Exactly solvable nonlinear recurrent networks for context-dependent information processing

C. H. Stock, S. Lahiri, A. H. Williams, S. Ganguli

Summary. An increasingly common approach to confronting large-scale neural recordings during complex tasks involves training large neural networks to solve the same task. This has led, for example, to highly accurate models of the ventral visual stream [**4**], prefrontal cortex [**1**], and retina [**2**]. However, such network models themselves can be so complex that they defy intuitive understanding, raising profound questions about the nature of explanation in systems neuroscience. For example, what design principles govern how the connectivity and dynamics of these networks endow them with their computational capabilities? Moreover, what would a codification of these design principles even look like? Here we follow a cue from Feynman's dictum: "What I cannot create, I do not understand." In essence, if we really understand how connectivity and dynamics give rise to a computation, we should be able to analytically write down, using pencil and paper, a variety of exactly solvable neural networks that accomplish a given task, instead of obtaining them indirectly through an opaque process of neural network training.

We accomplish this goal for a wide variety of context-dependent information processing tasks in which the same stimulus or behavioral condition can yield different actions depending on prior context. Our framework enables us to take such a task, specified as a set of input-dependent internal state transitions and actions, and generate an ensemble of nonlinear recurrent neural networks that accomplish the requisite transitions through their connectivity and dynamics. As an illustration of our method, we show how to analytically generate a diversity of networks that accomplish a canonical context-dependent task: the Wisconsin card sorting test. Overall, our newly discovered framework for generating exactly solvable nonlinear networks, through a process of analytical forward engineering, provides a foundation for extracting design principles governing how neural connectivity and dynamics can conspire to generate emergent computations.

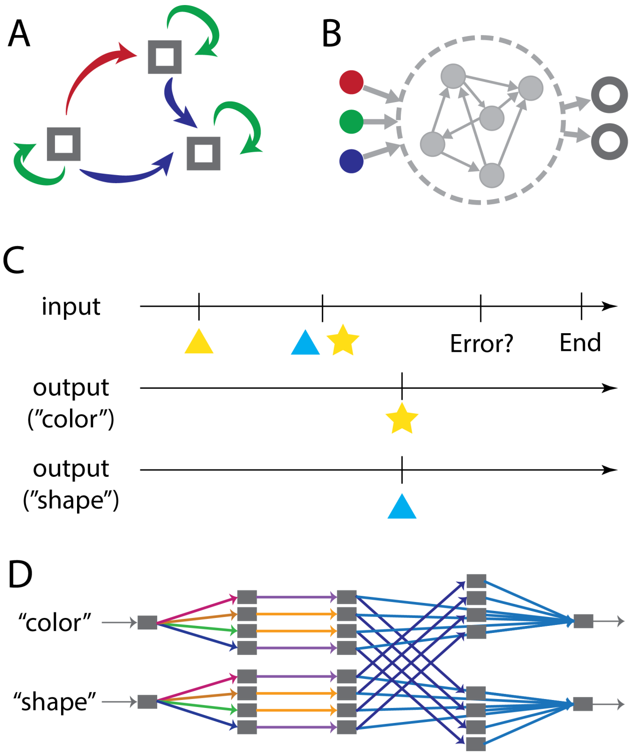

Figure 1: (A) Schematic of FSM. Colored arrows represent transitions under different input conditions. B) Schematic of RNN, including input, recurrent, and readout populations. (C) Timeline of inputs and rule-dependent outputs during a single trial of the WCST. (D) Graph of the FSM we use to perform the WCST, aligned with the timeline in (C). Colored arrows represent transitions under different input conditions.

Construction. We have found a constructive method, along with related theoretical guarantees, to recapitulate the computation implemented by any finite state machine–a general model of context-dependent information processing–in the dynamics of a nonlinear recurrent neural network. A *finite state machine* (FSM) consists of a finite set of states $\mathcal{S}$ (with $\mathrm{S} = |\mathcal{S}|$), a finite set of inputs $\mathcal{V}$ (with $\mathrm{V} = |\mathcal{V}|$), and a transition operator $T : \mathcal{S} \times \mathcal{V} \to \mathcal{S}$, such that the evolution of the state in discrete time is given by $s_{t+1} = T(s_t, v_t)$ for $s_t \in \mathcal{S}$ and $v_t \in \mathcal{V}$. A discrete-time *recurrent neural network* (RNN) consists of an analog N-dimensional state $\{\boldsymbol{x}_t \in \mathbb{R}^N\}_{t=0}^{\infty}$, driven by time-varying inputs $\{\boldsymbol{u}_t \in \mathbb{R}^M\}_{t=0}^{\infty}$, with dynamics $\boldsymbol{x}_{t+1} = \boldsymbol{F}(\boldsymbol{x}_t, \boldsymbol{u}_t) \equiv \boldsymbol{J}\varphi(\boldsymbol{x}_t) + \boldsymbol{W}\boldsymbol{u}_t$, where $\boldsymbol{J}$ is a recurrent weight matrix, $\boldsymbol{W}$ is an input weight matrix, and φ is a nonlinear scalar-valued function operating elementwise on $\boldsymbol{x}_t$. In our work we adopt $\varphi = \tanh(\cdot)$. Fig. 1(A,B) illustrates the FSM and RNN architectures.

