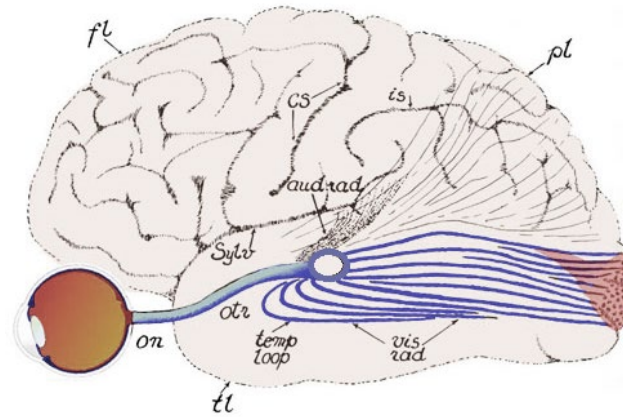
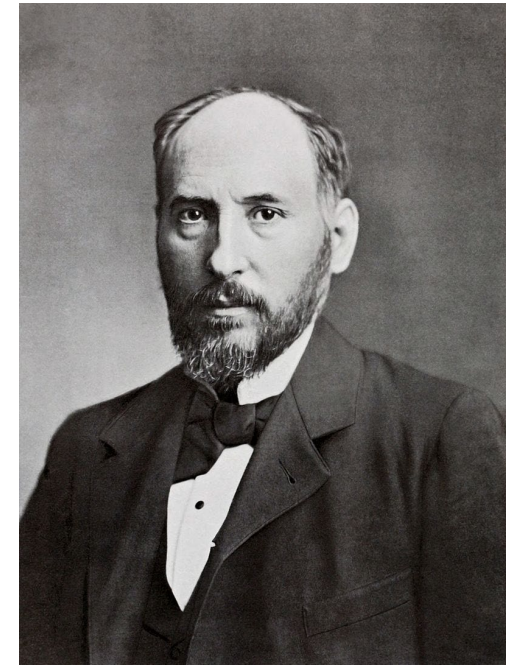
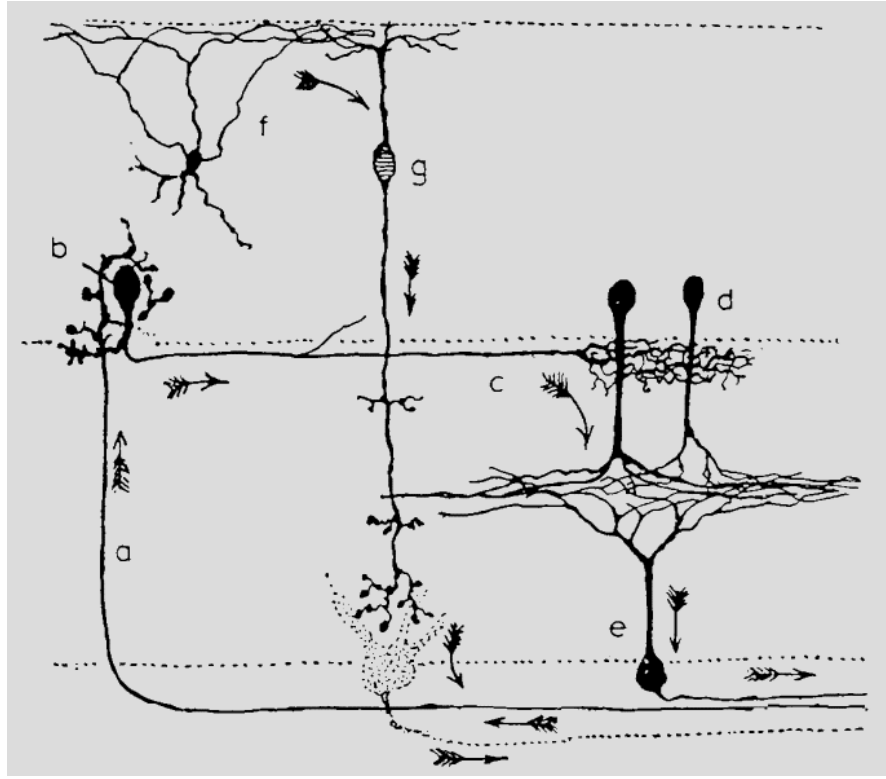
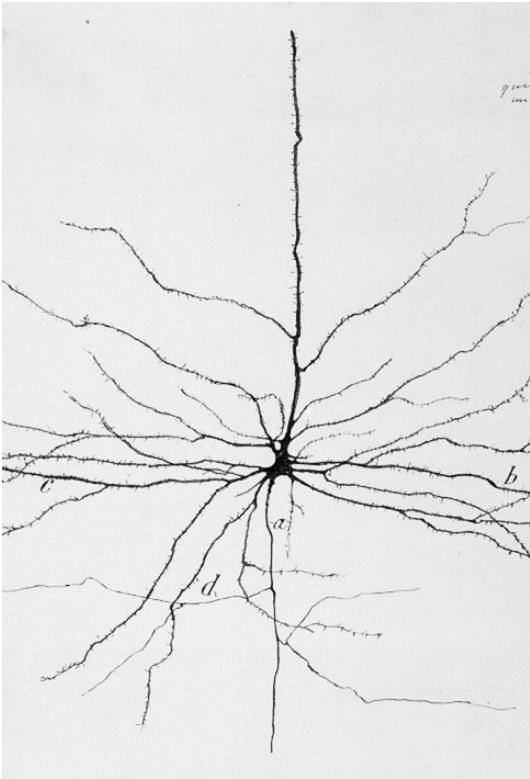


