Neuroscience Q7: PCA & Synaptic Learning Rules

The brain is a consummate learner; how is it that this mush of tissue is able to meaningfully adapt to changes in the world around it? Seminal work has discovered that one biochemical basis of learning is changes in the strength of synapses between neurons. The pre-synaptic neuron emits a spike, this triggers a chemical cascade to cross the synapse, which in turn triggers a voltage change in the postsynaptic neuron. Through changing many detailed features of the synapse, the same presynaptic spike can cause a larger or smaller effect on the postsynaptic neuron. Following a common approach, we will wrap all of this synpatic gobbledegook into the idea of synaptic weight, a single number that stands in for the strength of a connection between neurons, and changing this will be our model of learning. Let's think about a simple neuron with firing rate $y_t \in \mathbb{R}$ connected to a sensory population with firing rates $x_t \in \mathbb{R}^N$, via weight vector $w \in \mathbb{R}^N$:

$$y_t = w_t^T x_t \tag{8.2}$$

x_t represents the firing of the input neurons at time t, and w_t represents the synaptic weights. When you're born your neurons don't know how to arrange themselves (even if you're a fly, which you're not), so each of these synpatic weights needs to change to make the system as a whole do something interesting. A classic approach was suggested by Donald Hebb in the 40s: if two neurons are co-firing a lot they should be connected ("Fire together, wire together!") Let's formalise that in our model:

$$\Delta w_t = \eta x_t y_t \tag{8.3}$$

i.e. the change in a synaptic weight at time t is equal to a learning rate, η, times by the product of pre and post-synaptic firing rates. This rule says that if two neurons activate together, they should become more connected!

1. Assume that η is very small. This means that before the weights change very much you will have likely seen many different input-output pairs (x_t, y_t). Rather than computing the weight change for a particular stream of pairs, you can take the average over the distribution. Further, assume the input data are mean zero. Show that, on average, Hebbian weight updates are given by:

   $$\langle \Delta w_t \rangle = \eta C_{xx} w_t \tag{8.4}$$

   Where C_{xx} is the input data covariance matrix.

2. You know a lot about the eigenstructure of covariance matrices. Use this to argue that Hebb's rule is actually a bad rule as it will lead weights to do uninteresting things, in particular, exploding.

3. Despite this, what interesting subspace will the weight vector increasingly lie in as it follows Hebb's rule for a long time?

Okay, so Hebb's rule is out. Thankfully a chap called Oja had a think about this and came up with a better rule. He proposed:

$$\Delta w_t = \eta(x_t y_t - y_t^2 w_t) \tag{8.5}$$

You can see the first term is just Hebb's rule, but clever old Oja added a second term. This second term is going to stop the Hebbian bit from exploding, as you might already be able to guess from its form (it shrinks you along the current weight vector - good for stability!). It also obeys the key tenet of synaptic learning rules - locality! A synapse is a physical object, a mass of chemicals somewhere in your brain. It only has information about what is happening nearby, i.e. what the presynaptic neuron is doing, what the postsynaptic neuron is doing, and what happening within the synapse. You can't ask a synapse in your visual cortex to change weights according to a rule that depends on the synapses in your cerebellum. (Meditate and justify to yourself that Oja's rule is indeed local)

4. Do the averaging over datapoints again and show that the averaged learning rule can be written:

$$\Delta w_t = \eta(C_{xx}w_t - (w_t^T C_{xx} w_t)w_t) \tag{8.6}$$

5. First lets show this solves exploding weight problem. Study the behaviour of $||w_{t+1}||_2 - ||w_t||_2$. Since the learning rate is low only consider terms up to order η (i.e. because η is so small treat all the terms multiplied by η^2, η^3 etc. as 0). Show that the dynamics cause the length of the weight vector to converge to 1.

6. Study the dynamics of the components of w in the eigenbasis. If all the eigenvalues of C_{xx} are distinct show that only eigenvectors are fixed points under these dynamics. (What happens if C_{xx} has repeated eigenvalues?)

Wow, this is pretty cool, these random neurons are extracting principal components! But which principal component...?

You showed that all the eigenvectors were fixed points of the dynamics, i.e. if you start at an eigenvector you will stay there. Now what happens if you are perturbed slightly from one of these eigenvectors? Does the dynamics push you back to where you started (called a stable fixed point) or does it push you away (called unstable). The best analogy to think of here is something like a freely swinging pendulum. This has two fixed points: the pendulum is pointed straight up, or straight down. In either situation the forces balance, and it is a fixed point of the dynamics. But, if you perturb slightly from these two fixed points, very different things occur. For obvious reasons, we only really care about the fixed stable points, the system will never settle on the unstable ones. Let's find the stable fixed points of Oja's rule.

7. Assume the weight vector is sitting on an eigenvector, and is perturbed slightly by some noise vector of small magnitude. Decompose the noise vector in the eigenbasis and study the dynamics of the weight vector to first order in the noise. Show that the fixed point is only stable if the weight vector is sitting on the largest eigenvector.

Look at that, that neuron is extracting the first principal component of the data!! And from such a simple local rule!

Some interesting extensions to this are networks of neurons that extract many principal components, or how to choose η based on the rate of change of the distribution of data. Both of these were the topics of theoretical neuroscience papers from the last 20 years. There are learning rules that make the network do all sorts of things. That said, our understanding is very nascent, and all the best neural network models of the brain do not use such unsupervised learning rules.