bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks

Thomas Miconi
The Neurosciences Institute
La Jolla, CA 92037, USA
miconi@nsi.edu

Abstract

Neural activity during cognitive tasks exhibits complex dynamics that flexibly encode task-relevant variables. Recurrent neural networks operating in the near-chaotic regime, which spontaneously generate rich dynamics, have been proposed as a model of cortical computation during cognitive tasks. However, existing methods for training these networks are either biologically implausible, and/or require a continuous, real-time error signal to guide the learning process. The lack of a biological learning method currently restricts the plausibility of recurrent networks as models of cortical computation. Here we show that a biologically plausible learning rule can train such recurrent networks, guided solely by delayed, phasic rewards at the end of each trial. Networks operating under this learning rule successfully learn nontrivial tasks requiring flexible (context-dependent) associations, memory maintenance, nonlinear mixed selectivities, and coordination among multiple outputs. Furthermore, applying this method to learn various tasks from the experimental literature, we show that the resulting networks replicate complex dynamics previously observed in animal cortex, such as dynamic encoding of task features, switching from stimulus-specific to response-specific representations, and selective integration of sensory input streams. We conclude that recurrent neural networks offer a plausible model of cortical dynamics during both learning and performance of flexible behavior.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Recent evidence suggests that neural representations are highly dynamic, encoding multiple aspects of tasks, stimuli, and commands in the joint fluctuating activity of interconnected populations of neurons, rather than in the stable activation of specific neurons (Meyers et al. 2008; Barak, Tsodyks, and Romo 2010; Stokes et al. 2013; Churchland et al. 2012; Raposo, Kaufman, and Churchland 2014). Models based on recurrent neural networks (RNN), operating in the near-chaotic regime, seem well-suited to capture similar dynamics (Jaeger 2001; Maass, Natschläger, and Markram 2002; Buonomano and Maass 2009; Sussillo and Abbott 2009). For this reason, such models have been used to investigate the mechanisms by which neural populations solve various computational problems, including working memory (Barak et al. 2013; Rajan, Harvey, and Tank 2016), motor control (Sussillo et al. 2015; Laje and Buonomano 2013; Hennequin, Vogels, and Gerstner 2014), and perceptual decision-making (Mante et al. 2013).

However, the methods commonly used to train these recurrent models are generally not biologically plausible. The most common training methods are based on supervised learning, in which a non-biological algorithm (usually a form of backpropagation or regression) minimizes the difference between the network's output and a target output signal (Pearlmutter 1995; Jaeger 2001; Sussillo and Abbott 2009; Song, Yang, and Wang 2016; Rajan, Harvey, and Tank 2016). Besides the non-biological nature of these algorithms, the requirement for a constant supervisory signal is in stark contrast with most behavioral tasks, in which the only source of information about performance are time-sparse rewards that are usually delayed with regard to the actions that caused them.

A more biologically plausible form of learning is reward-modulated Hebbian learning: during ongoing activity, each synapse accumulates a *potential* weight change according to classical Hebbian learning, by multiplying pre- and post-synaptic activities at any time and accumulating

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

this product over time. These potential weight changes are then multiplied by a global reward signal, which determines the *actual* weight changes. This method, inspired by the effect of dopamine on synaptic plasticity, has been successfully demonstrated and analyzed in feedforward or weakly connected spiking (Izhikevich 2007; Florian 2007; Frémaux, Sprekeler, and Gerstner 2010) and firing-rate (Soltoggio and Steil 2013) networks. However, simple reward-modulated Hebbian learning does not work for strongly-connected recurrent networks that can generate complex trajectories of the type discussed here (Fiete, Fee, and Seung 2007) (see also Appendix).

A method that successfully trains arbitrary recurrent networks is the so-called *node-perturbation* method (Fiete and Seung 2006; Fiete, Fee, and Seung 2007). This method consists in applying small perturbations to neural activity, then calculating potential weight changes by multiplying the "normal" (non-perturbative) inputs by the perturbations (rather than by post-synaptic output, as in Hebbian learning). These potential weight changes are then multiplied by a reward signal to provide the final weight changes. This method was successfully applied to feedforward networks to model birdsong learning (Fiete, Fee, and Seung 2007) and our own previous results show that it is also successful for chaotic recurrent neural networks (T. Miconi, arXiv:1507.08973 [q-bio.NC]). Interestingly, this method is largely similar to the well-known REINFORCE algorithm, which is widely used in reinforcement learning (Williams 1992; Mnih et al. 2014; Peters and Schaal 2008; Kober, Bagnell, and Peters 2013) (see in particular Eq. 11 in (Williams 1992)).

However, node perturbation is non-Hebbian (since it multiplies two types of inputs, rather than pre- and post-synaptic activities) and requires information that is not local to the synapse (namely, the perturbatory inputs, which must somehow be kept separate from the "normal" inputs). Thus, it is not obvious how node-perturbation could be implemented in biological neural networks. Legenstein and colleagues (Hoerzer, Legenstein, and Maass 2014; Legenstein et al. 2010) showed that, under certain conditions, node-perturbation could be made more biologically plausible by leveraging moment-to-moment fluctuations in post-synaptic activity: by keeping a

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

running average of recent activity and subtracting it from the current instantaneous response at any time, we obtain a "high-pass" filtered trace of post-synaptic activity, which can be used as a proxy for the exploratory perturbations of post-synaptic activity (but see below). This can then be multiplied by the pre-synaptic inputs, and the final accumulated product is then modulated by a reward signal to recreate the node-perturbation method in a more biologically plausible, Hebbian manner (i.e. as a product of pre-synaptic and post-synaptic activities rather than between two input sources) (Legenstein et al. 2010). This method can successfully train chaotic recurrent neural networks (Hoerzer, Legenstein, and Maass 2014). Unfortunately, this method critically requires an instantaneous, real-time continuous reward signal to be provided at each point in time. The continuous, real-time reward signal is necessary to allow the subtraction method to extract task-relevant information and counter the effect of spurious deviations introduced by the running-average subtraction process (see Appendix). This is in contrast with most tasks (whether in nature or in the laboratory), which only provide sparse, delayed rewards to guide the learning process.

In summary, to our knowledge, there is currently no biologically plausible learning algorithm that can successfully train chaotic recurrent neural networks with realistic reward regimes. This limitation may restrict the potential plausibility of recurrent neural networks as operational models of actual cortical networks.

Here we introduce a novel reward-modulated Hebbian learning rule that can be used to train recurrent networks for flexible behaviors, with reward occurring in a delayed, one-time fashion after each trial, as in most animal training paradigms. This method is Hebbian and uses only synapse-local information, without requiring instantaneous reward signals (see Methods). We apply our method to several tasks that require flexible (context-dependent) decisions, memory maintenance, and coordination among multiple outputs. By investigating the network's representation of task-relevant aspects over time, we find that trained networks exhibit complex dynamics previously observed in recordings of animal frontal cortices, such as dynamic encoding of task features (Meyers et al. 2008; Stokes et al. 2013; Jun et al. 2010), switching from

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

stimulus-specific to response-specific representations (Stokes et al. 2013), and selective integration of sensory input streams (Mante et al. 2013). We conclude that recurrent networks endowed with reward-modulated Hebbian learning offer a plausible model of cortical computation and learning, capable of building networks that dynamically represent and analyze stimuli and produce flexible responses in a way that is compatible with observed evidence in behaving animals.

Results

Description of the learning rule

Here we provide an overview of the plasticity rule used in this paper. A full description is provided in Methods. Furthermore, we provide an extensive discussion of the rule at a more intuitive level in the Appendix.

Our model networks are fully-connected continuous-time recurrent neural networks operating in the early chaotic regime, which allows them to autonomously generate rich dynamics while still being amenable to learning (Sompolinsky, Crisanti, and Sommers 1988; Sussillo and Abbott 2009; Jaeger 2001; Maass, Natschläger, and Markram 2002). The activity of neuron *i* is governed by the following equations:

$$\tau \frac{dx_i}{dt} = -x_i(t) + \sum_{j=1}^{N} J_{j,i} r_j(t) + \sum_{k=1}^{M} B_{k,i} u_k(t) \qquad \text{[Equation 1]}$$

$$r_i(t) = tanh(x_i(t)) \qquad \text{[Equation 2]}$$

where x_i is the total excitation (or "potential") of neuron i, r_i is its actual response (or "firing rate" / activity), $J_{j,i}$ is the connection weight from neuron j to neuron i, $u_k(t)$ is the current value of each of the M external inputs to the network, and $B_{k,i}$ is the connection weight from external

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

input k to neuron i (τ is the relaxation time constant of neuron activation). One or more neurons in the network are arbitrarily designated as the "output" neurons, and their responses at any given time are used as the network's response (these neurons are otherwise identical to all others). For the simulations reported here, networks include 200 neurons (400 for the motor control task).

Synapses between neurons are modified according to a novel form of reward-modulated Hebbian learning, which we now describe.

First, in order to produce exploratory variation in network responses across trials, each neuron in the network occasionally receives a random perturbation $\Delta_i(t)$ to its activation. During a trial, at every time step, every synapse from neuron *i* to neuron *j* accumulates a *potential* Hebbian weight change (also called eligibility trace (Izhikevich 2007)) according to the following equation:

$$e_{i,j}(t) = e_{i,j}(t-1) + S(\, r_i(t-1) * (x_j(t) - \bar{x}_j)\,) \qquad \text{[Equation 3]}$$

Remember that r_i represents the output of neuron *i*, and thus the current input at this synapse. x_j represents the activation of neuron *j* and $\underline{x}_j$ represents a short-term running average of x_j, and thus $x(t)$ - $\underline{x}$ tracks the fast fluctuations of neuron output. Thus this rule is essentially Hebbian, based on the product of inputs and output (fluctuations). Importantly, *S* is a monotonic, supralinear function; in this paper we simply used the cubic function $S(x) = x^3$, though the particular choice of function is not crucial as long as it is supralinear (see Methods and Appendix). Note that the eligibility trace for any synapse is accumulated over the course of a trial, with each new timestep adding a small increment to the synapse's eligibility trace / potential weight change.

