18 Synaptic weights from Receptive Fields and Stimulus Reconstruction

18.1 General description of receptive fields

We consider a phenomenological description of the stimulus that causes a neuron to fire. Our description will be general, although as a matter of practice it is simplest to think in terms of visual objects, i.e., a pattern of illumination that evolves over time and space. The receptive field forms a kernel, such that the spike rate of the cell is the temporal convolution of the stimulus with the receptive field and the spatial overlap of the stimulus with the receptive field.

We define the inhomogeneous spike rate as $r(t)$. This is the rate that goes into, per se, a Poisson rate expression where the probability of no spikes in the interval $[0,t]$ and one spike in the interval $(t,t+dt]$ is $P(t) = r(t) \cdot \exp \left(- \int_{0}^{t} dt' r(t') \right)$. Then

$$r(t) = f \left[I_o + \int_{-\infty}^{\infty} d\vec{x} \int_{-\infty}^{t} dt'I(\vec{x},t')R(\vec{x},t-t') \right]$$ (18.1)

where $f[\cdot]$ is the nonlinear input-output relation, $I(\vec{r},t)$ is the stimulus or input, $R(\vec{x},t)$ is the receptive field with $\vec{x}$ the two-dimensional spatial vector, and $I_o$ is the baseline input.

When the stimuli driven part of the input is small compared to $I_o$, we can expand $g[\cdot]$ in a Taylor series and write

$$r(t) \simeq r_o + f' \int_{-\infty}^{\infty} d\vec{x} \int_{-\infty}^{t} dt'I(\vec{x},t')R(\vec{x},t-t')$$ (18.2)

where $r_o = f[I_o]$ and $f' = \frac{df}{dI} \bigg|_{I=I_o}$ (18.3)

so that the firing rate is a linear function of the stimulus. This allows us to focus on the receptive field without worrying about the nonlinearity $f[\cdot]$. A review by Chichilnisky (2001) shows an example for the case where the response is separable, i.e., where it can be written as the product of a spatial pattern times a temporal waveform, e.g., $R(\vec{r},t) = u(\vec{r})v(t)$; this review also addresses the assignment of both
\( R(\vec{x},t) \) and \( f[\cdot] \) when the stimulus driven part of the input is not small compared to \( I_o \).

To gain some insight into the general response properties of neurons, we recall that a matrix can always be expanded in terms of its eigenvectors by a singular valued decomposition. In terms of the notion for the receptive field, we have

\[
R(\vec{x},t) \equiv \sum_{n=1}^{\text{rank}(R)} \lambda_n u_n(\vec{x}) v_n(t)
\]

(18.4)

where the functions \( u_n(\vec{x}) \) form an orthonormal basis set in space and \( v_n(t) \) for an orthonormal basis set in time. The eigenvalues for these basis sets are given by \( \lambda_n^2 \) and, of course, are ordered so that \( \lambda_1 > \lambda_2 > \lambda_3 \cdots \). In this case the receptive field is not separable, as first discussed by the work of McClean and Palmer (1989) and analyzed in some detail in the work of Golomb, Kleinfeld, Reid, Shapley and Shraiman (1994).

Then

\[
r(t) \simeq r_o + f' \sum_{n=1}^{\text{rank}(R)} \lambda_n \int_{-\infty}^{\infty} d^2 \vec{x} \ u_n(\vec{x}) \int_{-\infty}^{t} dt' S(\vec{x},t') \ v_n(t-t').
\]

(18.5)

Now suppose that the stimulus is separable, as is often the case in primary sensory areas. For example, in vision our eyes shift from position to position about five times a second. In this case we may write

\[
I(\vec{x},t) \equiv X(\vec{x})T(t).
\]

(18.6)

The spatial part of the stimulus that each mode “sees” is given by the overlap integral of the spatial pattern of the stimulus with the spatial pattern of each mode,
Figure 3: Focal attention on faces causes the visual gaze to be maintained at key locations for 100 ms. From Yarbus

\begin{equation}
U_n = \int_{-\infty}^{\infty} d^2 \vec{x} X(\vec{x}) \ u_n(\vec{x}).
\end{equation}

where the $U_n$ are scalars. In this case the $u_n(\vec{x})$ act as the weights and the $U_n$ are the output of say a dendritic branch as opposed to the entire cell.

