A Neuronal Mechanism for Supervision of Hebbian Learning

Christian D. Swinehart and L.F. Abbott*

Abstract

In supervised learning, networks develop properties required to perform a task on the basis of an error signal. Models of supervised learning typically assume that synapse-specific error signals can be sent to individual synapses within a network, something that is biologically implausible. We assume that a supervisory signal, in the form of ordinary excitatory and inhibitory input, is sent only to the neurons of a network, modifying their response properties. This form of supervision is sufficient to guide Hebbian synaptic plasticity, allowing networks to learn tasks that neither unsupervised Hebbian plasticity nor supervised neuronal response modification can achieve individually.

Summary

Hebbian plasticity is an effective means of making neural circuits selective to patterned and correlated input in an unsupervised manner. However, many tasks require the development of particular responses on the basis of their usefulness, not on the basis of input correlations. In such situations, Hebbian synaptic plasticity faces a chicken-and-egg problem. Neural circuits must perform properly for a correlation-based rule to generate appropriate synaptic modifications, but synapses must be appropriately adjusted for circuits to perform properly in the first place. This problem must be solved by supervision.

In typical supervised learning applications, an error signal is detected and used to guide synaptic plasticity in such a way that network performance improves. This approach is effective, but it appears unrealistic when applied to biological systems because it requires that independent supervisory information be conveyed individually to each of the synapse in the network. Although massive feedback projections exist, few of them terminate in the type of three-element synapses that would be required for this type of synapse-specific supervision.

Although feedback projections that could carry supervisory information do not appear to be targeted to individual synapses, they are certainly targeted to individual neurons. We therefore propose that the supervising signal is a pattern of ordinary excitatory and inhibitory input sent to *neurons*, not to synapses. In this work, we do not consider the nature of the supervisor circuit that is the generator of these inputs.

*Volen Center and Department of Biology, Brandeis University, Waltham MA 02454

We simply assume that it monitors behavior and generates appropriate error signals. Instead, we focus on two key questions: Does providing error-related excitation and inhibition to the neurons of a network provide a sufficient signal to guide Hebbian learning in a supervised learning task? Can modification of neuronal responses by supervisory input combined with unsupervised Hebbian synaptic plasticity lead to the performance of tasks that neither one by itself could manage? The answer to both questions is yes.

We have examined response-modulation-based supervision by applying it to function approximation, a standard neural network learning task. In this task, a network is provided with different stimulus values, and it must respond by setting the firing rate of its output units to specified functions of these values. We modeled the supervisory modulation of the neurons of the network as individual changes in their firing rate versus input current relationships. The modulations took the form of either a shift, that is, a raising or lowering of the threshold, or a change in the slope of the sigmoidal neuronal response curve used for the neurons. It is well established that excitatory and inhibitory inputs can achieve both of these modulations. Synaptic strengths were modified only by an unsupervised Hebbian mechanism with a standard multiplicative constraint on the sum of the squares of the synaptic weights. It should be stressed that, by itself, this form of Hebbian learning is completely incapable of developing network connections that successful perform the function approximation task.

When we considered the case of a single output neuron learning to represent one function, response modulation alone was able to put the network in a modulated state in which it could successfully reproduce the target function. When Hebbian learning was added to this network, we found that the response-modulated neurons of the network were able to transfer the information about how to perform the task to the synapses. Ultimately, after Hebbian plasticity had equilibrated, the supervisory modulation could be removed without compromising performance. We suggest that this corresponds to the experience of initially performing a new task only with a good deal of effort and concentration (the modulatory response phase), but then, with repetition, performing it more effortless (the synaptic modification phase).

The situation becomes more interesting when we consider multiple output neurons being required to represent different functions. Here we run into what might be seen as a fundamental limitation of learning through neuronal instead of synaptic modification — there are many fewer neurons than synapses. Indeed, it is easy to find situations in which response modulation, even with ideal supervision, cannot construct a network that can perform the multiple-function-representation task. Neither can unsupervised Hebbian learning achieve this task. However, we find that a combination of supervised response modulation and unsupervised Hebbian learning results in the rapid learning of multiple output functions, even in cases where supervised response modulation alone fails dismally. This result arises because small biases imparted by the supervisor to the neurons can be amplified by the Hebbian synaptic process to yield a trained network.

We make this bias amplification argument more rigorous by analyzing how Hebbian learning proceeds when it is indirectly supervised by response modulation in

the manner we are proposing. Hebbian modification with a multiplicative constraint, such as we are considering, leads to synaptic weights that are aligned with the principal eigenvector of the correlation matrix of the presynaptic responses. In the function approximation task, this eigenvector, in the absence of response modulation, has all of its elements equal to each other, leading to a useless form of synaptic modification. Even a small amount of response modulation, introduced through the supervisory inputs, is sufficient to shift this uniform principle eigenvector to something that points in the right general direction for useful synaptic modification. Then, as the synapses are modified and network performance improves, the principle eigenvector is further tuned to produce the desired result.

In conclusion, we have shown that supervised learning tasks can be performed in a biologically plausible manner in which neurons rather than synapses receive a supervisory signal carried by conventional excitatory and inhibitory inputs.