At the end of each trial, a certain reward R is issued to the network, based on the network's performance for this trial as determined by the specific task. From this reward, the system computes a *reward prediction error* signal, as observed in physiological experiments, by subtracting the expected reward for this trial $\underline{R}$ (generally a running average of previously

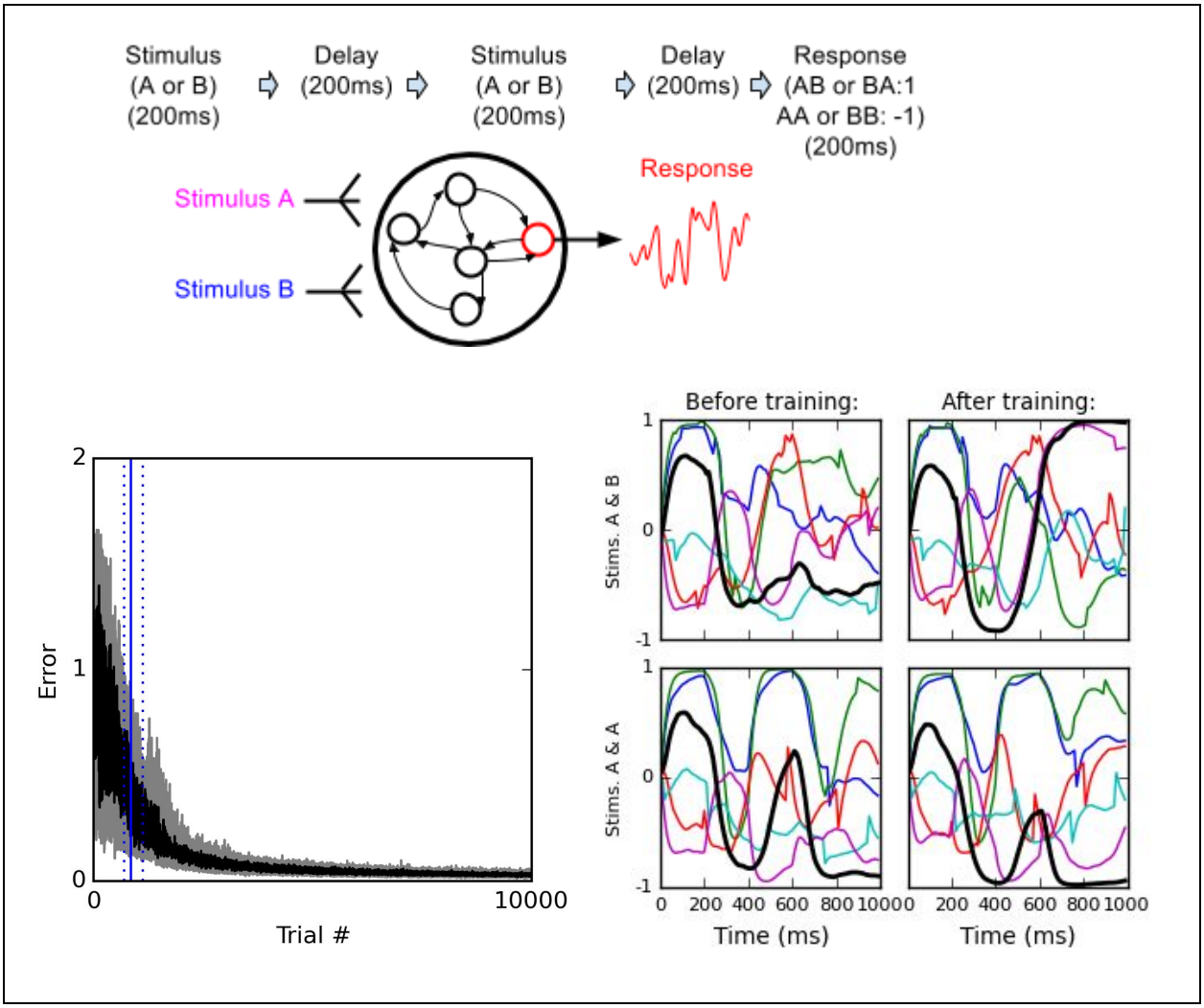
bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

received rewards; see Methods) from the actually received reward R. This reward-prediction signal is used to modulate the eligibility trace, producing the actual weight change:

$$\Delta J_{i,j} = \eta\, e_{i,j}(R - \overline{R}) \qquad \text{[Equation 4]}$$

where η is a learning rate constant. Together, equations 3 and 4 fully determine the learning rule described here. See Methods and Appendix for a complete description.

Task 1: Delayed nonmatch-to-sample task



bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1. Delayed nonmatch-to-sample task. Top: task description. The network is exposed to two successive stimuli, with an intervening delay. The task is to produce output -1 if the two stimuli were identical (AA or BB), or 1 if they were different (AB or BA); the output of the network is simply the activity of one arbitrarily chosen "output" neuron, averaged over the last 200 ms of the trial. Bottom, left: time course of response error for each successive trial, as a function of the trial number (dark curve: median over 20 runs; gray area: inter-quartile range). The solid vertical line indicate the median number of trials needed to reach the criterion of 95% correct responses over 100 successive trials (843 trials); dotted vertical lines indicate the inter-quartile range (692-1125 trials). Performance (i.e., magnitude of the response error) continues to improve after reaching criterion and reaches a low, stable residual asymptote. Bottom, right: Activities of 6 different neurons, including the output neuron (thick black line), for two stimulus combinations, before training (left) and after training (right). Note that neural traces remain highly dynamical even after training.

The first task considered here is a simple delayed nonmatch-to-sample problem (Figure 1). In every trial, we present two brief successive inputs to the network, with an intervening delay. Each input can take either of two values, labelled A and B respectively. The task is to determine whether the two successive inputs are identical (AA or BB), in which case the network should output -1; or different (AB or BA), in which case the network should output 1 (see Methods for a detailed description). This simple task exhibits several interesting features. First, it is arguably the simplest possible flexible decision task: on sensing the second stimulus, the network must flexibly produce a different response depending on the identity of the first stimulus. Second, because the intervening delay is much longer than the neural time constant, the network must maintain some memory of the first stimulus before the second stimulus arises. Third, to solve this task, some neurons in the network must necessarily possess some form of nonlinear mixed selectivity (note that the problem is in essence a delayed exclusive-or problem), a hallmark of neural activities in prefrontal cortices (Rigotti et al. 2013).

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The networks consistently learn to perform the task with high accuracy. Figure 1 shows the time course of the median error over 20 training runs, each starting with a different randomly initialized network. Shaded area indicates 1st and 3rd quartile over the 20 runs. To define a measure of successful performance, we set a criterion of 95% "correct" responses (i.e., correct sign of the output cell activity) over 100 successive trials ($p<10^{-20}$ under random choice, binomial test). The median time to criterion across 20 runs is 843 trials (inter-quartile range: 692-1125). Response error reliably converges towards a very low residual value.

How does the network represent and maintain traces of incoming stimuli? One possibility is that certain neurons encode stimulus identity by maintaining a stable "register" value over time, such that the firing rate of certain cells directly specify stimulus identity in a relatively time-independent manner. By contrast, physiological studies suggest that neural coding during a working memory task is highly dynamic, with stimulus identity being represented by widely fluctuating cell activations, in such a way that the tuning of individual neurons significantly changes over the course of a trial (Meyers et al. 2008; Barak, Tsodyks, and Romo 2010; Stokes et al. 2013).

To analyze the encoding and maintenance of stimulus identity over time in the network, we used a cross-temporal classification approach (Meyers et al. 2008; Stokes et al. 2013; Dehaene and King 2016). We trained a maximum-correlation classifier to decode various task-relevant features (identity of 1st and 2nd stimulus, and final response), based on whole-population activity at any given time, and then used these time-specific classifiers to try and extract the same task-relevant features at all possible points in time. This method can detect not only whether the network encodes a certain task-relevant variable, but also whether the representation of this variable changes over time (see Methods).

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

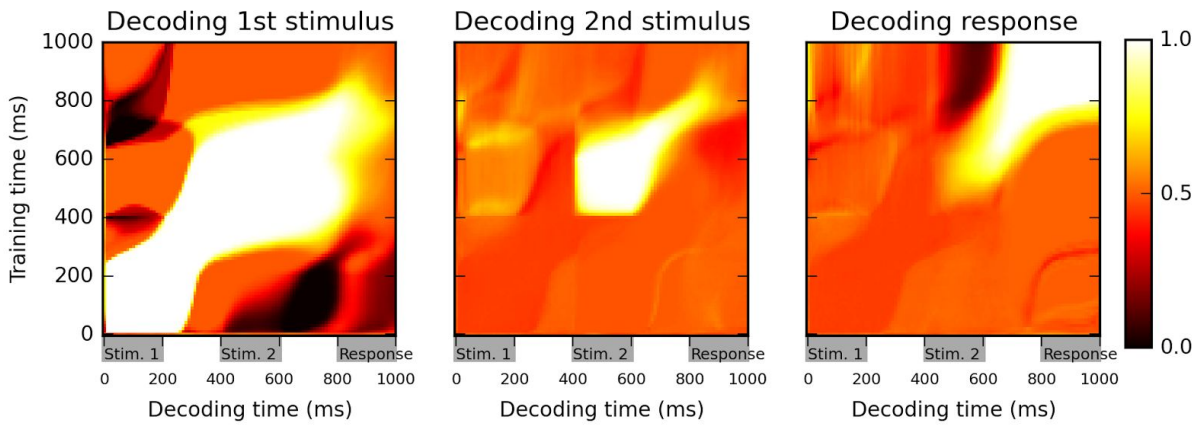


Figure 2. Cross-temporal classification performance reveals dynamic coding. Cross-temporal classification of 1st stimulus identity (left panel), 2nd stimulus identity (middle panel) and network response (right panel). Row i and column j of each matrix indicates the accuracy of a classifier, trained on population activity data at time i, in guessing a specific task feature using population activity data at time j (training and decoding data are always separate). While the network reliably encodes information about stimulus identity right until the onset of the response period (as shown by high accuracy values along the diagonal in left and middle panel), this information is stored with a highly dynamic encoding (as shown by low cross-classification accuracy across successive periods, i.e., "bottlenecks" with high accuracy on the diagonal but low accuracy away from the diagonal). Note that in both left and middle panel, stimulus identity information decreases greatly at the onset of the response period, reflecting a shift to a from stimulus-specific to response-specific encoding (see also Figure 3).

The results in Figure 2 suggest highly dynamic representation of stimuli by the network. For example, the identity of the first stimulus can be successfully decoded during both first and second stimulus presentation, as well as during the intervening delay, as shown by high classification accuracy values on the diagonal during this entire period (Figure 2, left panel). However, the cross-temporal classification performance between these two periods, as seen on the off-diagonal areas (for example, in the areas at 0-200ms on one axis and 400-600 on the other, corresponding to training the classifier based on data from one stimulus presentation and

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

testing it on data from the other stimulus presentation) is essentially at chance level (accuracy ~ .5), or even below chance (dark patches). This suggests that while the network reliably encodes information about 1st-stimulus identity across the first 800ms of the trial, the way in which this identity is represented changes widely between successive periods within the trial. Similarly, 2nd-stimulus identity is maintained from its onset until the beginning of the response period, but in a dynamical manner (low off-diagonal, cross-temporal accuracy between the 400-600ms period and the 600-800ms period, in comparison to the high diagonal accuracy over the entire 400-800ms period).

