If one only wants to remember discrete items, such as a person's name, then point attractor networks (Hopfield, 1982) can do the job. However, such networks cannot remember a continuous stream of analog information, something that is critical for a wide range of tasks—from simple ones, like holding our eyes still, to far more complicated ones, like parsing a spoken sentence.

The difficulty of remembering a stream of analog information can already be seen in the problem of neural integration. An integrator is a simple memory device that accumulates and remembers its total input over time. For example, an oculomotor integrator localized in the brainstem and cerebellum maintains a memory of eye position by integrating eye velocity signals (Robinson, 1989). A hallmark characteristic of any neural integrator is that a brief pulse of input leads to a graded pattern of persistent activity that long outlasts the pulse. For example, an efference copy of a brief eye movement command provided to the oculomotor integrator yields a persistent activity pattern that constitutes a memory trace of the new eye position, a trace that can be used to stabilize the eye. How can such graded activity be precisely maintained by forgetful neurons? One oft proposed theoretical solution makes use of the recurrent connectivity that exists among excitatory neurons (see e.g., Seung, 1996). Such connections mediate positive feedback, which prevents the decay of neuronal activity and allows neurons to persistently fire even in the absence of input. However, any solution that relies on positive feedback to maintain memory suffers from a severe fine-tuning problem: the amount of positive feedback must exactly balance the intrinsic decay tendencies of individual neurons. Too much feedback leads to runaway growth of neural activity, while too little leads to decay. In either case, memory of the total input is rapidly lost.

In this issue of *Neuron*, Goldman (2009) proposes an alternate mechanism for memory maintenance that does not rely on positive feedback, thereby elegantly circumventing the fine-tuning problem suffered by prior models of neural integration. Goldman begins by considering a purely feedforward chain of neurons (Figure 1A). An input pulse to the first neuron triggers a wave of feedforward activity that lasts up to a time proportional to the length of the chain, or about $N\tau$, where N is the number of neurons in the chain and τ is the intrinsic decay time constant of an individual neuron. Although each individual neuron responds transiently to the pulse (Figure 1B), the summed activity of the network resembles graded persistent activity whose magnitude is proportional to the pulse size. Thus, the network acts like an integrator, for up to N times the intrinsic decay time of single neurons. More generally, Goldman suggests that each node in Figure 1A could represent not a single neuron, but a group of neurons that act together as one stage of an N-stage feedforward network.

Of course one might object that purely feedforward stages of connectivity, as in Figure 1A, are at odds with the observation of strongly recurrent connectivity in many brain regions. However, Goldman and two other studies reviewed below all show that what looks like recurrent connectivity may not be: networks that appear, anatomically, to be strongly recurrent may nevertheless functionally behave like purely feedforward networks. We show a concrete example in Figures 1C–1F. The key idea is that instead of single neurons driving other single neurons in a feedforward chain as in Figure 1A, whole population *patterns* of activity can drive other population patterns of activity in a feedforward chain, without any activity pattern exerting positive feedback on itself (see Figure 1E, top.) This situation can occur in a fully recurrent network, as in Figure 1C. Such networks yield long transient dynamical patterns of activity in response to a pulse, which can, through a weighted sum, again yield stable persistent activity that lasts a time of about $N\tau$.

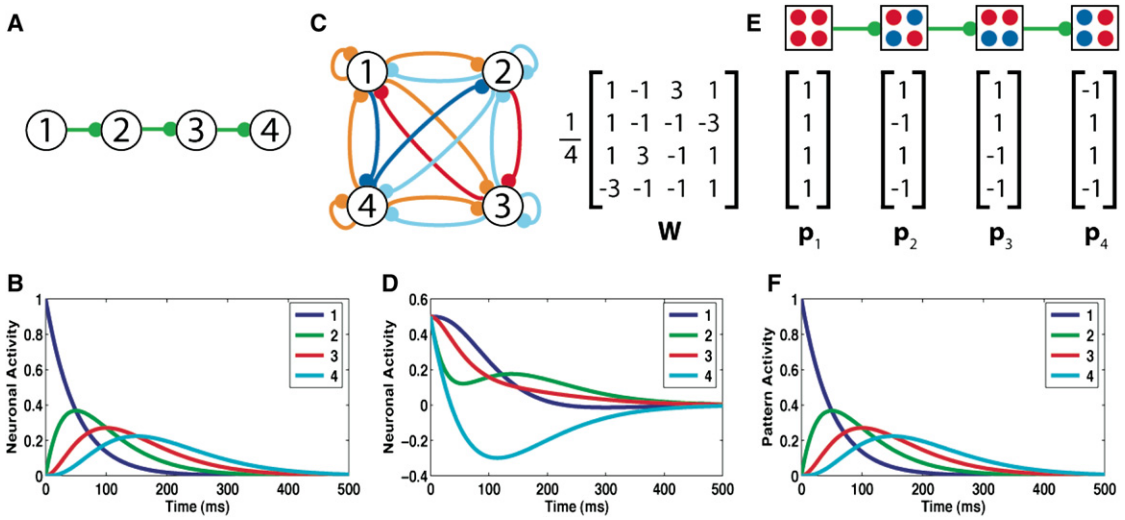

Figure 1. Functionally Feedforward Dynamics Hidden in a Recurrent Architecture

(A) A purely feedforward network of four neurons.