Some biologically-based theoretical considerations of the visual system



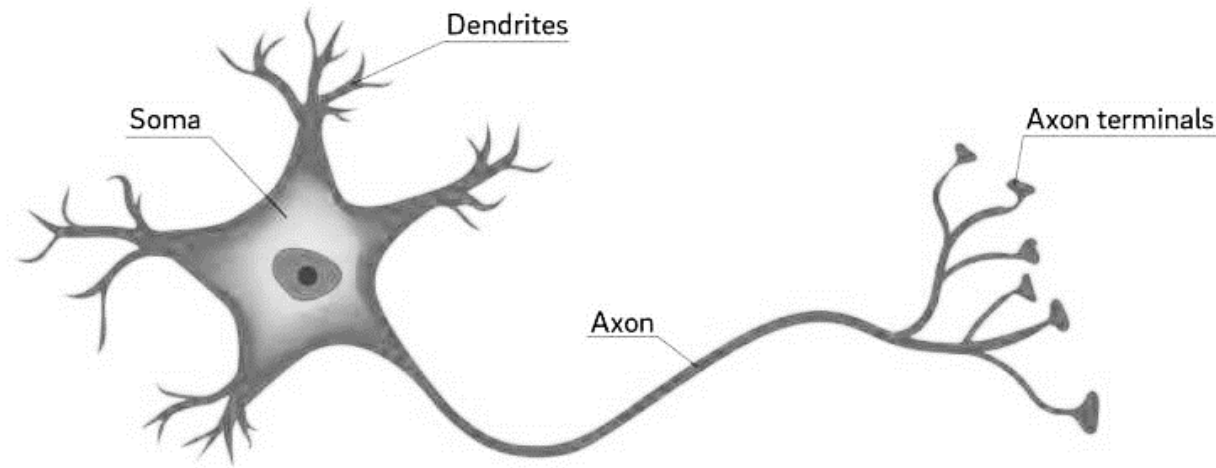
Ko Ho, CUHK

The neuron doctrine by Ramón Y Cajal



Nobel Prize in Physiology or Medicine 1906

The neuron doctrine by Ramón Y Cajal

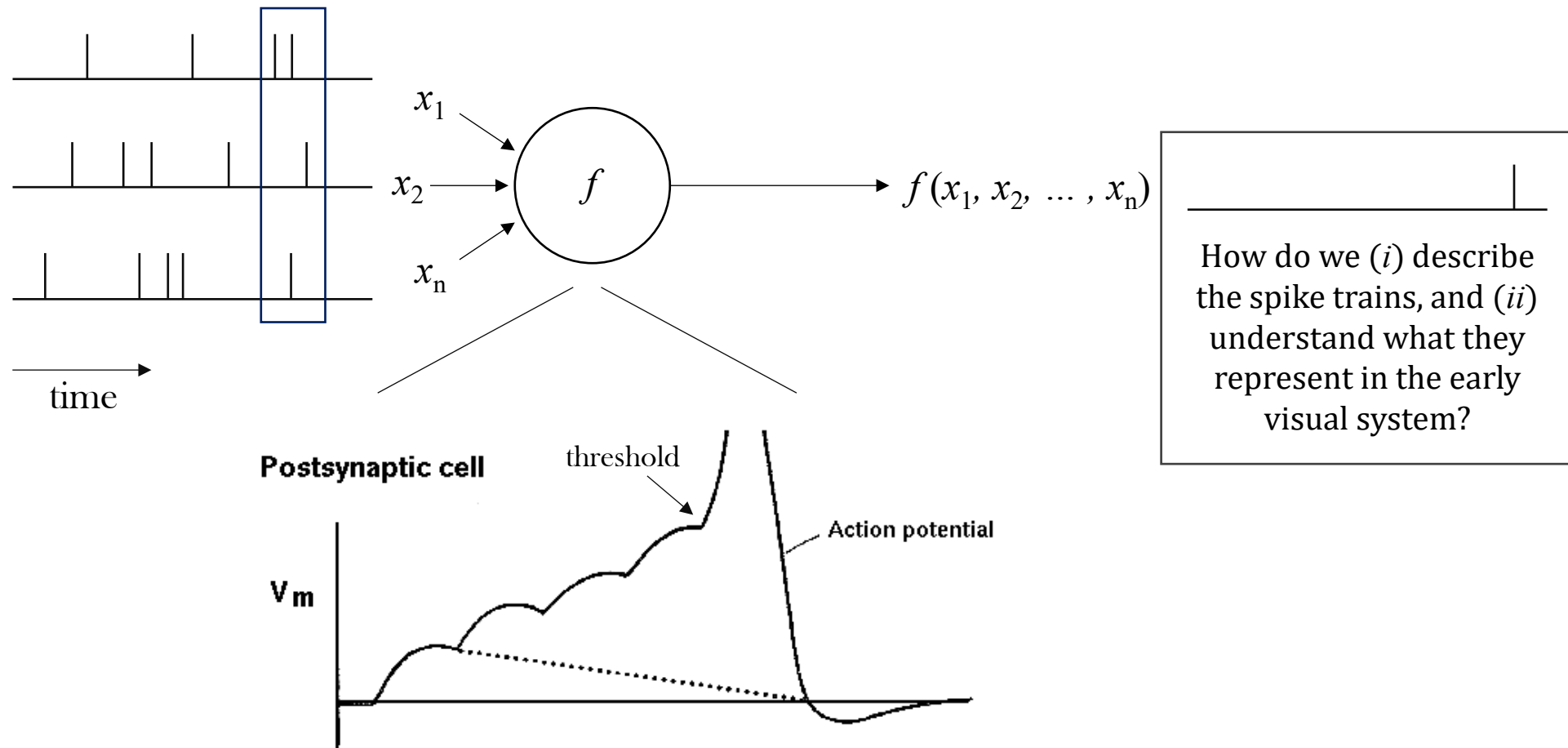


Neurons are the structural and functional units of the nervous system. Neurons are individual cells, not anatomically continuous to other neurons.