At a high level, the construction we propose maps states of the FSM to stable fixed points of the RNN dynamics, and adds input-dependent transitions between those fixed points such that the network dynamics exactly recapitulates the computations implemented by the FSM. The method takes advantage of the fact that all of the dynamical constraints we seek to enforce are linear in $\boldsymbol{J}$ and therefore can be solved via standard linear algebraic techniques.

In the first step, we solve for a $\boldsymbol{J}$ that encodes up to N specified fixed points. In particular we associate S fixed points $\boldsymbol{x}_1^f, \boldsymbol{x}_2^f, \cdots, \boldsymbol{x}_S^f \in \mathbb{R}^N$, in one-to-one correspondence with the S states of the FSM. We ensure these points are fixed points by choosing $\boldsymbol{J}$ via the pseudoinverse learning rule. In addition, in a proof not provided here, we have shown that if the fixed point activities are drawn uniformly at random from $\{-a, a\}$, then the fixed points are indeed *stable* when $\boldsymbol{u} = 0$ and $|a|$ is sufficiently large. Furthermore, this stability result is robust to perturbations from this choice of fixed point activities. We further engineer $\boldsymbol{W}$ and $\boldsymbol{J}$ to enforce input-driven transient dynamics that mimics the requisite transitions in the FSM. Now the end result is an ensemble of analytically defined RNNs, all of which can simulate an arbitrarily chosen FSM. In our construction, the number of required neurons grows only linearly in the number of inputs and number of states.

Example: Wisconsin card-sorting test. To demonstrate the relevance of our method to neuroscience, we implement a version of a well-known human psychophysical test known as the Wisconsin card sorting test (WCST). In the task, the subject is shown a sequence of cards that contain multimodal information (e.g. both colors and shapes), and must match them according to a rule (e.g. “match color” or “match shape”). During the test, the rule itself can switch, which the subject learns via an error signal. In particular, because of the rule dependence, the exact same stimulus sequence can lead to different behaviors on different trials. Thus, the task involves such cognitive processes as working memory, selective attention, rule-based reasoning, and context-dependent perceptual decision making; indeed, this wide range of behaviors can be modeled by a finite state machine. However, previous attempts to obtain neural networks that solve the WCST have relied on iteratively learning the network [**3**], and have not been able to specify fixed points directly. Our technique addresses both of these these shortcomings. Using 28 states and 40 input-dependent transitions, we obtain perfect performance on the task. Fig. 1(B,C) contains a schematic of the task and the FSM graph we use to implement the network.

Robustness of method. To answer the question of whether our method could be extended beyond this example to instantiate arbitrary FSMs in a fully automated fashion, we generated thousands of FSMs and tested the fidelity with which the constructed RNNs could recapitulate their computations. Specifically, we sampled 5,000 random FSMs with $S = 30$ states and $V = 15$ inputs, generated corresponding networks of $N = 500$ neurons, and tested the RNNs at every possible state transition in each network. Remarkably, the networks performed at 100% accuracy in every trial.

Conclusions. Characterizing the organizing principles by which neural connectivity and dynamics conspire to yield emergent computations is a major outstanding project in theoretical neuroscience. We have presented a framework for obtaining closed-form solutions for implementations of finite state machines under the dynamics of nonlinear recurrent neural networks. Thus, we are able to analytically study networks which exhibit complex behaviors, such as context-dependent computation, working memory, and rule-based reasoning. As numerical tools for optimizing recurrent neural networks proliferate in machine learning, we believe that forward engineering will continue to yield uniquely valuable and highly interpretable insights into the interplay of network dynamics, connectivity, and computation in the brain.

References. [**1**] Mante. *Nature* (2013). [**2**] McIntosh. *NIPS* (2016). [**3**] Rigotti. *Frontiers in computational neuroscience* 4 (2010). [**4**] Yamins. *PNAS* (2014).