Identifying Learning Rules from Neural Network Observables

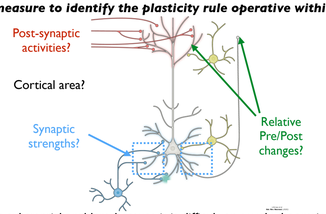
Aran Nayebi[1]*, Sanjana Srivastava[2]*, Surya Ganguli[3,5], Daniel L. K. Yamins[2,4,5]

Neurosciences PhD Program[1], Department of Computer Science[2], Applied Physics[3], Psychology[4], and Wu Tsai Neurosciences Institute[5]
Stanford University, Stanford, CA 94305

Motivation

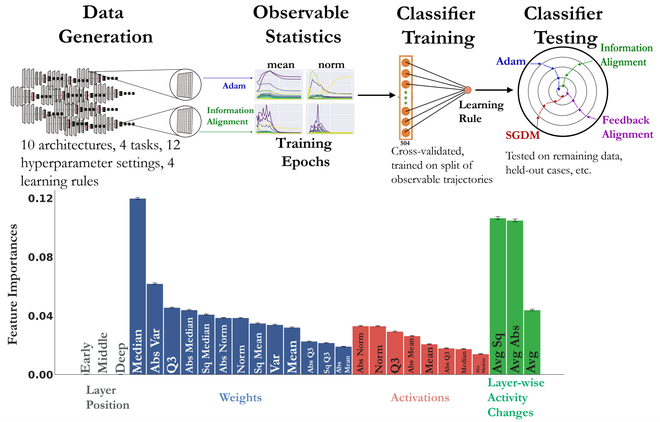
A major long-term goal of computational neuroscience will be to identify which learning rules actually drive learning in any given neural circuit. Along the route to this goal, it will be necessary to develop practically accessible experimental observables that can efficiently separate between hypothesized learning rules. So suppose you could observe a set of neurons in a circuit over the course of learning:

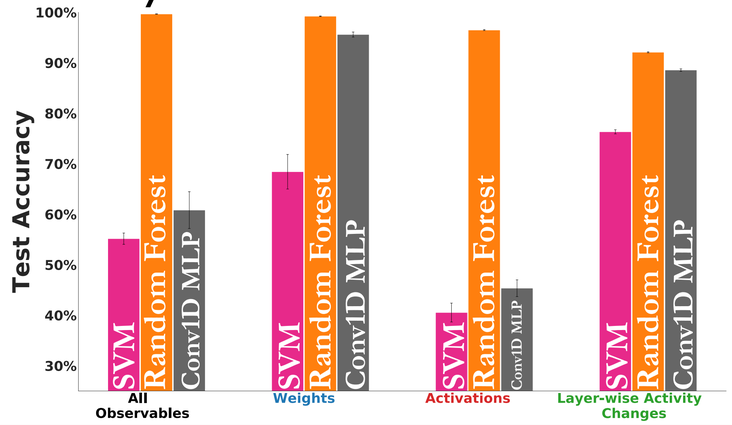
What would you measure to identify the plasticity rule operative within that circuit?

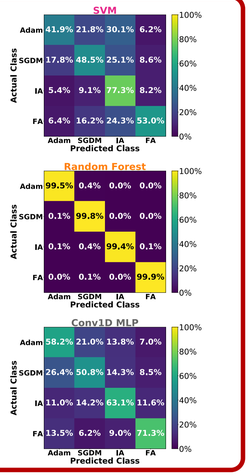


This turns out to be a substantial problem, because it is difficult on purely theoretical grounds to identify which patterns of neural changes arise from given learning rules, without also knowing the overall network architecture and loss function target (if any) of the learning system. We ask this question in ANNs, combining what we can measure in neuroscience with what we can conclude given that we know the ground truth learning rule in this setting.

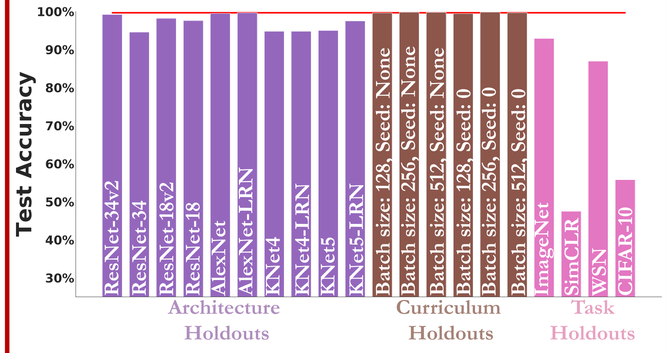
General Separability Problem *is* Tractable



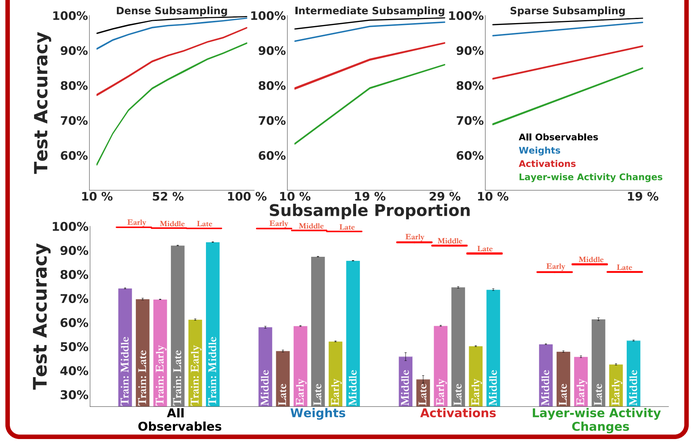




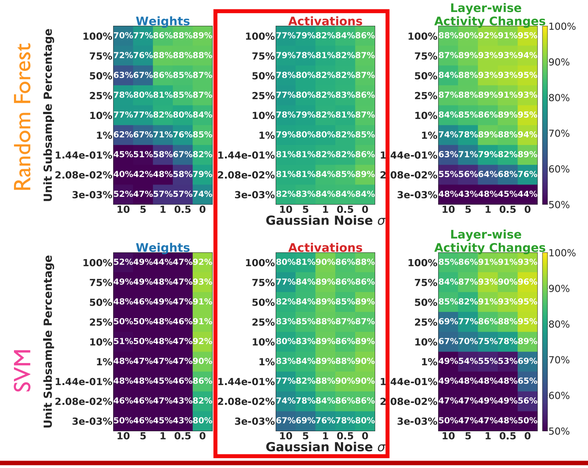
Holdouts of *Entire* Classes of Input Types (Strong Generalization to Architecture & Training Curriculum)



Sparse Subsampling *Across* Trajectory Robust to Undersampling



Activations Robust to Measurement Noise & Unit Undersampling



Conclusions

- We can identify learning rules *only* on the basis of aggregate statistics of observable measures: weights, activations, or layer-wise activity changes
- Simple (non-linear) classifier such as Random Forest generalizes across certain held-out classes of input types ("architecture" and "training curricula" holdouts)
- Measurements temporally spaced further apart are more robust to *trajectory undersampling*, for each observable measure
- Aggregate statistics across units of the network's activation patterns are most robust to *unit undersampling* and *measurement noise*
- **Hypothesis:** *in vivo* electrophysiological recordings of ***post-synaptic activities*** from a neural circuit on the order of several hundred units, frequently measured at wider intervals during the course of learning, may provide a good basis on which to identify learning rules

Experimental Collaborations Welcome!

Contact anayebi@stanford.edu if interested