Another feature of these plots is that the accuracy of stimulus identity decoding strongly decreases over the course of the "response" period (low values along the diagonal for the 800-1000ms in 1st and 2nd panel of Figure 2). This suggests that the network largely stops maintaining information about the specific identity of previous stimuli, and instead encodes solely the actual response, as shown by the very strong classification accuracy in the upper-right portion of the third panel.

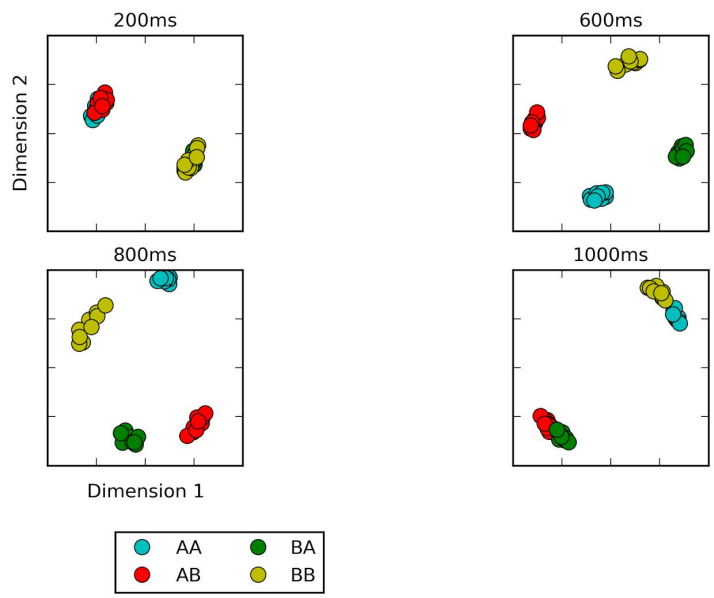


Figure 3. Multi-dimensional scaling plots of population activity reflect shifting encodings of task-relevant information. Population response vectors at various points in time (color-coded

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

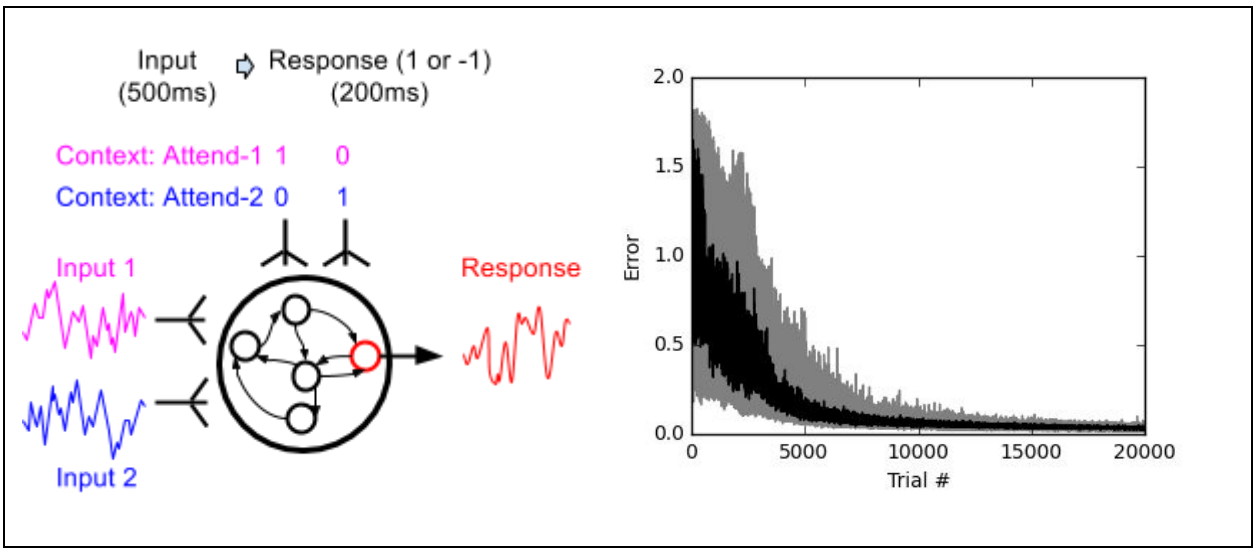
by stimulus combination) are projected in two dimensions while preserving distances between data points as much as possible, using multi-dimensional scaling. At the end of the first stimulus presentation (200ms), population states are firmly separated by first stimulus identity, as expected. After second stimulus presentation (600ms), all four possible stimulus combinations lead to clearly separated population activity states. However, population states corresponding to different responses start to cluster together at the onset of the response period (800ms). Late in the response period (1000ms), population trajectories corresponding to the same response (AA and BB, or BA and AB) have largely merged together, reflecting a shift from stimulus-specific to response-specific representation and a successful "routing" of individual stimulus-specific states to the adequate response-specific state.

To test this interpretation, following (Stokes et al. 2013), we produce Multi-dimensional scaling (MDS) plots of population activity at different time points and for different stimulus conditions (Figure 3). MDS attempts to find a two-dimensional projection such that the distance between any two data points is as similar as possible to their actual distance in the full-dimensional space: nearby (distant) population states should thus produce nearby (distant) points on the MDS plot. Early in the trial, all possible stimulus identity combinations generate different, consistent trajectories, indicating stimulus-dependent encoding. By the late response period, however (1000ms), the trajectories have essentially merged into two clusters, corresponding to the network response ("same" or "different") and largely erasing any distinction based on specific identity of either first or second stimulus. Thus, during the response period, the network flexibly moves from a stimulus-specific representation to a response-specific representation: the stimulus-specific response is flexibly routed to the appropriate, context-dependent response state, as previously observed in cortical activity during a flexible association task (Stokes et al. 2013).

Task 2: Flexible selective integration of sensory inputs

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

An important aspect of cognitive control is the ability to attend selectively to specific portions of the sensory input, while ignoring the rest, in a flexible manner. Recently Mante, Sussillo and colleagues have studied the neural basis of this ability in macaque monkey prefrontal cortex (Mante et al. 2013). Monkeys were trained to report either the dominant color or the dominant motion direction of randomly-moving colored dots. Thus, the same stimulus could entail different appropriate responses depending on current context (i.e. which modality - color or motion - was relevant for this trial). Furthermore, due to the noisy stimulus, the task required selective temporal integration of the relevant sensory input. In addition to neural recordings, Mante and colleagues also trained a recurrent neural network to perform the same task, using supervised learning based on Hessian-free optimization. By analyzing the trained network, they identified mechanisms for selective integration of flexibly-specified inputs in a single network (Mante et al. 2013). This task was also used as an example application by Song and colleagues for their recurrent network training framework (Song, Yang, and Wang 2016).



bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

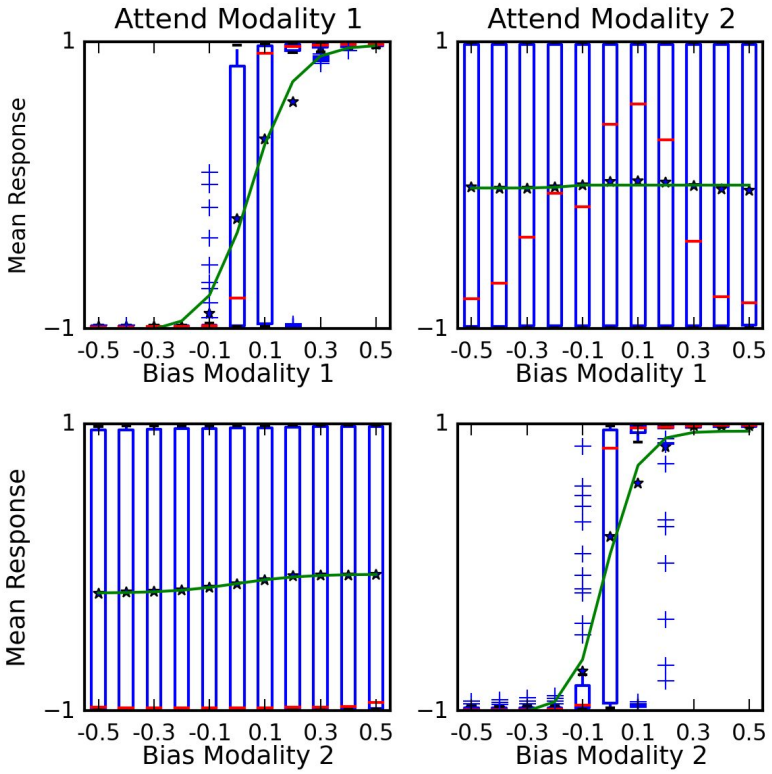


Figure 4. Selective integration task. Top: task description. The network receives two noisy inputs, simulating sensory information coming from two different modalities, as well as two "context" inputs to indicate which of the two sensory inputs must be attended. The task is to produce output 1 if the cued input has a positive mean, and -1 if the cued input has negative mean; this task requires both selective attention and temporal integration of the attended input. Bottom: Psychometric curves of network responses. Top-left panel: average response when context requires attending to modality 1, sorted by the bias of modality 1 inputs. The network response correctly tracks the overall bias of modality 1 inputs. Bottom-left panel: same data, but sorted by modality 2 bias. Network response is mostly unaffected by modality 2 bias, as expected since the network is required to attend to modality 1 only. Right panels: network responses when context requires attending to modality 2. Again, the network correctly identifies the direction of the relevant modality while mostly ignoring the irrelevant modality.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

We trained a network to perform the same task, using our proposed plasticity rule (see Figure 4). Our settings are deliberately similar to those described by Mante, Sussillo and colleagues. The network has two "sensory" inputs (representing the two stimulus modalities of motion and color), implemented as random (Gaussian) time series, with a randomly chosen mean for each trial; the mean of each time-series thus represent the "value" of the corresponding modality for this trial. In addition, two "context" inputs specify which modality is relevant for each trial. The network must produce output 1 if the context-indicated sensory input has positive mean, or -1 if it has negative mean (see Methods for a detailed description).