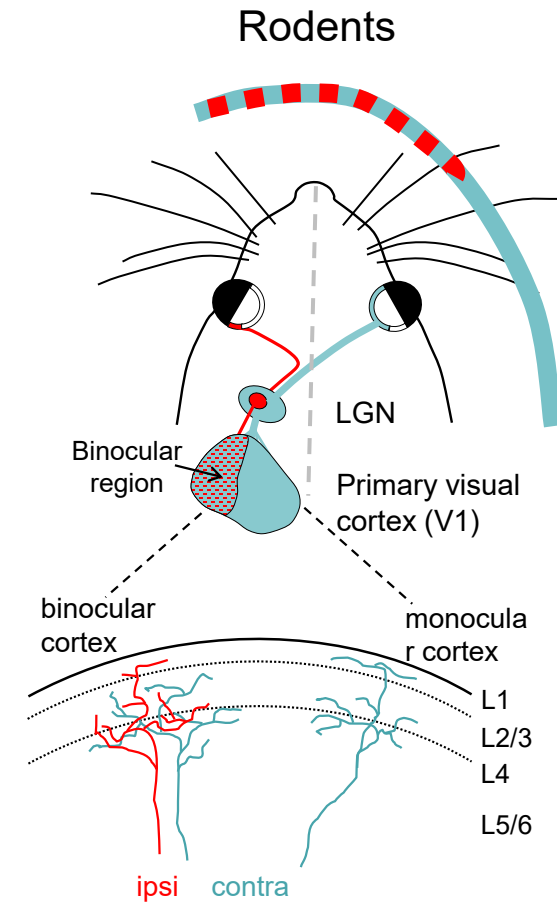
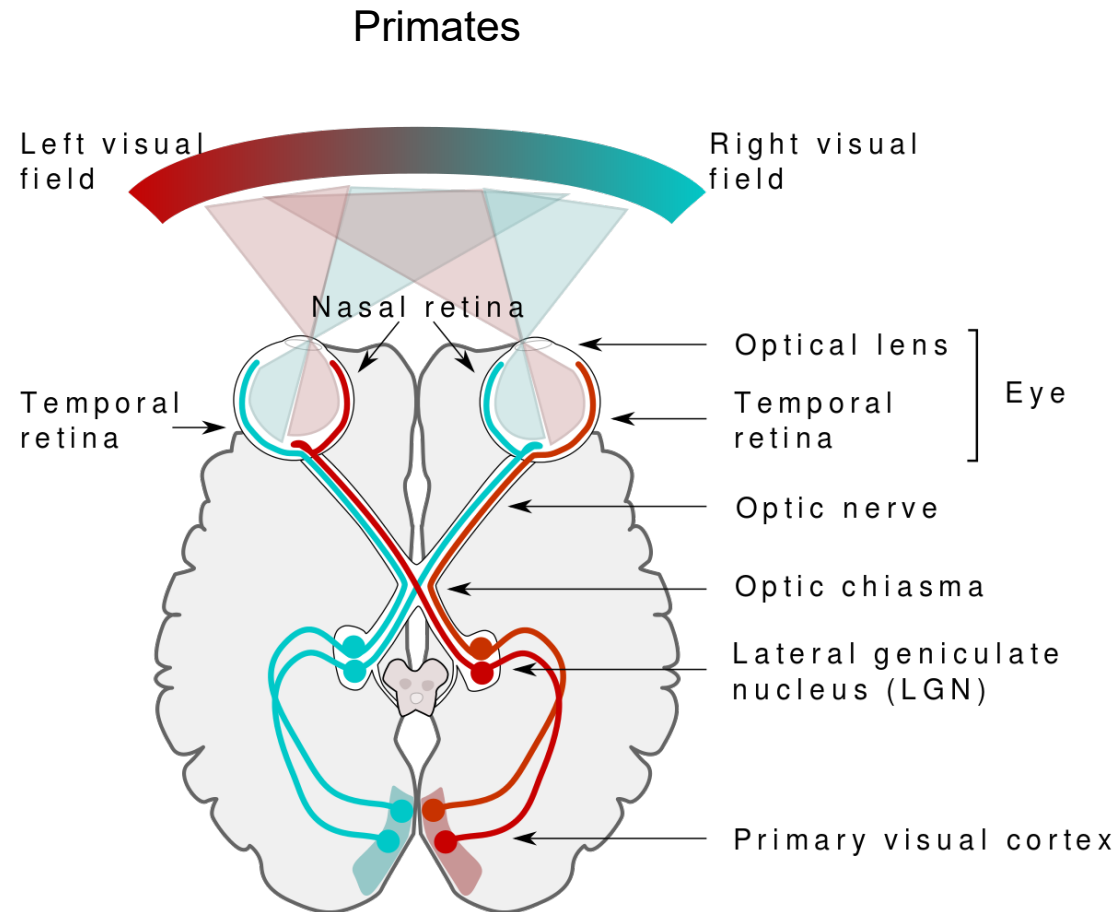
A neuron has three parts: dendrites, soma (cell body) and axon. The axon branches and makes close contact to dendrites or the soma of other neurons.

Direction of nerve impulse is due to the pre-established relations between the neurons and the initial position of the excitation; if the point of entry of the current varies, the excitation wave can go from the axon to the cell body or from an axonal branch to its main trunk; and the same can apply to dendrites.

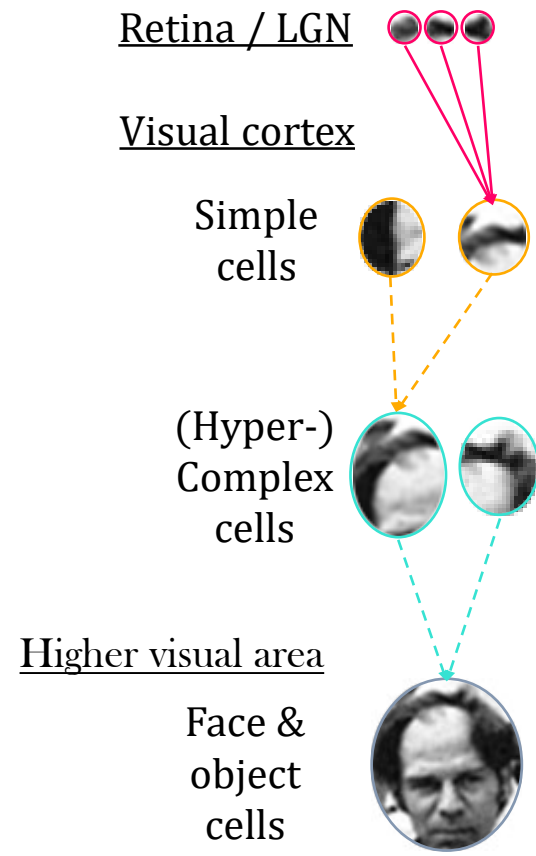
Neurons are inter-connected threshold logic units



General organization of the visual system: anatomy



General organization of the visual system: hierarchy of visual processing



Hubel & Wiesel
Nobel Prize 1981

Some notes before we carry on

Neurons spike stochastically, hence we often consider their responses in a **probabilistic** manner.