The time dependence of the stimulus is convoluted with each of the associated temporal modes to form the temporal evolution for that mode, \textit{i.e.},

\begin{equation}
V_n(t) = \int_{-\infty}^{t} dt' \ T(t') \ v_n(t - t').
\end{equation}

where the $V_n(t)$ are functions. We thus find

\begin{equation}
r(t) = r_o + f' \ \sum_{n=1}^{\text{rank}(R)} \lambda_n \ U_n \ V_n(t).
\end{equation}

so that each temporal waveform is weighted by the expansion coefficient for the receptive field and the spatial overlap of the mode with the stimulus. The point is that the temporal response of the neuron, given by $r(t)$, depends on the spatial
pattern of the input as well as the temporal evolution of the stimulus. This is what one calls, or some call, a "temporal code", i.e., the coding of different stimuli, even quasi-static stimuli, by different temporal patterns of spike rates. The inhomogeneous rate $r(t)$ may evolve in time as fast as the response of the sensory cells, such as retinal ganglion cells for the case of vision.

A final point is that the summation over modes typically contains only a few terms, not the full rank of the matrix $R$. The spatial coefficient $U_n$ has a signal-to-noise ratio that varies in proportion to $\lambda_n$ for the $n^{th}$ mode. Thus the above series is cut off after two or three terms as the signal dives below the noise. The SVD expansion can be used as a data compression scheme in the description of the receptive field. For magnocellular cells,

$$r(t) \approx r_o + [f'\lambda_1 U_1] V_1(t) + [f'\lambda_2 U_2] V_2(t).$$  

(18.10)
18.2 Digression on singular values decomposition

In the expansion

\[ R(\vec{x}, t) \equiv \sum_{n=1}^{\text{rank}(R)} \lambda_n u_n(\vec{x}) v_n(t) \]  \hspace{1cm} (18.11)

the functions satisfy the orthonormality constraints

\[ \int_{-\infty}^{\infty} d^2 \vec{x} \ u_n(\vec{x}) u_m(\vec{x}) = \delta_{nm} \]  \hspace{1cm} (18.12)

and

\[ \int_{-\infty}^{\infty} dt' \ v_n(t') v_m(t') = \delta_{nm}. \]  \hspace{1cm} (18.13)
We now consider the contraction of the receptive field matrices to form a symmetric correlation matrix, \( i.e., \)

\[
C(t, t') = \int_{-\infty}^{\infty} d^2 \vec{x} R(\vec{x}, t) R(\vec{x}, t')
\]

\[
\equiv \sum_{n=1}^{\text{rank}(R)} \sum_{m=1}^{\text{rank}(R)} \lambda_n \lambda_m \int_{-\infty}^{\infty} d^2 \vec{x} \ u_n(\vec{x}) u_m(\vec{x}) \ v_n(t) v_m(t')
\]

\[
= \sum_{n=1}^{\text{rank}(R)} \sum_{m=1}^{\text{rank}(R)} \lambda_n \lambda_m \delta_{nm} \ v_n(t) v_m(t')
\]

Then \( v_n(t) \) solves the eigenvalue equation

\[
\int_{-\infty}^{\infty} dt' C(t, t') v_n(t') = \sum_{m=1}^{\text{rank}(R)} \lambda_n^2 v_m(t) \int_{-\infty}^{\infty} dt' v_n(t') v_m(t') = \lambda_n^2 v_n(t)
\]

and the \( u_n(\vec{x}) \) are found from

\[
\int_{-\infty}^{\infty} dt' R(\vec{x}, t') v_n(t') = \sum_{m=1}^{\text{rank}(R)} u_m(\vec{x}) \int_{-\infty}^{\infty} dt' v_n(t') v_m(t') = u_n(\vec{x}).
\]

### 18.3 Learning connections

We consider the optimal assignments of weights to the coding of an input. Optimal here will means that we can then decode the input, i.e., predict the input, from the firing properties of the neurons. This does NOT mean that we can perfectly reconstruct or predict the form of the input from neuronal response; perception is only as good as the receptors.