Figure 4 shows the psychometric curves of a fully-trained network, that is, the mean response as a function of stimulus value. For either modality, we show separate psychometric curves for when this modality was the relevant one and when it was irrelevant. When trials are sorted according to the value of the relevant modality, responses form a steep sigmoid curve with a relatively sharp transition between -1 and +1 centered roughly at 0. By contrast, when trials are sorted according to the value of the irrelevant modality, responses are evenly distributed across the entire range. Thus, the network accurately responds to the relevant signal, while largely ignoring the irrelevant one in each context (Compare to Figure Extended Data 2 in (Mante et al. 2013)). This indicates that the network has learned not only to perform temporal integration of an ambiguous, stochastic input, but also to flexibly "attend" to different input streams depending on context.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

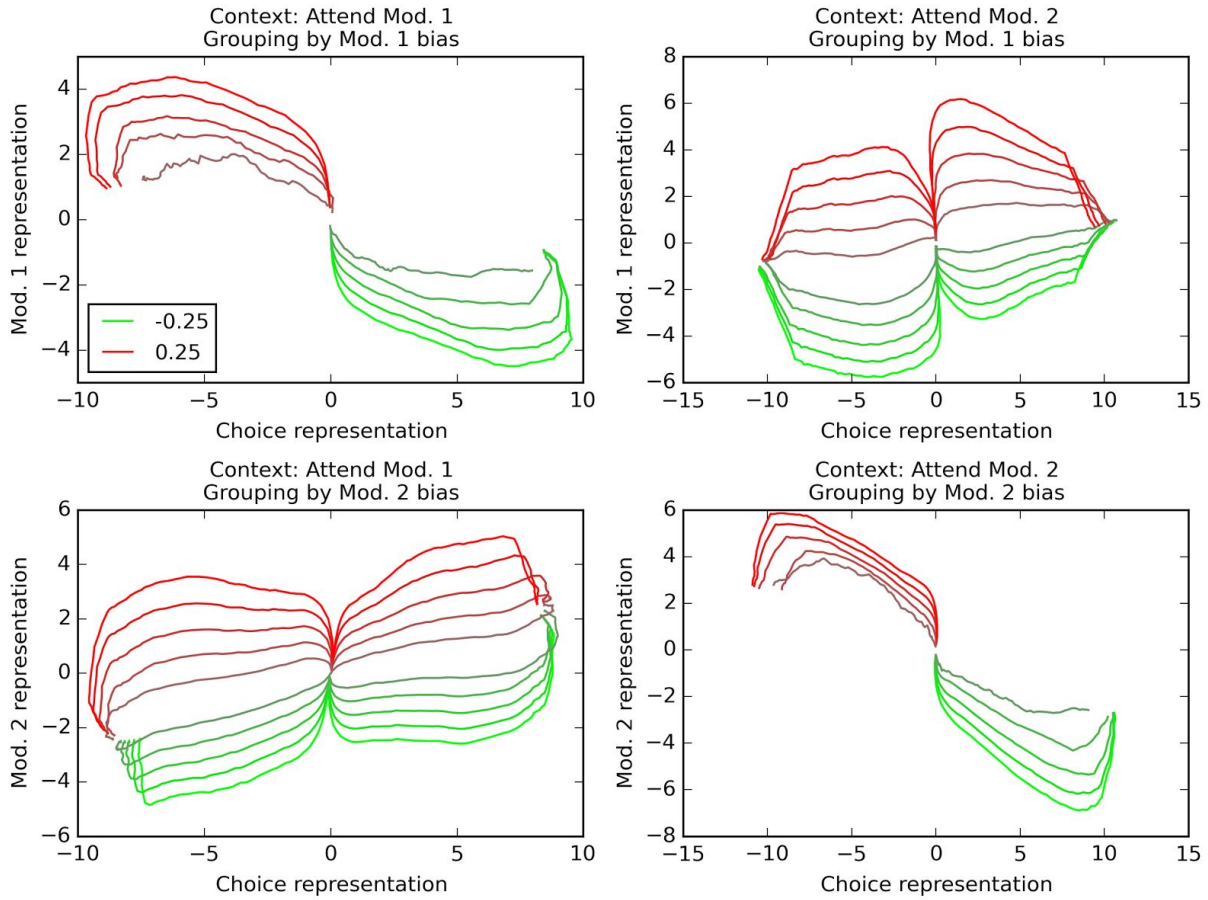


Figure 5. Orthogonal decoding of population activities. Population responses of trials having the same context, the same final choice and the same modality 1 bias (from -0.25 to 0.25) are averaged at each point in time, resulting in different trajectories (one per modality bias value), and projected over dimensions indicating how much the network encodes modality 1 value, modality 2 value, and final choice (the latter always on the x axis). Only correct trials are used (thus top-left and bottom-right panels only have 10 trajectories). The trajectories reveal that the network encodes both the relevant and the irrelevant modality, though only the relevant one is linked to final choice. See text for details.

How is information represented in the network over time? We use Mante & Sussillo's orthogonal decoding procedure, which seeks to extract independent measures of how various task features (stimulus values, context, decision) are encoded in the network (see Methods). Briefly, this

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

method consists in using multiple linear regression measure how much certain task-relevant features are being *independently* represented by the network at any time (see Methods for a detailed description). The results are shown in Figure 5 (compare to Figures 2, 5 in (Mante et al. 2013)). These trajectories plot the evolution of network information over time, for various combinations of context and task feature. Each trajectory represents the average population activity, at successive points in time, of all (correct) trials that have the same mean value for a certain modality and resulted in the same final choice; we project these averaged population trajectories along the orthogonal feature dimensions extracted by orthogonal decoding, and plot the resulting trajectories in feature dimension subspaces ('final choice' dimension is always used as the horizontal axis, while the vertical axis may be either of the two sensory modality dimensions).

As observed in cortical recordings (Mante et al. 2013), these trajectories reveal that *both* the relevant and the irrelevant modality are actually represented in the network: the trajectories for varying value of either modality form an ordered progression in the corresponding "modality" dimension (y-axis), even when that modality is irrelevant (bottom-left and top-right panels in Figure 5); however, only the relevant modality correlates with representation of final choice (compare panels where trajectories are separated by value of the relevant vs. irrelevant modality), in accordance with physiological observations (Mante et al. 2013). This confirms that the network has learnt to selectively integrate the context-indicated variable while discarding the irrelevant one for each trial.

Task 3: Controlling a musculoskeletal model of the human arm

In both of the previous tasks, the network output was a single response channel. However, flexible behavior often requires coordinating multiple outputs, especially during movement. To test whether our plasticity rule can produce coordinated multiplexed responses, we trained a network to control a biomechanical model of the human arm. The model is a custom

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

modification of the one described in (Saul et al. 2015) (itself an extension of (Holzbaur, Murray, and Delp 2005)) and uses the Thelen muscle model (Thelen 2003). The model implements the human upper skeleton, with 4 degrees of freedom (3 at the shoulder, 1 at the elbow), actuated by 16 muscles attached to the shoulder, chest, and upper and lower arm bones. Each of the 16 muscles is controlled by a specific network output cell. The task is to reach towards one of two spherical targets, located in front of the body on either side of the sagittal plane. The appropriate target ball for each trial is indicated by two input channels, set either to 1 and 0 (left-side target) or to 0 and 1 (right-side target) respectively for the entire duration of the trial (700 ms). No other inputs are provided to the system. The error at the end of each trial is measured by the absolute distance between the tip of the hand and the center of the target ball, plus a small penalty for total muscle activation over the entire trial. Note that while the target balls are symmetrically arranged with regard to the body, they are not symmetrical with regard to the right arm (which is the one we model): the right-side ball is closer than the left-side one, and thus reaching either target requires qualitatively different movements.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

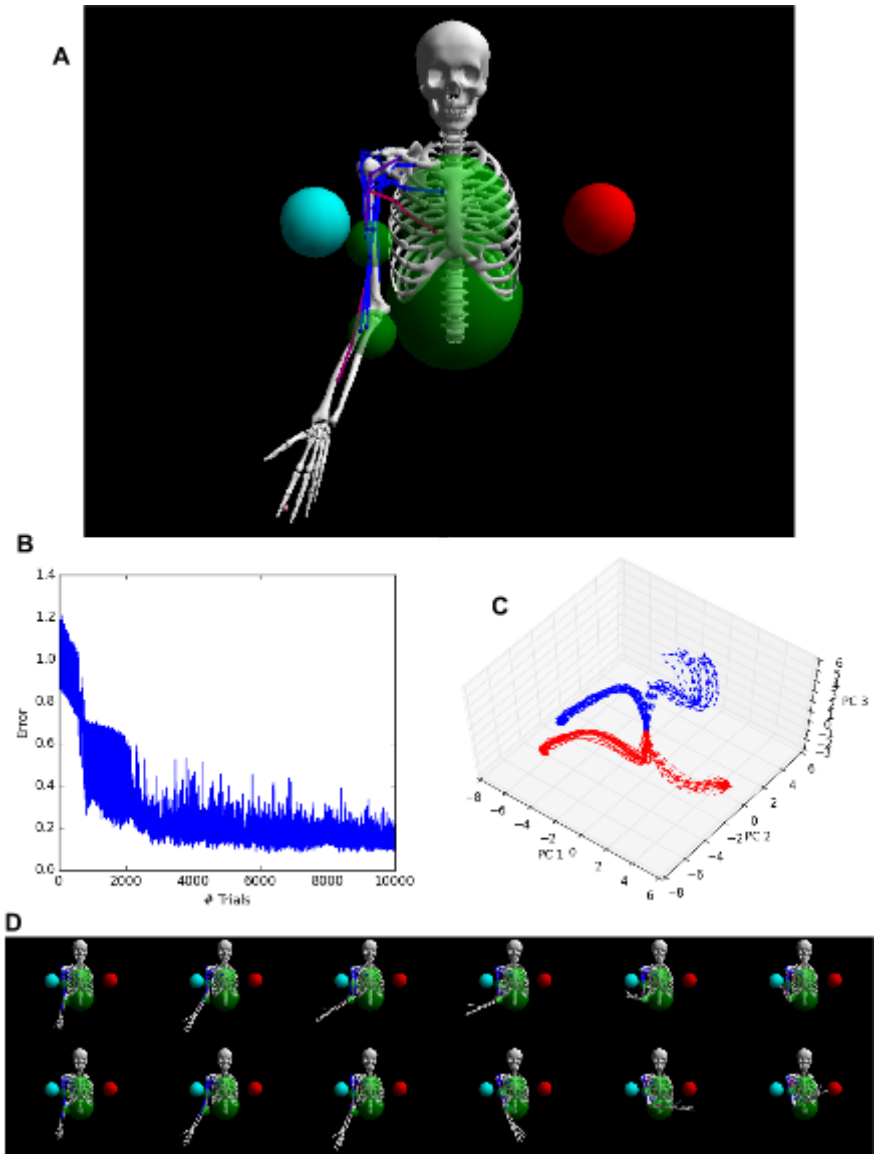


Figure 6. Controlling a biophysical model of the human arm. A: a model of the human upper skeleton with 4 degrees of freedom (shoulder and elbow joints), actuated by 16 muscles at the shoulder, chest and arm (colored strings: blue indicates low activation, red indicates high activation). The task is to reach either of two target balls (blue or red), depending on a context input. B: During training, the error (measured by the distance between tip of hand and target ball at the end of each trial) improves immediately and reaches a low residual plateau after about 3000 trials. C: Projection of network responses along the first three principal

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

components of network activity, both before and after training (blue: left-target, red: right-target, dotted lines: untrained network, solid lines: trained network). D: frame-by-frame illustrations of a right-target trial (top row) and a left-target trial (bottom row), after training.