(B) The response of this network to a pulse of input to neuron 1. Each neuron is a leaky integrator with an intrinsic time constant of 50 ms. Although the firing rate of the first neuron decays on this timescale, it excites a transient wave of feedforward activity that lasts more than 200 ms, or four times the intrinsic neuronal decay time.

(C) A recurrent network with a 4 × 4 connectivity matrix, **W**. The network is shown on the left with different colors for the four possible weights: $^{3}/_{4}$ (red), $^{1}/_{4}$ (orange), $-^{1}/_{4}$ (light blue), $-^{3}/_{4}$ (dark blue).

(D) An equal pulse of input to all four neurons in panel (C) yields a complex, transient pattern of activity across neurons that again lasts more than four times the intrinsic neuronal timescale. (Negative firing rates can be interpreted as firing below spontaneous levels.)

(E) The long transient arises because the network is a feedforward network in disguise: for $i = 1$, 2, and 3, **W** maps $\mathbf{p}_i$ to $\mathbf{p}_{i+1}$ ($\mathbf{W}\mathbf{p}_i = \mathbf{p}_{i+1}$), and, although not shown in the figure, **W** maps $\mathbf{p}_4$ to zero ($\mathbf{W}\mathbf{p}_4 = 0$). Each vector $\mathbf{p}_i$ is a pattern of activity across neurons shown on top (red neurons are active, blue suppressed).

(F) The same neuronal response in panel (D) can be plotted in terms of the amplitude of each activity pattern $\mathbf{p}_i$. Initially, all neurons have the same level of excitation, and only pattern $\mathbf{p}_1$ is present. However, over time activity is transferred in a feedforward wave from pattern to pattern, recovering dynamics identical to that of the purely feedforward network (note in particular that the traces in panels [B] and [F] are identical).

Is there an advantage to purely feedforward networks compared to purely feedback ones? One key issue, as mentioned above, is robustness to perturbations in connectivity. Goldman shows that functionally feedforward networks can tolerate large percentage changes in connectivity that would otherwise lead to instabilities in purely feedback networks. But also, functionally feedforward networks provide rich transient dynamical responses to inputs that could serve as a basis for more general temporal processing beyond simply the maintenance of persistent activity. Indeed a wide variety of cortical and hippocampal areas reveal rich dynamical patterns of activity during working memory tasks, rather than simple, static patterns of persistent activity (see the Introduction of Goldman [2009] for references). In particular, Goldman examines recordings from monkey prefrontal cortex during a working memory task (Batuev, 1994) and shows that functionally feedforward networks can fit the diversity of neuronal responses, whereas purely feedback networks, with single modes of activity, cannot.

An important point not touched upon in Goldman (2009) is the sensitivity of integration to noise. Indeed, the consideration of noise reveals an important limit on memory. For example, in the simple case of a feedforward chain of neurons as in Figure 1A, suppose that in addition to the input pulse of signal to neuron 1, each neuron in the chain also receives continuous background input, which we take to be noise. The network will integrate not only the signal, but also the noise. Since a neuron at a certain depth in the chain receives noise from all upstream neurons, the strength of noise accumulates linearly down the chain. The strength of the signal, however, stays constant as it propagates down the chain, since it enters only at the first neuron. Thus, the signal-to-noise ratio (SNR) decays inversely with time. When the SNR reaches 1, the network has effectively lost any memory about the size of the input pulse. Thus, the network's memory is limited by τ times the input SNR.

How can we get around the problem of noise? The answer can be found in a recent study by Ganguli et al. (2008). In that study, the authors investigated the ability of general networks, in a class closely related to that considered by Goldman, to remember a sequential stream of analog input in the presence of noise. They found that networks with purely feedback interactions are not able to remember their inputs beyond a time governed by the input SNR, no matter how large they are. Thus, (hidden) feedforward structure is necessary if a network's memory of past inputs is to last up to a time proportional to its size. However feedforward structure alone is not enough; Ganguli et al. showed that amplification between feedforward stages is also required to combat noise. This amplification cannot be achieved by amplifying the signals carried by single neurons since neurons have a limited range of firing rates. However, if the number of neurons in each stage grows sufficiently rapidly, one can achieve distributed signal amplification without saturating individual neurons. With such amplification, the memory of a network can grow indefinitely with the number of feedforward stages. Since noise

accumulates linearly in the number of stages, the number of neurons per stage must grow at least as fast to preserve the input SNR. This places a limit on the duration of time over which any network of N neurons can remember its input. Indeed, Ganguli et al. (2008) prove mathematically that no network of N neurons can remember an input stream for a time longer than $\sqrt{N}\tau$, in the presence of both strong noise and nonlinear saturation, and any network that approaches this limit must employ a (possibly hidden) distributed feedforward scheme.