Spikes have non-zero durations, yet we assume that's negligible, and we are principally concerned with the **timing and rate** of their occurrences.

For a given neuron, we assume that there is an underlying **stimulus-dependent rate function** that governs spike occurrences.

Assume that we give an animal a time-varying sensory stimulus, we define the neural response rate function $r(t)$ of a given neuron under observation as one that has the following property:

The expected spike count over a time interval $\langle n_{[t_0, t_0 + \Delta t]} \rangle$ is given by

$$\langle n_{[t_0, t_0 + \Delta t]} \rangle = \int_{t_0}^{t_0 + \Delta t} r(t) dt.$$

For small Δt ,

$$\langle n_{[t_0, t_0 + \Delta t]} \rangle \approx r(t_0) \Delta t.$$

When Δt is small enough such that within the interval at most one spike can occur (which is also biologically realistic), $\langle n \rangle$ gives the probability of spiking, i.e.

$$P(1 \text{ spike during the interval } [t_0, t_0 + \Delta t]) = r(t_0) \Delta t.$$

We can then have a model of spike generation, by assuming:

1. Spike occurrences are independent in non-overlapping time intervals.
2. The underlying process of spike generation, for each sufficiently small time interval $[t_0, t_0 + \Delta t]$, is governed by the following rule with an underlying rate function

$$P(1 \text{ spike during the interval } [t_0, t_0 + \Delta t]) = r(t_0)\Delta t.$$

This is termed a Poisson model of spike generation, due to the following result:

For a reasonably smooth $r(t)$ and a certain (not-too-large) interval $[t_0, t_0 + \Delta t]$,

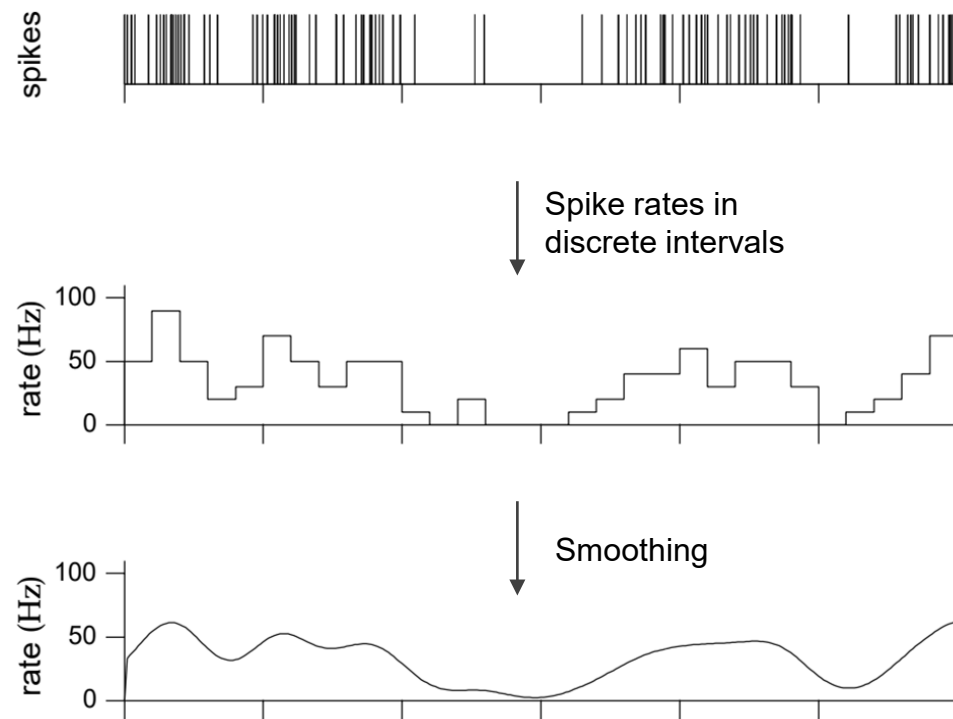
$$P(k \text{ spikes in } [t_0, t_0 + \Delta t]) \approx \frac{(r(t_0)\Delta t)^k}{k!} e^{-r(t_0)\Delta t} \sim \text{Pois}(k, r(t_0)\Delta t).$$

A Poisson model of spike generation also has the following properties:

1. For a given time interval, the Fano factor (ratio of variance to mean) of spike rate equals one, this is a reasonable approximation in many cases, but not for all.
2. The inter-spike time interval (ISI) distribution is predicted to be exponential, which is biologically never true due to the presence of refractory period immediately following each spike. This can be modelled better by a renewal process.

Nonetheless, the model provides fairly good approximations in many situations, allowing us to simulate neural responses well.

A typical experimental example of obtaining $r(t)$:



Our next question: how does one model the stimulus-dependence of the neural response rate function $r(t)$?

We first consider a linear filter model for the estimated rate function $r_{est}(t)$,

$$r_{est}(t) = r_0 + \int_0^\infty s(t - \tau)D(\tau)d\tau,$$

where r_0 is the baseline firing rate, $s(t)$ represents a time-varying stimulus, and $D(\tau)$ is a filter.

With a given model's output $r_{est}(t)$ and an experimentally measured $r(t)$, we attempt to minimize the difference between $r_{est}(t)$ and $r(t)$ wrt $D(\tau)$. For this purpose, a metric to use would be the mean squared error over the recording time interval $[0, T]$, given by

$$E = \frac{1}{T} \int_0^T [r_{est}(t) - r(t)]^2 dt.$$

(cont')

$$\begin{aligned} E &= \frac{1}{T} \int_0^T [r_{est}(t) - r(t)]^2 dt \\ &= \frac{1}{T} \int_0^T \left[r_0 + \int_0^\infty s(t - \tau) D(\tau) d\tau - r(t) \right]^2 dt \end{aligned}$$

Discretizing E we get

$$\begin{aligned} E &= \frac{\Delta t}{T} \sum_{i=0}^{T/\Delta t} \left[r_0 + \sum_{k=0}^{\infty} s_{i-k} D_k \Delta t - r_i \right]^2 \\ \frac{\partial E}{\partial D_j} &= \frac{2\Delta t}{T} \sum_{i=0}^{T/\Delta t} \left[r_0 + \sum_{k=0}^{\infty} s_{i-k} D_k \Delta t - r_i \right] s_{i-j} \Delta t \end{aligned}$$