Results are shown in Figure 6. Initially, as expected, the untrained network performs random, aimless movements, resulting in high initial error (Figure 6B). Performance improves almost from the start of the training process, reaching a low residual error after about 3000 trials. To visualize the impact of training on the dynamics of population activity, we project the population activity over its first three principal components at successive points in time over the course of each trial, using 16 trials for either target context, both before and after training (i.e., 64 trials in total). The projection shows that training considerably alters network trajectories (Figure 6C). The fully trained network correctly reaches the adequate target according to context (Figure 6D).

Discussion

This paper makes three contributions:

1- We introduce a biologically plausible learning algorithm that can train a recurrent neural network to learn flexible (input-specified) tasks, using only time-sparse, delayed rewards and synapse-local information to guide learning.

2- We show that this rule can train network for relatively complex tasks, requiring memory maintenance, selective attention, and coordination of multiple outputs.

3- We show that the trained networks exhibit features of neural activity observed in primate higher cortex during similar tasks. In particular, we demonstrate highly dynamic population-wide encoding of task-relevant information, as observed in neural recordings (Meyers et al. 2008; Stokes et al. 2013; Barak, Tsodyks, and Romo 2010); and we show that selective integration of sensory inputs occurs as described in both observational and modelling studies of primate prefrontal cortex during a similar selective attention task (Mante et al. 2013). In our view, the fact that these features of cortical activity arise spontaneously in networks trained with a

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

biologically plausible rule (as opposed to training the network to directly reproduce observed neural activity traces) increases the plausibility of recurrent neural networks as a model of cortical computation, during both performance and learning of cognitive tasks.

Our proposed plasticity rule implements reward-modulated Hebbian learning between inputs, outputs, and rewards, with the crucial introduction of a supralinear amplification applied to the Hebbian plasticity increments (see Methods). In other words, we posit that plasticity is dominated by large correlations of inputs and outputs, while smaller ones are relatively ignored. This hypothesis of non-linear effects in Hebbian plasticity allows our rule to support robust learning in highly dynamic networks, without requiring non-Hebbian plasticity between segregated driving and perturbatory inputs (Fiete, Fee, and Seung 2007), or a continuous, real-time reward signal (Legenstein et al. 2010; Hoerzer, Legenstein, and Maass 2014) (see Methods and Appendix). We note that this suggestion is similar to the independent proposal of so-called thresholded Hebbian rules (Soltoggio and Steil 2013), in which plasticity is only triggered if the Hebbian product reaches a certain threshold.

The flexible, dynamic coding observed in prefrontal activity has led to suggestions that cortex implements “silent” memory traces by using short-term synaptic plasticity (Barak, Tsodyks, and Romo 2010; Stokes 2015). Short-term synaptic plasticity clearly plays an important role in neural responses, and may well play an important role in maintaining a “hidden internal state” of the network (Buonomano and Maass 2009). However, our network does not implement short-term synaptic plasticity; no weight modification occurs during the course of a trial (all learning occurs between trials), and all the decoding results reported above were obtained with frozen synaptic weights. Our results suggest that the highly dynamic activities spontaneously produced by near-chaotic recurrent networks can be harnessed to produce the dynamic encodings observed in experiments, using only sparse, delayed rewards and biologically plausible plasticity rules. Thus, while short-term synaptic plasticity clearly affects neural responses, it may not be required to explain the highly dynamic nature of working-memory encodings.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

It is unlikely that cortical connectivity should be drastically and finely remodeled through a long training process for any new task. For example, while monkeys require extensive training to perform decision tasks, human subjects can quickly perform new tasks simply by verbal instruction. Rather, it is more likely that the process of slow, reward-modulated synaptic modification in cortical circuitry depicted here reflects the learning of functional networks capable of implementing a certain *type* of task (or cognitive ability), which must then be activated and parameterized for each instance of the task. The latter process of flexible task specification is likely to involve not just other cortical areas, but also the basal ganglia and dopamine system. Elucidating the interactions between cortical, limbic, and dopaminergic structures is an important future task for the study of flexible behavior and its neural implementation.

Methods

Model description

Here we provide a full description of our model and proposed plasticity rule, with an emphasis on implementation details. In the Appendix, we provide an extended discussion at a more intuitive level. Note that the source code for all simulations reported here is available online at http://github.com/ThomasMiconi/BiologicallyPlausibleLearningRNN.

As mentioned above, our model is a fully-connected continuous-time recurrent neural network of N neurons, governed by the classical RNN equations (Sompolinsky, Crisanti, and Sommers 1988; Sussillo and Abbott 2009; Jaeger 2001; Maass, Natschläger, and Markram 2002):

$$\tau \frac{dx_i}{dt} = -x_i(t) + \sum_{j=1}^{N} J_{j,i} r_j(t) + \sum_{k=1}^{M} B_{k,i} u_k(t)$$

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$r_i(t) = tanh(x_i(t))$$

where x_i is the excitation (or "potential") of neuron i, r_i is its response (or "firing rate" / activity), $J_{j,i}$ is the connection weight from neuron j to neuron i, $u_k(t)$ is the current value of each of the M external inputs to the network, and $B_{k,i}$ is the connection weight from external input k to neuron i (τ is the relaxation time constant of neuron activation). In addition, four arbitrarily chosen neurons have a constant activation x=1 and thus provide a bias input to other neurons. There is no separate feedback or output network. Instead, one or more neurons in the network are arbitrarily designated as the "output" neurons, and their responses at any given time are used as the network's response (these neurons are otherwise identical to all others). J is initialized with weights taken from a normal distribution with mean 0 and variance g^2/N, while the input weights $B_{k,i}$ are fixed and taken from a uniform distribution over the [-1,1] interval. Activations x_i are initialized at the start of every trial with uniform noise in the [-0.1, 0.1] range. For the simulations reported here, N=200 (400 for the motor control task), τ=30ms, and g=1.5. Note that the latter value places the networks in the early chaotic regime, where the long-term behavior generally remains non-periodic (Sompolinsky, Crisanti, and Sommers 1988).

Synapses between neurons are modified according to a novel form of reward-modulated Hebbian learning, as described below. First, in order to produce exploratory variation in network responses across trials, each neuron in the network occasionally receives a random perturbation $\Delta_i(t)$ to its activation; these perturbations are not segregated from "normal" inputs (in contrast to (Fiete and Seung 2006; Fiete, Fee, and Seung 2007)). Note that $\Delta_i(t)$ might also represent random noise, or a "teaching" signal from a different area. In this paper, $\Delta_i(t)$ is taken from a uniform distribution within the [-.5, .5] range, occurring randomly and independently for each neuron with a mean rate of 3Hz.

During a trial, at every time step, every synapse from neuron *i* to neuron *j* accumulates a *potential* Hebbian weight change (also called eligibility trace (Izhikevich 2007)) according to the following equation:

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$e_{i,j}(t) = e_{i,j}(t-1) + S(\, r_i(t-1) * (x_j(t) - \bar{x}_j)\,) \qquad \text{[Equation 3]}$$

Remember that r_i represents the output of neuron i, and thus the current input at this synapse. x_j represents the activation of neuron j and $\underline{x}_j$ represents a short-term running average of x_j, and thus $x(t)$ - $\underline{x}$ tracks the fast fluctuations of neuron output. Thus this rule is essentially Hebbian, based on the product of inputs and output (fluctuations). Crucially, S is a monotonic, supralinear function of its inputs; in other words, we posit that the plasticity mechanism is dominated by large increments, and tends to suppress smaller ones. The particular choice of S is not critical, as long as it is supralinear. In this paper we simply used the cubic function $S(x) = x^3$. Sign-preserving squaring $S(x) = x|x|$ also gives satisfactory results; however, simply using the identity function fails to produce learning. The supralinear amplification of plasticity events allows our learning rule to successfully learn from instantaneous deviations of activity, using only sparse, delayed rewards, without requiring a continuous, real-time reward signal (Legenstein et al. 2010; Hoerzer, Legenstein, and Maass 2014); see Discussion and Appendix.

Note that the eligibility trace for any synapse is accumulated over the course of a trial, with each new timestep adding a small increment to the synapse's eligibility trace / potential weight change.

At the end of each trial, a certain reward R is issued to the network, based on the network's performance for this trial as determined by the specific task. From this reward, the system computes a *reward prediction error* signal, as observed in physiological experiments, by subtracting the expected reward for this trial $\underline{R}$ (see below for computation of $\underline{R}$) from the actually received reward R. This reward-prediction signal is used to modulate the eligibility trace into an actual weight change:

$$\Delta J_{i,j} = \eta\, e_{i,j}(R - \bar{R}) \qquad \text{[Equation 4]}$$

where η is a learning rate constant, set to 0.5 for all simulations described here.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

To compute the reward prediction error signal (R-$\underline{R}$), we need to estimate the expected reward $\underline{R}$. Following (Frémaux, Sprekeler, and Gerstner 2010), we simply maintain a running average of recent rewards for trials of the same type (where trial type is determined by the combination of inputs). As (Frémaux, Sprekeler, and Gerstner 2010) pointed out, it is important that separate traces should be maintained for each trial type, so as to provide an accurate estimation of the expected reward $\underline{R}$ for each trial. Thus, after the *n*-th trial of a given type, $\underline{R}$ is updated as follows:

$$\overline{R}(n) = \alpha_{trace}\,\overline{R}(n-1) + (1-\alpha_{trace})\,R(n) \qquad \text{[Equation 5]}$$

Where R(n) is the reward for this trial, and $\underline{R}$(n-1) was the expected reward after the previous trial of the same type. In all simulations, $\alpha_{trace} = 0.33$.

To stabilize learning, we clip the weight modifications for each trial to have a maximum absolute value of 10^{-4} (across experiments, roughly 10% of all potential weight modifications exceed this value and are clipped).