Functionally feedforward structures solve several theoretical puzzles, but are they used in the brain? Interestingly, evidence that they are can be found in an elegant study by Murphy and Miller (2009), also in this issue of *Neuron*. These authors were motivated by a very different puzzle than temporal memory, one involving selective amplification of cortical activity patterns in V1. As is well known, an oriented stimulus yields a sensory-evoked pattern of activity across V1 in which cells with orientation preferences similar to that of the stimulus have high activity while the rest have low activity. Kenet et al. (2003) examined spontaneous V1 activity in an anesthetized cat in the absence of a stimulus and found that this activity resembled sensory-evoked patterns more often than chance. One possible explanation for the resemblance between spontaneous and evoked activity is the selective amplification of orientation maps from unstructured inputs through positive feedback loops (Goldberg et al., 2004). Basically, neurons with similar orientation preferences excite each other, leading to the amplification of orientation map-like activity. However, networks that amplify inputs through positive feedback do so at a price: they respond slowly to their inputs. Intuitively this is because activity must propagate multiple times through the recurrent loops in order to be amplified, and such propagation takes time. But significant slowing down does not seem to occur in the data in Kenet et al. (2003); spontaneous activity in V1 fluctuates on a timescale comparable to its inputs. Of course, strong and rapid amplification could occur in one feedforward step, but where could such a step exist in the V1 recurrent circuitry? Murphy and Miller propose that such a feedforward step would naturally be hidden in a ubiquitous feature of cortical circuitry: strongly excitatory circuits balanced by equally strong inhibition. For example, any fluctuation that tilts the balance in favor of excitation would transiently drive both excitatory and inhibitory populations, but eventually increased inhibition would restore the balance. This yields a hidden feedforward single-stage amplifier in which small differential patterns of excitatory and inhibitory firing drive large common patterns of firing. Although the authors do not rule out purely feedback mechanisms of sensory map amplification in V1, they argue convincingly that transient amplification through this feedforward mechanism should play an important role alongside traditional feedback mechanisms.

Through a remarkable and simultaneous convergence of ideas, the three studies discussed in this preview have highlighted the importance of hidden feedforward connectivity in recurrent architectures, from the three different but related perspectives of neuronal integration, sequence memory, and sensory amplification. Perhaps the most fascinating test of these ideas would be the direct observation of such feedforward connectivity hidden within the oncoming rush of connectomics data. More generally, beyond the realm of neuroscience, dynamical systems governed by hidden feedforward structures are known in the physics and mathematics literature as *nonnormal* dynamical systems. Due to their rich and long-lasting transient behavior, models of nonnormal dynamics have been invoked to explain many varied and subtle aspects of our natural world, from the transition to turbulence in fluid mechanics to population growth patterns in ecology (Trefethen and Embree, 2005). The studies discussed here are among the first to connect the general theory of nonnormal dynamics to the field of neuroscience and, as such, provide intriguing hypotheses for how network connectivity may yield rich emergent dynamics capable of bridging the gap between biophysics and cognition.

REFERENCES

Batuev, A.S. (1994). Acta Neurobiol. Exp. *54*, 335–344.

Ganguli, S., Huh, D., and Sompolinsky, S. (2008). Proc. Natl. Acad. Sci. USA *105*, 18970–18975.

Goldberg, J.A., Rokni, U., and Sompolinsky, H. (2004). Neuron *42*, 489–500.

Goldman, M.S. (2009). Neuron *61*, this issue, 621–634.

Hopfield, J.J. (1982). Proc. Natl. Acad. Sci. USA *79*, 2554–2558.

Kenet, T., Bibitchkov, D., Tsodyks, M., Grinvald, A., and Arieli, A. (2003). Nature *425*, 954–956.

Murphy, B.K., and Miller, K.D. (2009). Neuron *61*, this issue, 635–648.

Robinson, D.A. (1989). Annu. Rev. Neurosci. *12*, 33–45.

Seung, H.S. (1996). Proc. Natl. Acad. Sci. USA *93*, 13339–13344.

Trefethen, L.N., and Embree, M. (2005). Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators (Princeton, NJ: Princeton University Press).