Setting $\frac{\partial E}{\partial D_j} = 0$,

$$\begin{aligned} \sum_{i=0}^{T/\Delta t} (r_i - r_0) s_{i-j} \Delta t &= \sum_{i=0}^{T/\Delta t} \sum_{k=0}^{\infty} s_{i-k} D_k s_{i-j} (\Delta t)^2 \\ &= \sum_{k=0}^{\infty} D_k \sum_{i=0}^{T/\Delta t} s_{i-k} s_{i-j} (\Delta t)^2 \end{aligned}$$

Taking $\Delta t \rightarrow 0$ again, and supposed $s(t)$ is white noise stimulus,

$$\begin{aligned} \int_0^T (r(t) - r_0) s(t - \tau) dt &= \int_0^{\infty} D(\tau') \int_0^T s(t - \tau') s(t - \tau) dt d\tau' \\ \frac{1}{T} \int_0^T r(t) s(t - \tau) dt &= \int_0^{\infty} D(\tau') \frac{1}{T} \int_0^T s(t - \tau') s(t - \tau) dt d\tau' \end{aligned}$$

Denoting Q the correlation function, we can write

$$Q_{rs}(-\tau) = \int_0^\infty D(\tau') Q_{ss}(\tau' - \tau) d\tau'.$$

Since $s(t)$ is white noise stimulus, $Q_{ss} = \delta(t)\sigma^2$, and

$$\begin{aligned} Q_{rs}(-\tau) &= \int_0^\infty D(\tau') \delta(\tau) \sigma^2 d\tau' \\ &= \sigma^2 D(\tau) \\ D(\tau) &= \frac{Q_{rs}(-\tau)}{\sigma^2}. \end{aligned}$$

Which means that the optimal filter $D(\tau)$ is given by the reverse correlation of the rate function and a white noise stimulus (scaled by dividing by the variance of stimulus).

How do we apply this result then? Intuitively thinking, the reverse correlation function weighs the stimulus by the magnitude of response.

Instead of computing correlation between neural response rate and stimulus, can we simply average the stimulus over a time window preceeding spikes?

As it turns out, we can, since the spike-triggered average stimulus s_{avg} is computed as

$$\begin{aligned} s_{avg}(\tau) &= \left\langle \frac{1}{k} \sum_{i=1}^k s(t_i - \tau) \right\rangle \\ &= \left\langle \frac{1}{k} \sum_{i=1}^k \int_0^T s(t - \tau) \delta(t - t_i) dt \right\rangle \\ &= \left\langle \frac{1}{k} \int_0^T s(t - \tau) \sum_{i=1}^k \delta(t - t_i) dt \right\rangle \end{aligned}$$

(cont')

$$\begin{aligned}s_{avg}(\tau) &\approx \frac{1}{\langle k \rangle} \int_0^T s(t - \tau) \langle \sum_{i=1}^k \delta(t - t_i) \rangle dt \\ &= \frac{T}{\langle k \rangle} \left[\frac{1}{T} \int_0^T s(t - \tau) r(t) dt \right] \\ &= \frac{Q_{rs}(-\tau)}{\langle r \rangle}\end{aligned}$$

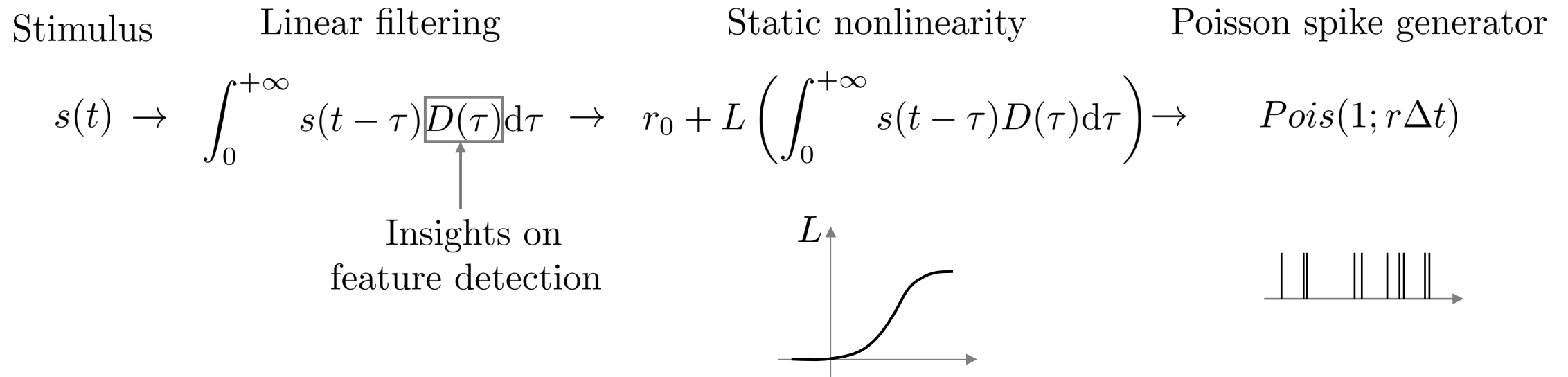
Recall that

$$D(\tau) = \frac{Q_{rs}(-\tau)}{\sigma^2},$$

we thus have

$$D(\tau) \approx \frac{\langle r \rangle s_{avg}}{\sigma^2}.$$

The model typically used in neuroscience is termed the linear-nonlinear-Poisson cascade model, as we simulate the spike generation process by passing the neural response rate function obtained from linear filtering through an additional static nonlinearity, followed by a Poisson spike generation process, i.e.