We take the inputs as \( I_i(t) \). This could correspond to the value of the output of a photoreceptor or, in a synthetic world, to the pixel of a camera at a time \( t \). We take the output of a neuron as \( V_i(t) \), where \( V \) is a positive number. Then

\[
V_i(t) = \sum_k W_{ik} I_k(t)
\]
where $W_{ik}$ connects input $I_k$ to the $i$-th neuron. This just corresponds to a linear, one level network, more properly a Perceptron. We are going to keep the clunky form of double indices and sums to maintain clarity, much as one could write $V_i(t)$ as the inner product $V_i(t) = W_i^T I(t)$.

How do we choose the $W_{ik}$s? We can choose them to be the $V_i(t)$s the best predictor of the input. Then we look at the reconstruction algorithm, given by

$$I_k(t) = \sum_i W_{ik} V_i(t)$$

(18.18)

Let’s form a quadratic error function and minimize the argument that minimizes the error. we are going to do this in two ways. First, to get the optimal output and see if we recover our original input scheme. This will act as a form of self consistency. Second, to get the optimal form of the connection strengths, i.e., the get a rule for the online learning of new inputs.

First, let’s find the output that minimizes the error at each time point, and for each reconstructed input, where

$$\text{Error}(i,k,t) = \| I_k(t) - W_{ik}(t-1)V_i(T) \|^2$$

(18.19)

$$= \left[ \| I_k(t) \|^2 - 2I_k(t)W_{ik}(t-1)V_i(t) + \| W_{ik}(t-1)V_i(T) \|^2 \right]$$

$$= \left[ \| I_k(t) \|^2 - 2X_k(t)W_{ik}(t-1)V_i(t) + \| W_{ik}(t-1) \|^2 V_i^2(t) \right].$$

where the time variable for the $W_{ik}$ refers to the extent of updating.

Let’s find the argument $V_i$ that minimizes the error, i.e., $\arg\min$. Then

$$V_i(t) = \arg\min_{V_i} \left[ \sum_k \text{Error}(i,k,t) \right]$$

(18.20)

$$= \arg\min_{V_i} \sum_k \left[ -2I_k(t)W_{ik}(t-1)V_i(t) + \| W_{ik}(t-1) \|^2 V_i^2(t) \right]$$

$$= \arg\min_{V_i} \sum_k \left| \sum_k W_{ik}(t-1)I_k(t) - V_i(t) \right| \left| \sum_k W_{ik}(t-1) \right|^2$$

$$= \frac{\sum_k W_{ik}(t-1)I_k(t)}{\| \sum_k W_{ik}(t-1) \|^2}.$$

This is the same result we had previously to within a normalization. The output of each cell is the inner product of the input with the weights. So we are self-consistent.

Let’s find a rule for the argument $W_{ik}(t)$ that minimizes the error, i.e., $\arg\min_{W_{i,k}}$. 


over all time. Then

\[
W_{ik}(T) = \arg \min_{W_{ik}} \left[ \sum_{t=1}^{T} \text{Error}(i, k; t) \right]
\]  

\( (18.21) \)

\[
= \arg \min_{W_{ik}} \sum_{t=1}^{T} \left[ -2I_k(t)W_{ik}(t-1)V_i(t) + \| W_{ik}(t-1) \|^2 V_i^2(t) \right]
\]

\[
= \arg \min_{W_{ik}} \sum_{t=1}^{T} \left\| \frac{I_k(t)V_i(t)}{V_i(t)} - W_{ik}(t-1) \right\|^2 V_i^2(t)
\]

\[
= \frac{\sum_{t=1}^{T} I_k(t)V_i(t)}{\sum_{t=1}^{T} V_i^2(t)}.
\]

so we see that the weights are the cross-correlation of the input with the output. To get an incremental rule, we note

\[
W_{ik}(T - 1) = \frac{\sum_{t=1}^{T-1} I_k(t)V_i(t)}{\sum_{t=1}^{T-1} V_i^2(t)}
\]