Delayed nonmatch-to-sample task

The first task considered here is a simple delayed nonmatch-to-sample problem (Figure 1). In every trial, we present two brief successive inputs to the network, with an intervening delay. Each input can take either of two values, labelled A and B respectively. The task is to determine whether the two successive inputs are identical (AA or BB), in which case the network should output -1; or different (AB or BA), in which case the network should output 1. We specify the input stimuli by using two different input channels u1 and u2; the identity of the input stimulus is determined by which channel is activated (i.e., for stimulus A, u1=1 and u2=0; for stimulus B, u1=0 and u2=1; remember that each input channel u_k is transmitted to the network by its own independent set of weights - see Methods). In every trial, the first stimulus is presented for 200

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ms, then after a 200 ms delay the second stimulus is presented for 200 ms. Outside of input presentation periods, both input channels are set to 0. The trial goes on for an additional 400ms, thus each trial is 1000ms long. The network's overall response is determined by the activity of the arbitrarily chosen output neuron over the last 200 ms of the trial (the so-called "response" period). The overall error for this trial is the average *absolute* difference between the network's output (that is, the activity of the output neuron) and the target response (1 or -1 depending on presented stimuli), over these last 200ms.

For details on how the network activity was analyzed for Figures 2 and 3, see below.

Selective integration of context-cued sensory inputs

This task was introduced by Mante and colleagues (Mante et al. 2013). In this study, monkeys looked at randomly-moving colored dots, in which both the value and coherence of motion direction and dot color varied from trial to trial. Monkeys had to report the dominant motion direction, or the dominant color, according to current task conditions; thus, the same stimulus could entail different appropriate responses depending on current context. Furthermore, due to the noisy stimulus, the task required temporal integration of sensory input. Importantly, the authors showed that prefrontal neurons registered inputs from both the relevant and the irrelevant modality; however, inputs from the irrelevant modality had no long-term impact on neural activities, while inputs from the relevant modality were selectively integrated over time. Thus, only information from the relevant modality contributed to the final decision.

Our settings are deliberately similar to those described by Mante, Sussillo and colleagues in their neural network implementation of the task, and Song and colleagues in their own implementation (Song, Yang, and Wang 2016). The network has two "sensory" inputs (representing the two stimulus modalities of motion and color) and two "context" inputs (which specify which of the two modalities must be attended to). The sensory inputs are noisy time-series, centered on a specific mean which indicates the "value" of this input for this trial. More precisely, each of the

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

two sensory inputs is sampled at each time step from a Gaussian variable with variance 1 and a mean, or bias, randomly set at either -0.5 or 0.5 for each trial (this is for the learning phase; for the testing phase used to generate the psychometric curves in Figure 4, the bias is varied in increments of 0.1 from -0.5 to 0.5, inclusive). The mean/bias of the Gaussian (positive or negative) represents the "direction" or "value" of the corresponding sensory input (left vs. right, or red vs. green). The context inputs are set to 1 and 0, or 0 and 1, to indicate the relevant modality for this trial. The goal of the network is to determine whether the sensory input in the relevant modality has positive or negative mean.

Sensory inputs are presented for the first 500ms of the trial, followed by a 200ms response period during which all sensory inputs are set to 0. The expected response of the network is 1 if the relevant sensory input has positive mean, and -1 otherwise; thus the same sensory input can entail different appropriate responses depending on the context. As for the previous task, the network's response for a trial is the firing rate of the arbitrarily chosen output cell, and the error for a trial is the average absolute difference between the firing rate of this output cell and the appropriate response for this trial (either -1 or 1) over the 200ms response period.

For details of network analysis (Figure 5), see below.

Analysis of network activity

1- Decoding of network information in a delayed nonmatch-to-sample task

In the delayed nonmatch-to-sample task, we used a cross-temporal classification analysis (Meyers et al. 2008; Stokes et al. 2013) to investigate how fully trained networks encode information over time (Figure 2). The interpretation of these cross-temporal decoding accuracy matrices is that they tell us not only whether the network encodes a certain task-relevant variable, but also whether it uses similar representations to encode this variable at different points in time. If we train one such classifier using data at time *t* in some trials, and then use it to decode

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

population activity from data at the same time *t* in other trials, then decoding accuracy measures how strongly the network encodes the feature at that time point *t*. However, when the decoder is trained on data at time t_{learn} and then applied to population activity data at time t_{decode}, the resulting accuracy measures the stability in the network's "neural code" for this feature across both time points, i.e., how similarly the decoded feature is represented by the network across these time points. If representations are similar across both time points (that is, if the network use similar patterns to represent each possible value of the feature across both time points), then classifiers successfully trained with population activities at time t_{learn} should also produce accurate decoding of population activities at time t_{decode}. By contrast, if the network uses different representations/encoding of task features at these two time points, cross-temporal accuracy should be poor; this should be represented as "bottlenecks" of high accuracy on the cross-temporal decoding plots, whereby information is high along the diagonal (i.e. the feature is indeed encoded by the network at that given time), but away-from-diagonal (cross-temporal) decoding accuracy is low. This is precisely what we observe in Figure 2.

We follow the maximal-correlation classifier approach described in (Meyers et al. 2008) as closely as possible. Briefly, we want to measure how well a certain task-relevant feature (identity of 1st presented stimulus, or identity of 2nd presented stimulus, or final response) can be predicted by observing network activity at time t_1, using a classifier trained on network activity at time t_2. First, we sample the activation of each neuron, every 10 ms, for each trial. This data is stored in a matrix of 100 rows and 200 columns, indicating the activities (firing rates) of all 200 neurons at each of the 100 sampling times. We first generate 80 trials (20 per possible condition, where "condition" is defined as one of the 4 possible stimulus combination: AA, AB, BA or BB) with a trained network. The time course of neural activity will differ somewhat between successive trials, even for identical conditions, due to noise. Then we iterate the following procedure. For each of all 4 possible conditions, we randomly choose half the trials as "training" trials, and the other half as "testing" or "decoding" trials. The training trials corresponding to the same category that we are trying to decode (for example, all stimuli having the same 1st presented stimulus) are averaged together, pointwise, for each neuron and each time point, giving

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

a "prototype" matrix of activation for each neuron at each timepoint under this category. This training data allows us to decode the category of each testing trial, at each point in time, using maximum-correlation classification, in the following way. We compute the Pearson correlation between each row of each "testing" trial and each row of each "prototype" trial. Each such correlation between row *i* of a testing trial and row *j* of a training category-average tells us how much the population activity at time *i* in the testing trial resembles the average population activity at time *j* for this particular category. We can then select the category for which this correlation is maximal, at each training/testing timepoint pair, as the "decoded" category for each testing trial. For each testing trial, this provides a 100x100 matrix of decoded categories (one for each pair of training and testing timepoints). Of course, each testing trial belongs to only one category, so only one possible answer is correct, and thus we can compute another 100x100 matrix of binary values, indicating whether the decoded category at a given point in the decoding matrix (i.e., for any given pair of training and testing timepoints) is correct. The average of these "correctness matrices", over all testing trials, provides the accuracy in cross-temporal decoding of this category for every training/testing pair of timepoints. We iterate this whole procedure 100 times and average together the resulting "correctness" matrices. The resulting 100x100 matrix indicates at each row *i* and column *j* the proportion of times that the decoded category for population activity at timepoint *j* was correct, using training data from timepoint *i*. This is the matrix shown in each of the panels in Figure 2 (one for each of the three categories to be decoded).

2- Orthogonal decoding of network information during a selective integration task

For the selective integration task, we used the analysis method introduced by Mante, Sussillo and colleagues (Mante et al. 2013), and also used by Song and colleagues (Song, Yang, and Wang 2016) (see Figure 5). Intuitively, the purpose of this method is to estimate how much information the network encodes about different task feature (input value, context, final choice, etc.) *independently* from each other.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

After generating multiple trials under various conditions (context -- that is, relevant modality -- and bias for each modality) with a fully trained network, we regress the activity of each neuron over the values of features of interest (context, value of each modality, and final choice) for each trial. This gives us a set of weights for each neuron, one for each feature, representing how much each feature influences the neuron's firing rate. We then "switch views" by grouping together all such weights for any given feature (200 weights - one per neuron). This in turn produces vectors in neuron population space, along which the feature is in a sense maximally represented (notice that this is quite different from, and not equivalent to, the simpler idea of simply regressing each feature over the firing rates of the neurons across trials). We then orthogonalize these vectors using QR decomposition, to ensure that these representations are as independent from each other as possible. Projecting population activity at a given time over the resulting vectors approximates the network's current estimate of the corresponding feature at that time. For successive time slices, we average network activity vectors corresponding to the same value of bias in a certain modality, a certain attended modality, and a certain final choice. We refer the reader to (Mante et al. 2013) for a complete description of the method.

We project population activity, averaged within various groups of trials, at each point in time, over these decoding axes. The trials are grouped according to final choice, value of one modality (either modality 1 or modality 2), and current context (i.e., relevant modality), and the population activity at each point in time is averaged across all trials within each group. When the resulting averages are projected over the orthogonal feature vectors, they produce trajectories, indicating the network's encoded value for each feature, at each point in time, for trials of this group. Only correct trials are used, and thus certain combinations are impossible (for example, positive value of modality 1 bias, while attending modality 1, with a final choice of -1); this is reflected in the top-left and bottom-right panels of Figure 6, which contain half as many trajectories as the top-right and bottom-left panels.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Appendix

We provide an extended discussion of the learning rule presented here (and other related learning rules) in the Appendix enclosed in the Supplementary Text.

Acknowledgements

We thank W. Einar Gall for useful comments and suggestions. We thank Vishwa Goudar for helpful discussions. We thank H. Francis Song for important insight regarding the computation of state-space trajectories in Figure 5. This work was supported by Neurosciences Research Foundation through funding from The G. Harold and Leila Y. Mathers Charitable Foundation and the William and Jane Walsh Charitable Remainder Unitrust, for which we are grateful.

References

Barak, Omri, David Sussillo, Ranulfo Romo, Misha Tsodyks, and L. F. Abbott. 2013. "From Fixed Points to Chaos: Three Models of Delayed Discrimination." *Progress in Neurobiology* 103 (April): 214–22.

Barak, Omri, Misha Tsodyks, and Ranulfo Romo. 2010. "Neuronal Population Coding of Parametric Working Memory." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (28): 9424–30.

Buonomano, Dean V., and Wolfgang Maass. 2009. "State-Dependent Computations: Spatiotemporal Processing in Cortical Networks." *Nature Reviews. Neuroscience* 10 (2): 113–25.