For visual stimuli / images, we can similarly model neural response rate by

$$r_{est}(t) = r_0 + L \left(\int_0^\infty \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} s(x, y, t - \tau) D(x, y, \tau) dx dy d\tau \right).$$

Where the spike-triggered average is now

$$s_{avg}(x, y, \tau) = \frac{Q_{rs}(x, y, -\tau)}{\langle r \rangle}.$$

While the optimal filter is similarly given by

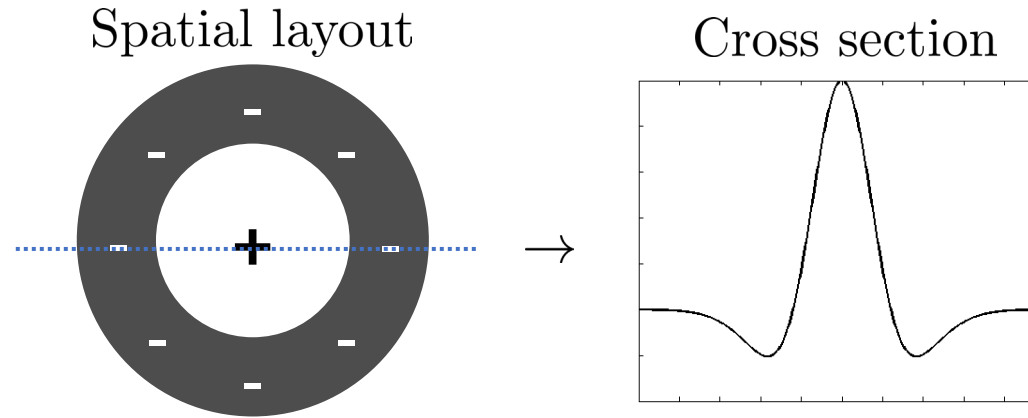
$$D(x, y, \tau) = \frac{\langle r \rangle s_{avg}(x, y, \tau)}{\sigma^2}.$$

In a subset of neurons, $D(x, y, \tau)$ is spatiotemporally separable, i.e.

$$D(x, y, \tau) = \boxed{D_s(x, y)} D_t(\tau).$$

↑
Spatial feature

Fitting the LNP models to neural responses of the retinal ganglion cells (RGCs, i.e. retinal output neurons) and the lateral geniculate nucleus, we obtain spatial receptive field (RF) (spatial profile of the optimal filter D) with the following structure for ON-center cells (with polarity reversed for OFF-center cells):



Which can be fitted well by the difference-of-Gaussian (DOG) function

$$D_s(x, y) = \pm \left(\frac{1}{2\pi\sigma_{\text{cen}}^2} \exp \left(-\frac{x^2 + y^2}{2\sigma_{\text{cen}}^2} \right) - \frac{\lambda}{2\pi\sigma_{\text{sur}}^2} \exp \left(-\frac{x^2 + y^2}{2\sigma_{\text{sur}}^2} \right) \right),$$

where λ controls the relative magnitudes of the center and surround field contributions.

Whereas in the primary visual cortex (V1), a subset of neurons have RFs with the following spatial structure:

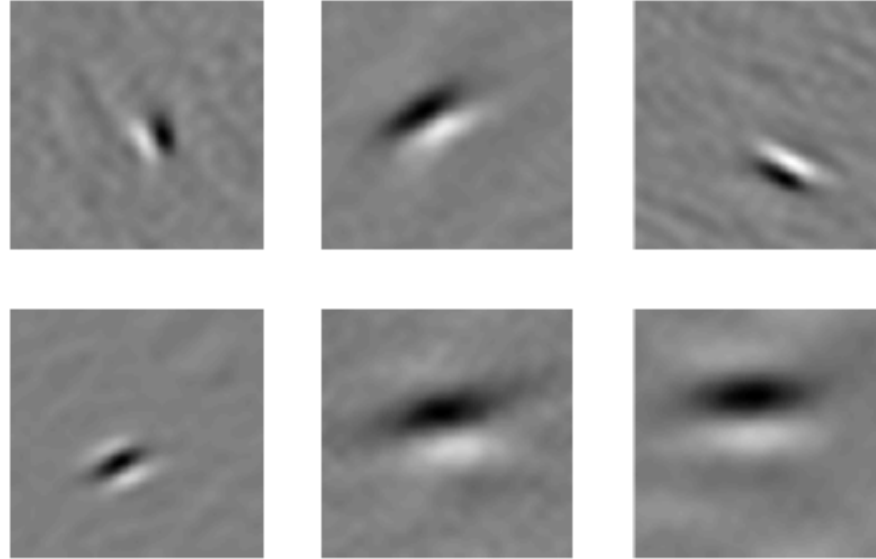


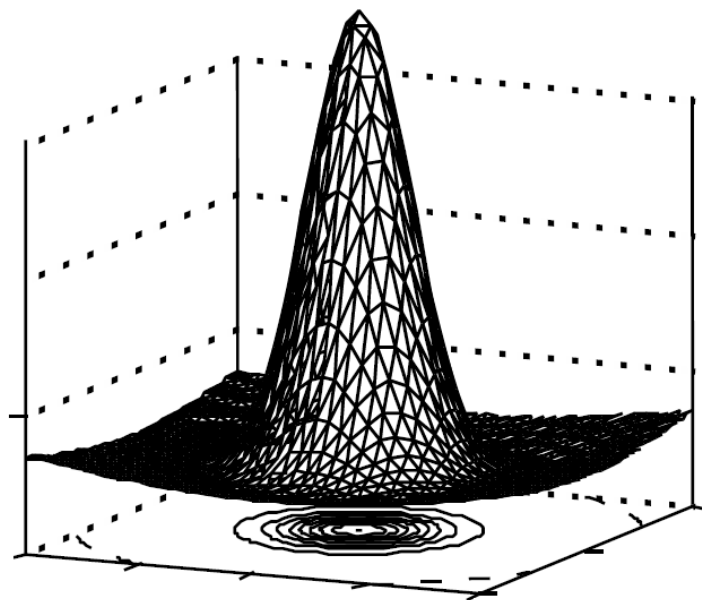
Image courtesy: Dario Ringach (UCLA) in Hyvarinen et al (2007)

That can be well-described by the Gabor function

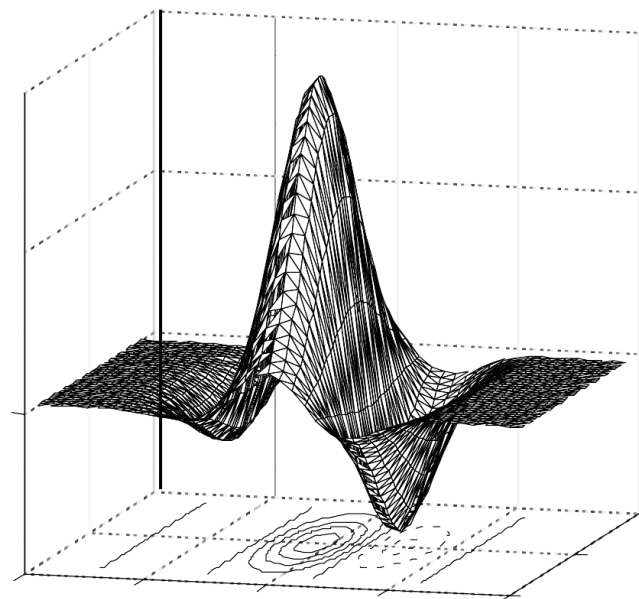
$$D_s(x, y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}\right) \cos(kx - \phi).$$

Why use these kinds of filters or feature detectors?