\( (18.22) \)

\[
= \frac{\sum_{t=1}^{T} I_k(t)V_i(t) - I_k(T)V_i(T)}{\sum_{t=1}^{T} V_i^2(t) - V_i^2(T)}.
\]

so

\[
\sum_{t=1}^{T} I_k(t)V_i(t) = W_{ik}(T - 1) \sum_{t=1}^{T} V_i^2(t) - W_{ik}(T - 1)V_i^2(T) + I_k(T)V_i(T)
\]

\( (18.23) \)

Combining terms, we get our learning rule

\[
\Delta W_{ik}(T) \equiv W_{ik}(T) - W_{ik}(T - 1)
\]

\( (18.24) \)

\[
= \frac{V_i(T) [I_k(T) - W_{ik}(T - 1)V_i(T)]}{\sum_{t=1}^{T} V_i^2(t)}.
\]

The change in weight has a contribution that appears like a Hebb rule, except that the change decrements sustained plasticity. The rule depends only of pre- and post-synaptic activity and the previous value of the weight; all of this is biologically plausible. The normalization by the square of the output is worry some; presumably one must add a feature that cuts this term off after some long time so that the denominator is simply a term that depends of average (square) activity.

The learning rule is the starting point for work by Oja (1982) showing that such a rule leads to weights that approximate the first principle component of the input. Let’s plug in for \( V_i(T) \),

\[
\Delta W_{ik}(T) = \frac{\sum_{m} W_{im}(T-1)C_{mk}}{\| \sum_{k} W_{ik}(T-1) \|^{2}} - \frac{\sum_{m,n} W_{im}(T-1)C_{mn}W_{in}(T-1)}{\| \sum_{k} W_{ik}(T-1) \|^{2}} W_{ik}(T - 1)
\]

\( (18.25) \)

\[
= \frac{\| \sum_{k} W_{ik}(T-1) \|^{2} \sum_{m} W_{im}(T-1)C_{mk} - \sum_{m,n} W_{im}(T-1)C_{mn}W_{in}(T-1)W_{ik}(T - 1)}{\sum_{m,n} \sum_{t=1}^{T} W_{im}(T-1)C_{mn}W_{in}(T-1)}
\]

8
where

$$C_{mn} = \frac{1}{T} \sum_{t=1}^{T} I_m(t) I_n(t). \quad (18.26)$$

is the correlation matrix of the inputs and may be assumed to achieve a set of roughly constant values. Switching back to matrix notation

$$\frac{dW(t)}{dt} = \|W(t)\|^2 CW(t) - \left[ W^T(t)CW(t) \right] W(t) \quad (18.27)$$

we see that the expression is in the form of the Ojas (1982) rule for which the $W$ will select the dominant eigenvector of the correlation matrix or equivalently the first principle component of the correlation matrix. In steady state, $dW(t)/dt = 0$, which leads to the eigenvalue equation

$$CW = \frac{W^T CW}{\|W\|^2} W \quad (18.28)$$

where $W = W(t \to \infty)$. The weights vector $W$ will be dominated by the leading eigenvector of the correlation matrix of the inputs, $C$. The associated eigenvalue is just $W^T CW / \|W\|^2$; this can be readily checked using the same approach that we used to show that a linear recurrent system can only store one memory, i.e., the dominant eigenvector.

Rather impressively, Chklovskii recently showed that this approach applied to an input space of odorants yields a matrix that matches the response of selected olfactory receptor neurons cells.

**Figure 9:** Observed versus calculated synaptic weights for responses in the fly. From Chklovskii