Churchland, Mark M., John P. Cunningham, Matthew T. Kaufman, Justin D. Foster, Paul Nuyujukian, Stephen I. Ryu, and Krishna V. Shenoy. 2012. "Neural Population Dynamics during Reaching." *Nature* 487 (7405): 51–56.

Dehaene, Stanislas, and Jean-Remi King. 2016. "Decoding the Dynamics of Conscious Perception: The Temporal Generalization Method - Springer." In *Micro-, Meso- and Macro-Dynamics of the Brain*, edited by G. Buzsaki and Y. Christen. Springer.

Fiete, Ila R., Michale S. Fee, and H. Sebastian Seung. 2007. "Model of Birdsong Learning Based on Gradient Estimation by Dynamic Perturbation of Neural Conductances." *Journal of Neurophysiology* 98 (4): 2038–57.

Fiete, Ila R., and H. Sebastian Seung. 2006. "Gradient Learning in Spiking Neural Networks by Dynamic Perturbation of Conductances." *Physical Review Letters* 97 (4): 048104.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Florian, Răzvan V. 2007. "Reinforcement Learning through Modulation of Spike-Timing-Dependent Synaptic Plasticity." *Neural Computation* 19 (6): 1468–1502.

Frémaux, Nicolas, Henning Sprekeler, and Wulfram Gerstner. 2010. "Functional Requirements for Reward-Modulated Spike-Timing-Dependent Plasticity." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (40): 13326–37.

Hennequin, Guillaume, Tim P. Vogels, and Wulfram Gerstner. 2014. "Optimal Control of Transient Dynamics in Balanced Networks Supports Generation of Complex Movements." *Neuron* 82 (6): 1394–1406.

Hoerzer, Gregor M., Robert Legenstein, and Wolfgang Maass. 2014. "Emergence of Complex Computational Structures from Chaotic Neural Networks through Reward-Modulated Hebbian Learning." *Cerebral Cortex* 24 (3): 677–90.

Holzbaur, Katherine R. S., Wendy M. Murray, and Scott L. Delp. 2005. "A Model of the Upper Extremity for Simulating Musculoskeletal Surgery and Analyzing Neuromuscular Control." *Annals of Biomedical Engineering* 33 (6): 829–40.

Izhikevich, Eugene M. 2007. "Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling." *Cerebral Cortex* 17 (10): 2443–52.

Jaeger, Herbert. 2001. "The 'echo State' Approach to Analysing and Training Recurrent Neural Networks – with an Erratum note1." GMD 148. German National Research Center for Information Technology. http://web.info.uvt.ro/~dzaharie/cne2013/proiecte/tehnici/ReservoirComputing/EchoStatesTechRep.pdf.

Jun, Joseph K., Paul Miller, Adrián Hernández, Antonio Zainos, Luis Lemus, Carlos D. Brody, and Ranulfo Romo. 2010. "Heterogenous Population Coding of a Short-Term Memory and Decision Task." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (3): 916–29.

Kober, Jens, J. Andrew Bagnell, and Jan Peters. 2013. "Reinforcement Learning in Robotics: A Survey." *The International Journal of Robotics Research*, August. doi:10.1177/0278364913495721.

Laje, Rodrigo, and Dean V. Buonomano. 2013. "Robust Timing and Motor Patterns by Taming Chaos in Recurrent Neural Networks." *Nature Neuroscience* 16 (7): 925–33.

Legenstein, Robert, Steven M. Chase, Andrew B. Schwartz, and Wolfgang Maass. 2010. "A Reward-Modulated Hebbian Learning Rule Can Explain Experimentally Observed Network Reorganization in a Brain Control Task." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (25): 8400–8410.

Maass, Wolfgang, Thomas Natschläger, and Henry Markram. 2002. "Real-Time Computing without Stable States: A New Framework for Neural Computation Based on Perturbations." *Neural Computation* 14 (11): 2531–60.

Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. "Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex." *Nature* 503 (7474): 78–84.

Meyers, Ethan M., David J. Freedman, Gabriel Kreiman, Earl K. Miller, and Tomaso Poggio. 2008. "Dynamic Population Coding of Category Information in Inferior Temporal and Prefrontal Cortex." *Journal of Neurophysiology* 100 (3): 1407–19.

Mnih, Volodymyr, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. 2014. "Recurrent Models of Visual Attention." In *Advances in Neural Information Processing Systems 27*,

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 2204–12. Curran Associates, Inc.

Pearlmutter, B. A. 1995. "Gradient Calculations for Dynamic Recurrent Neural Networks: A Survey." *IEEE Transactions on Neural Networks / a Publication of the IEEE Neural Networks Council* 6 (5): 1212–28.

Peters, Jan, and Stefan Schaal. 2008. "Reinforcement Learning of Motor Skills with Policy Gradients." *Neural Networks: The Official Journal of the International Neural Network Society* 21 (4): 682–97.

Rajan, Kanaka, Christopher D. Harvey, and David W. Tank. 2016. "Recurrent Network Models of Sequence Generation and Memory." *Neuron* 90 (1): 128–42.

Raposo, David, Matthew T. Kaufman, and Anne K. Churchland. 2014. "A Category-Free Neural Population Supports Evolving Demands during Decision-Making." *Nature Neuroscience* 17 (12): 1784–92.

Rigotti, Mattia, Omri Barak, Melissa R. Warden, Xiao-Jing Wang, Nathaniel D. Daw, Earl K. Miller, and Stefano Fusi. 2013. "The Importance of Mixed Selectivity in Complex Cognitive Tasks." *Nature* 497 (7451): 585–90.

Saul, Katherine R., Xiao Hu, Craig M. Goehler, Meghan E. Vidt, Melissa Daly, Anca Velisar, and Wendy M. Murray. 2015. "Benchmarking of Dynamic Simulation Predictions in Two Software Platforms Using an Upper Limb Musculoskeletal Model." *Computer Methods in Biomechanics and Biomedical Engineering* 18 (13): 1445–58.

Soltoggio, Andrea, and Jochen J. Steil. 2013. "Solving the Distal Reward Problem with Rare Correlations." *Neural Computation* 25 (4): 940–78.

Sompolinsky, Haim, A. Crisanti, and H. J. Sommers. 1988. "Chaos in Random Neural Networks." *Physical Review Letters* 61 (3): 259.

Song, H. Francis, Guangyu R. Yang, and Xiao-Jing Wang. 2016. "Training Excitatory-Inhibitory Recurrent Neural Networks for Cognitive Tasks: A Simple and Flexible Framework." *PLoS Computational Biology* 12 (2): e1004792.

Stokes, Mark G. 2015. "'Activity-Silent' Working Memory in Prefrontal Cortex: A Dynamic Coding Framework." *Trends in Cognitive Sciences* 19 (7). Elsevier: 394–405.

Stokes, Mark G., Makoto Kusunoki, Natasha Sigala, Hamed Nili, David Gaffan, and John Duncan. 2013. "Dynamic Coding for Cognitive Control in Prefrontal Cortex." *Neuron* 78 (2): 364–75.

Sussillo, David, and L. F. Abbott. 2009. "Generating Coherent Patterns of Activity from Chaotic Neural Networks." *Neuron* 63 (4): 544–57.

Sussillo, David, Mark M. Churchland, Matthew T. Kaufman, and Krishna V. Shenoy. 2015. "A Neural Network That Finds a Naturalistic Solution for the Production of Muscle Activity." *Nature Neuroscience* 18 (7): 1025–33.

Thelen, Darryl G. 2003. "Adjustment of Muscle Mechanics Model Parameters to Simulate Dynamic Contractions in Older Adults." *Journal of Biomechanical Engineering* 125 (1): 70–77.

Williams, Ronald J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." *Machine Learning* 8 (3-4). Kluwer Academic Publishers: 229–56.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Text

Appendix

Discussion of learning algorithms in recurrent neural networks

Here we attempt to provide a more intuitive explanation of how the learning rule actually works. We also discuss existing learning algorithms for strongly connected RNNs, including the node-perturbation method (Fiete and Seung 2006) and the Hoerzer-Legenstein-Maas method (Hoerzer, Legenstein, and Maass 2014).

Why does node-perturbation / REINFORCE work?

Node-perturbation consists in applying perturbations to the outputs of neurons, then computing an eligibility trace equal to the accumulated product of the (non-perturbative, "normal") inputs by the perturbations, and multiplying this eligibility trace by a baseline-subtracted reward signal to obtain the actual synaptic modifications. Fiete and Seung (Fiete and Seung 2006; Fiete, Fee, and Seung 2007) showed formally that this method descends the gradient of error over the weights.

Interestingly, this method is largely equivalent to the classical REINFORCE algorithm (see in particular Equation 11 in (Williams 1992), and discussion in (Fiete and Seung 2006)), which is widely used in reinforcement learning. This method can also be shown to descent the gradient of error over the weights with a very different approach (Kober, Bagnell, and Peters 2013; Peters and Schaal 2008). As such, node-perturbation essentially a form of policy gradient search, to use the terminology of reinforcement learning (Kober, Bagnell, and Peters 2013; Peters and Schaal 2008).

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Our previous experiments showed that the node-perturbation method can successfully learn complex tasks in strongly-connected, chaotic RNN (see T. Miconi, arXiv:1507.08973 [q-bio.NC]). The plasticity rule described in this paper is largely an attempt at implementing this method with biologically plausible mechanisms, without requiring an instantaneous, real-time reward signal (as in (Hoerzer, Legenstein, and Maass 2014) - see below).

Why is node-perturbation learning so efficient in learning correct trajectories? While the mathematical derivation of node-perturbation and REINFORCE are well established (Fiete and Seung 2006; Peters and Schaal 2008), here we seek a more intuitive understanding of how this method works.

Briefly, node-perturbation learning imposes a phasic perturbation on the network, then "incorporates" the effect of this perturbation into the network weights if the resulting trajectory turns out to produce a higher reward than expected - or conversely, incorporates the "opposite" of this perturbation if the resulting trajectory turned out to produce lower reward than expected. The effect of the weight modification is to ensure that, next time the same input is presented, the neurons will respond in a way that will be a bit more similar to what the perturbed response was. Conversely, if the trajectory turned out to lead to a lower reward, the weight modification will be "anti-incorporated" into the weights, making sure that the next presentation of the same inputs will be nudged away from the direction of the perturbation.