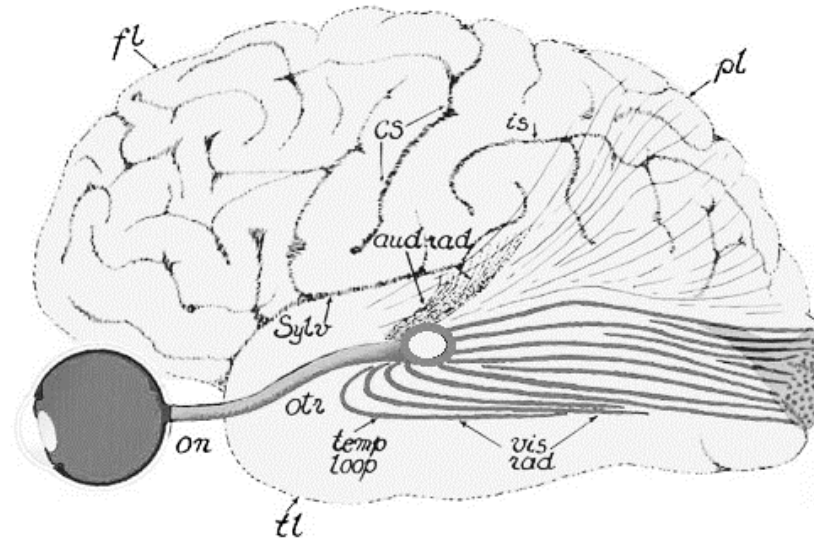
Difference of Gaussian



Gabor



Consider the anatomy of the retina and optic nerve:



The eyes should preferably be located at a relatively peripheral location, and able to move freely independent of the head.

Thus the best design is to attach a bundle of nerves (the optic nerve) for it to send out signals, which must also be thin.

Consequentially, there is a need to transmit signals compactly and efficiently.

Consider the form of visual representation in V1

Action potentials are metabolically costly, while sensory representation have to be more explicit in certain ways to be utilized by higher-order areas of the brain to extract meanings.

In another words, it is desirable to form a representation such that fewer active units are needed to describe a certain scene.

To make representation more explicit, the most straightforward thing to do next, is connecting the dots detected by the retina/LGN, and reperesent edges instead.

Two seminal hypotheses made by Horace Barlow FRS (1921 - 2020):

The redundancy reduction hypothesis (1961)

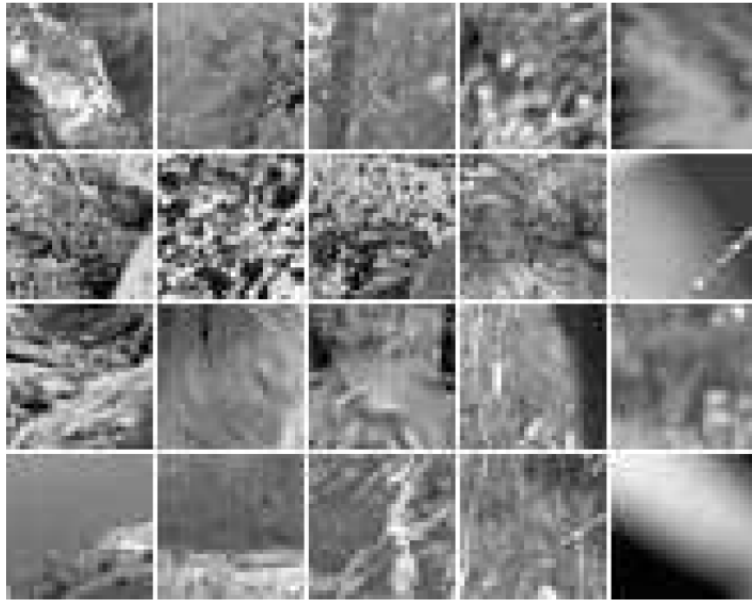
Neurons should encode sensory information in ways whereby statistical dependencies among them are minimized for best utilization of resources for transmission (i.e. adopting an *efficient coding* strategy).

The neuron doctrine of perception (1972)

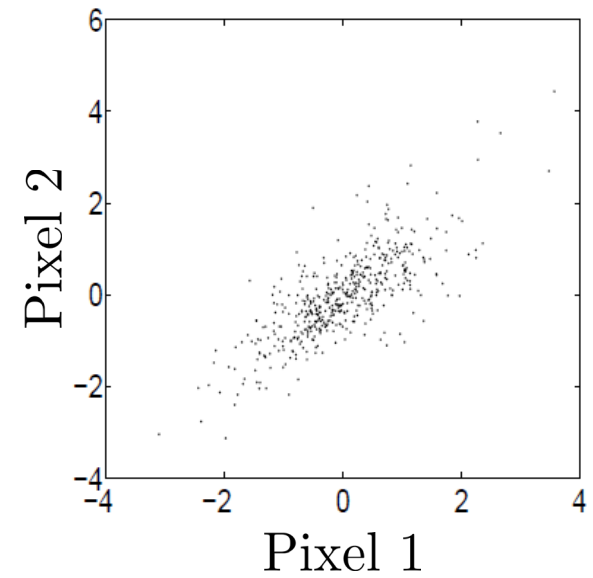
Neural representations are organized to describe sensory stimuli with few active units that are nonredundant (i.e. adopting a *sparse coding* strategy), and such representations are matched to the statistical properties of natural stimuli.

The real purpose eventually, is to model the *causes* of the stimuli, while the efficient coding and sparse coding strategies are means under specific circumstances for later-stage extraction of relevant information.

