Hebbian Learning

- Hebbian learning is a classic theory of neuroscience. The basic idea is that *cells that fire together wire together*. This means that the synaptic strength between two neurons tends to get strengthened whenever the two neurons both fire, There is experimental evidence that variants of this algorithm occur in Aplysia (Kandel).
- As we will see in later lectures, most neural network learning algorithms involve some variant of Hebbs rule. Indeed there have been recent attempts to make neural networks learn in a local manner which emphasizes Hebbs rule. This applies to a large range of learning algorithms including deep neural networks, clustering algorithms like k-means, the Boltzmann machine and many others. This applies both to supervised and unsupervised learning algorithms. This is perhaps not surprising and, to some extent follows from the fact that most learning algorithms are minimizing a loss function by stochastic gradient descent (to be discussed later).
- Hebbian learning also arises in more symbolic learning and memorization algorithms (L. Valient Circuits of the Mind) where one goal is to store memories by binding concepts together.
- This lecture will describe a simple application of Hebbs rule to learn the receptive fields of neurons in an unsupervised manner.

Unsupervised learning of the receptive fields.

- We now introduce unsupervised neural network algorithms for learning receptive fields. This section is based on computational studies performed in the 1980's (Linsker, 1986a,b; Yuille et al., 1989), see (Zhaoping, 2014) for other references. These studies are based on modifications of the Hebb learning rule, which has some experimental support. Exercise demo (12.3.1) illustrates principal component analysis and Oja's rule (Oja, 1982).
- The basic findings are that center-surround, orientation selective, quadrature pairs, and disparity sensitive cells (precursors to cells that can estimate depth from binocular stereo) could all be obtained by variants of the same learning rule. Analysis of these findings suggest that this is partly due to the shift invariance of images.

Unsupervised learning by Hebb's rule (I)

- We first describe a simple unsupervised learning model for a single cell (Oja, 1982). The output $S(t)$ of the cell is a function of time t and is a weighted sum of the inputs $I_i(t)$, where the weights $\omega_i(t)$ are functions of time and are updated by Oja's rule (Oja, 1982):

$$S(t) = \sum_j w_j(t) I_j(t),$$

$$\frac{dw_i(t)}{dt} = S(t)\{I_i(t) - S(t)w_i(t)\}. \quad (7)$$

- The first term (Hebbs) increases the strength of a weight w_i if its input $I_i(t)$ is positively correlated with the output $S(t)$ (i.e., $< S(t)I_i(t) >> 0$), while the second term decreases the value of all weights by an amount proportional to their strength.
- This can be expressed as a single update equation:

$$\frac{dw_i(t)}{dt} = \sum_j w_j I_i(t) I_j(t) - \sum_{jk} w_i w_j w_k I_j(t) I_k(t). \quad (8)$$

Unsupervised learning by Hebb's rule: Analysis (I)

- Next we assume that the weights w_i change at a slower rate than the input images. This enables us to replace the terms $I_i(t)I_j(t)$ with their expectation $K_{ij} =< I_i(t)I_j(t) >$, which is the correlation function of the input. This gives:

$$\frac{dw_i(t)}{dt} = \sum_j w_j K_{ij} - \sum_{jk} w_i w_j w_k K_{jk}. \tag{9}$$

- The fixed points of this equation, the values of w such that $\frac{dw_i(t)}{dt} = 0$, can be shown to be eigenvectors of the correlation function K_{ij}. A slight modification gives an update rule (Yuille et al., 1989) that converges to the global minimum of the cost function:

$$E(\vec{w}) = -(1/2)\sum_{i,j} K_{ij} w_i w_j + (k/4)(\sum_i w_i^2)^2$$

Unsupervised learning by Hebb's rule: Analysis (II)

- The global minimum corresponds to the biggest eigenvalue of K_{ij}. If the correlation function K_{ij} decreases with distance, then the biggest eigenvalue is at frequency 0, so the cell is not tuned to any frequency. But if the correlation function has the shape of a Mexican hat, then the biggest eigenvalue has a nonzero frequency, which implies that the cell is orientated (Yuille et al., 1989).
- The correlation function of natural images does decrease spatially, but Linsker (1986a,b) showed that correlation functions similar to the Mexican hat arise if this learning procedure is applied to a sequence of layers.
- This analysis yields receptive fields that are sinusoids, and hence have no spatial fall-off, which is unrealistic. But receptive fields of neurons are limited by the geometrical positions of the dendrites. If these constraints are included, then the algorithms converge to receptive fields that are similar to Gabor functions.

Unsupervised learning by Hebb's rule: Conclusion

- It is interesting that Hebbian Learning can learn receptive field properties which are somewhat similar to those found experimentally. The approach can be extended to learn filters for binocular stereo algorithms (see previous lectures). Zhaoping Li's book gives a detailed discussion of these algorithms. It is also possible for unsupervised algorithms to learn some of the more salient structures of V1 such as the ocular dominance patterns, the spatial layout of receptive fields (filterbanks), and even the cytochrome-oxidase blobs that carry color channels.
- These studies are very impressive, but perhaps not fully satisfying. It is, as yet, impossible to test that the visual system evolves using these unsupervised learning algorithms. Moreover, many of the findings are post-hoc (after the fact and it would be better, as in physics, to predict a phenomenon before you observe it.
- Certain findings, like Gabor filters, may arise chiefly because images are shift-invariant. As a cynic once said "it would be more interesting if you found a learning algorithm that didn't predict Gabor filters".