Consider what happens when a single perturbation z(t) is applied to neuron j at time t. Under node-perturbation, the weight modification dW at a given incoming synapse of neuron j is equal to x(t) * z(t) * DR, where x(t) is the current pre-synaptic input at the synapse, z(t) the perturbation received by the neuron, and DR the "net" reward received at the end of the trial (centered to mean zero over many trials). Suppose that DR is positive (the effect of the perturbation was "good") and z was positive (the neuron received a positive perturbation). The effect of this dW will then be to add a positive multiple of x(t) to the weight Wj,i. As a result, the input weight vector of neuron j will become more correlated with the current population response

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

vector X(t); this will result in a higher response of j next time the input X(t) is presented. Since the perturbation z(t) was positive, this is exactly what is needed to replicate the effect of the perturbation, that is, increase the firing of j. Conversely, if z(t) was negative, the effect would be to subtract a multiple of X(t) from the input weight vector of neuron j, which would reduce the correlation between X(t) and its weight vector (possibly making it more negative if it already was), and therefore reduce the response of j next time input X(t) is presented - again, mimicking the effect of the (negative) perturbation.

The converse applies when DR is negative (that is, when the resulting trajectory was "bad"): the resulting modification will lead to incorporating the opposite of the received perturbation into the incoming weight vector.

Thus, node-perturbation "translates" the trajectory perturbation into a weight modification that would reproduce this perturbed trajectory, then incorporates this modification if the perturbation was "good", or its opposite if the perturbation turned out to be "bad".

Why does simple reward-modulated Hebbian learning not *work (for strongly-connected recurrent networks)?*

Node-perturbation is not Hebbian. In reward-modulated Hebbian learning (RMHL), weight modifications are proportional to the product of pre- and post-synaptic activities, later multiplied by a reward signal. However, node-perturbation computes the product of pre-synaptic activity by *perturbations*, rather than the full post-synaptic activity. As explained above, this allows the synaptic change to "incorporate" the effect of the perturbation if this perturbation led to a higher reward, or incorporate its opposite if the perturbation led to a lower reward.

The problem is that there is no obvious biological mechanism by which the synapse can somehow isolate the "perturbation" inputs from the "normal" inputs. Furthermore, synaptic

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

plasticity is thought to result from a product of inputs and outputs (Hebbian), but node-perturbation requires it to be computed by a product between two inputs ("normal" and "perturbative"). This leads to the question of how node-perturbation could be approximated in a biologically plausible, Hebbian manner - a question to which the present paper attempts to provide a plausible answer.

At first sight it might seem that the two are easily reconciled. Let us write the total reward-modulated Hebbian synaptic weight change after a trial as $x * (y + z) * DR$, where x is pre-synaptic activity, y is pre-perturbation post-synaptic activity, z is the perturbation and DR is the zero-mean reward signal. In the limit of slow weight modifications, for successive presentations of a same input, the total weight change will be $\sum x * (y + z) * DR = \sum (x*y*DR + x*z*DR)$ (where summation is taken across successive trials). Because xy is roughly constant across successive presentations of the same input (for slow weight modification) and DR is centered to have zero mean, the first term vanishes, leaving only the second term x*z*DR. But this remaining term is exactly equal to the equation for node-perturbation learning, as described above. Thus, intuitively, the bulk of the post-synaptic response should "cancel out" due to the zero-mean centering of the reward signal, leaving only the perturbation-proportional term: the only change between RMHL and node-perturbation should thus be a change in variance, rather than in the asymptotic weight value.

However, while the foregoing is valid when inputs and outputs are constant for a given trial (e.g. for very short trials), it does not hold for temporally extended trials in which the input and/or outputs are highly dynamic. This is especially the case when the outputs at time t influence the inputs at times t'>t, as is the case in strongly recurrent networks of the type studied above. In this case, x*y is not constant over successive presentations of the same trial; indeed, a small perturbation may lead to a drastic, unpredictable change in future activity y within the trial. Therefore, the x*y*DR term does not vanish from the summation, despite the fact that DR is zero-centered.

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Under node-perturbation learning, these potentially large trial-to-trial fluctuations in intrinsic activity are ignored, because only the perturbation itself is used in the plasticity rule, which allows its incorporation in the weights. But in reward-modulated Hebbian learning, these fluctuations between successive presentations of the same input will typically dwarf the perturbations and thus dominate the synaptic changes, creating unpredictable weight changes that are unrelated to the direction of the perturbation (a related point is made by Legenstein and colleagues (Legenstein et al. 2010)).

Why does our Hebbian learning rule work with sparse, delayed rewards?

In the present article, we first compute fluctuations of post-synaptic activity by subtracting a very short-term running average (trace) of ongoing activity (Eq. 3). This idea was introduced by Legenstein and colleagues (Legenstein et al. 2010) and applied to recurrent neural networks by Hoerzer and colleagues (Hoerzer, Legenstein, and Maass 2014).

The rationale of this method is that, under reasonable assumptions, fast fluctuations of output activity should be dominated by the exploratory perturbations; thus, by positing that Hebbian plasticity is based on these fluctuations of post-synaptic outputs rather than the full post-synaptic outputs, we can re-create node-perturbation in a purely Hebbian manner.

However, a difficulty of this method is that it requires a continuous, instantaneous reward signal at each point in time (Hoerzer, Legenstein, and Maass 2014). This eliminates a major advantage of reinforcement learning, namely, the ability to learn from sparse, delayed rewards.

The reason why real-time reward signals are needed is that simply subtracting a running average from ongoing activity does not perfectly isolate external perturbations, due to spurious relaxation effects. For example, after a single positive perturbation is applied, the running average is now elevated, and the difference between ongoing activity will now be negative, until the average

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

decays down to the ongoing value. These negative relaxation terms will be accumulated into the Hebbian product and cancel the positive, perturbation-related initial term.

To take a simple example, suppose we apply a single positive perturbation to a flat signal. The perturbation itself cause a sharp positive deviation from the previous activity. Thus, as expected, subtracting ongoing activity from recent average correctly isolates the received perturbation. But after the perturbation has occurred, the running average (which includes the recent perturbation) is now elevated, and thus the next few time steps of (unperturbed) activity will be lower than the trace, creating spurious negative deviations. If the signal remains flat, the sum of negative deviations will in fact equal and fully cancel out the positive deviation caused by the perturbation itself, causing spurious weight changes in the opposite direction of the ones created by the perturbation (assuming slowly changing inputs in comparison to the perturbations).

In the Legenstein / Hoerzer method, this problem is addressed by the assumption of a continuous, real-time reward signal, which *also* undergoes subtraction of a short-term running average. Because the same effect will occur in both the activity trace and the continuous reward trace, the negative "spurious" relaxation deviations in both traces, multiplied with each other, will produce a *positive* addition to the synaptic changes, reinforcing the one created by the perturbation rather than cancelling it. Thus, in the Legenstein/Hoerzer method, an instantaneous real-time reward signal allows learning to occur based on fluctuations of post-synaptic activity.

By contrast, the learning rule described in this paper does not require a continuous, real time reward signal. As shown above, it can learn from sparse, delayed rewards at the end of each trial. The crucial hypothesis in our learning rule is that weight increments are accumulated supralinearly (the S function in Equation 3), essentially positing that synaptic plasticity amplifies larger events and ignores smaller ones.

The reason this hypothesis permits learning with delayed rewards is as follows. First, suppose that we apply a supralinear amplification to the post-synaptic fluctuations themselves (i.e. the

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

running-average-subtracted post-synaptic outputs), rather than to the overall eligibility trace. By positing a supralinear amplification of the post-synaptic fluctuations, we magnify the large deviations from the running average caused by the actual perturbations, and suppress the smaller, countering relaxation deviations that accumulate as the running average falls back towards the underlying signal value. As a result, the perturbation itself is successfully isolated from ongoing activity, without requiring either an explicit separation between "normal" and "perturbative" inputs or a continuous, real-time reward signal; while the spurious relaxation effects are suppressed. The product of perturbations by incoming pre-synaptic activities can now be accumulated over the course of a trial, and multiplied by a single overall reward value at the end of each trial, reproducing the effect of node-perturbation learning in a purely Hebbian manner, and without requiring an instantaneous reward signal at each point in time.

However, it is not easy to see how a biological plasticity mechanism could selectively apply a nonlinear transformation to the fluctuations of post-synaptic outputs. It seems less onerous to assume that the supralinear amplification applies directly to the plasticity increments themselves, which is equivalent to positing that plasticity is dominated by larger Hebbian events and minimizes small ones. This results in the plasticity rule proposed here (Equations 3 and 4). This plausible approximation turns out to produce successful learning, as described above.

Note that this is very similar to recently-proposed thresholded Hebbian rules, whereby plasticity is only triggered by events in which the Hebbian product reaches a certain threshold (Soltoggio and Steil 2013). A supralinear amplification offers a smoother amplification of larger Hebbian events, by comparison to the all-or-nothing effect of a threshold; however, the overall effect is similar: ignore small, possibly incidental correlations of input and output, but retain the larger ones, which are more likely to be informative.

References

Fiete, Ila R., Michale S. Fee, and H. Sebastian Seung. 2007. "Model of Birdsong Learning Based on Gradient Estimation by Dynamic Perturbation of Neural Conductances." *Journal of*

bioRxiv preprint doi: https://doi.org/10.1101/057729; this version posted September 29, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Neurophysiology 98 (4): 2038–57.

Fiete, Ila R., and H. Sebastian Seung. 2006. "Gradient Learning in Spiking Neural Networks by Dynamic Perturbation of Conductances." *Physical Review Letters* 97 (4): 048104.

Hoerzer, Gregor M., Robert Legenstein, and Wolfgang Maass. 2014. "Emergence of Complex Computational Structures from Chaotic Neural Networks through Reward-Modulated Hebbian Learning." *Cerebral Cortex* 24 (3): 677–90.

Kober, Jens, J. Andrew Bagnell, and Jan Peters. 2013. "Reinforcement Learning in Robotics: A Survey." *The International Journal of Robotics Research*, August. doi:10.1177/0278364913495721.

Legenstein, Robert, Steven M. Chase, Andrew B. Schwartz, and Wolfgang Maass. 2010. "A Reward-Modulated Hebbian Learning Rule Can Explain Experimentally Observed Network Reorganization in a Brain Control Task." *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30 (25): 8400–8410.

Peters, Jan, and Stefan Schaal. 2008. "Reinforcement Learning of Motor Skills with Policy Gradients." *Neural Networks: The Official Journal of the International Neural Network Society* 21 (4): 682–97.

Soltoggio, Andrea, and Jochen J. Steil. 2013. "Solving the Distal Reward Problem with Rare Correlations." *Neural Computation* 25 (4): 940–78.

Williams, Ronald J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning." *Machine Learning* 8 (3-4). Kluwer Academic Publishers: 229–56.