Natural image statistics: pixel intensity correlation



Neighboring pixels
 $\Rightarrow$



Natural image statistics: power / amplitude spectra

$$I(x, y)$$

Fourier transform
 $\Rightarrow$

$$\text{Amplitude spectrum: } \sqrt{|\hat{I}(u, v)|}$$



A



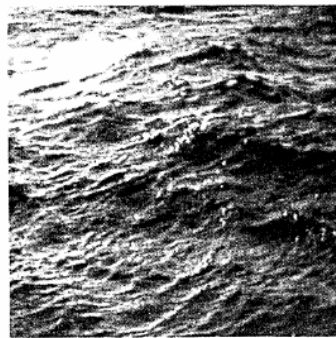
B



C



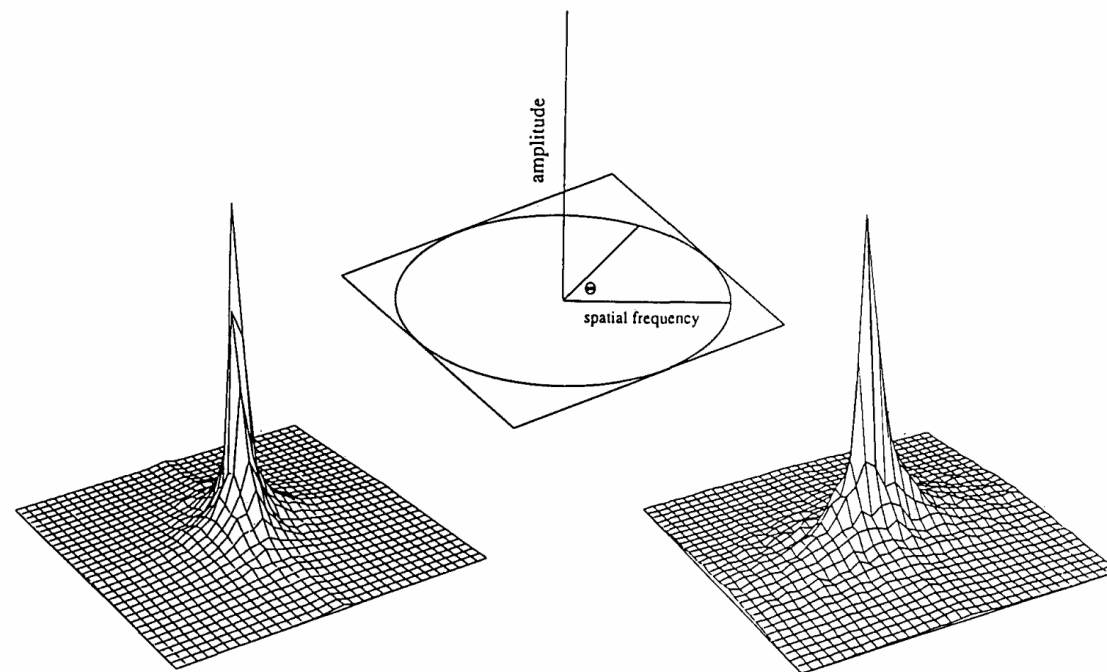
D



E



F



Field (1987)

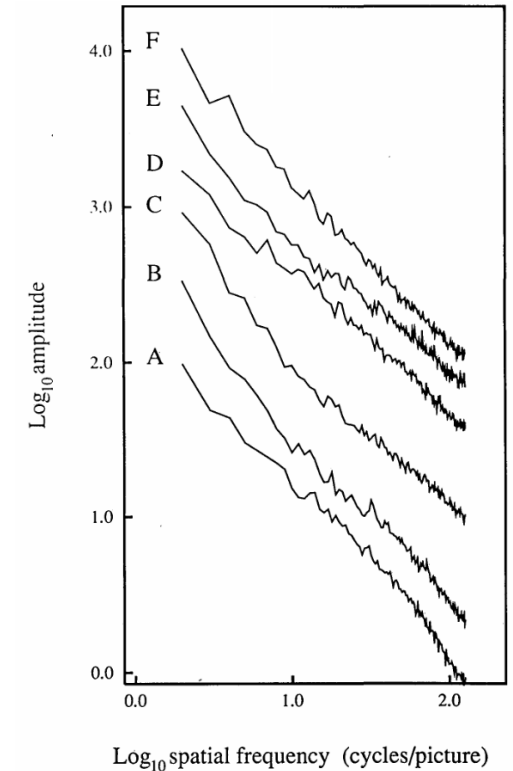
Natural image statistics: power / amplitude spectra

An observation:

$$\log \sqrt{|\hat{I}(u, v)|} \approx -\log f + c(\theta),$$

for many natural scenes, where $f = \sqrt{u^2 + v^2}$ is the spatial frequency, and $c(\theta)$ is a scaling factor.

Implying: the scale-invariance of natural images.



Field (1987)

Why the scale-invariance? Suppose I_1 is a natural image whose power spectrum fulfills (equivalent to the relationship stated in the previous slide)

$$|\hat{I}_1(u, v)| = \frac{A(\theta)}{f^2},$$

where $A(\theta) = e^{2c(\theta)}$.

Let I_2 be a rescaled version of I_1 , then

$$\begin{aligned} I_2(x, y) &= I_1(\alpha x, \alpha y) \\ \hat{I}_2(u, v) &= \mathcal{F} [I_1(\alpha x, \alpha y)] \\ |\hat{I}_2(u, v)| &= \frac{1}{\alpha^2} \left| \hat{I}_1\left(\frac{u}{\alpha}, \frac{v}{\alpha}\right) \right| \\ &= \frac{1}{\alpha^2} \frac{A(\theta)\alpha^2}{f^2} \\ &= \frac{A(\theta)}{f^2} \end{aligned}$$

Thus the power / amplitude spectra of the rescaled I_1 remain unchanged.

Next we consider the amplitude spectrum of output from retina. Assuming that the retina consists of an array of translational-invariant filters D_s (which is biologically reasonable), the output of retina is described by

$$\begin{aligned} R(x, y) &= \int \int I(x - x', y - y') D_s(x', y') dx' dy' \\ &= I(x, y) * D_s(x, y) \end{aligned}$$

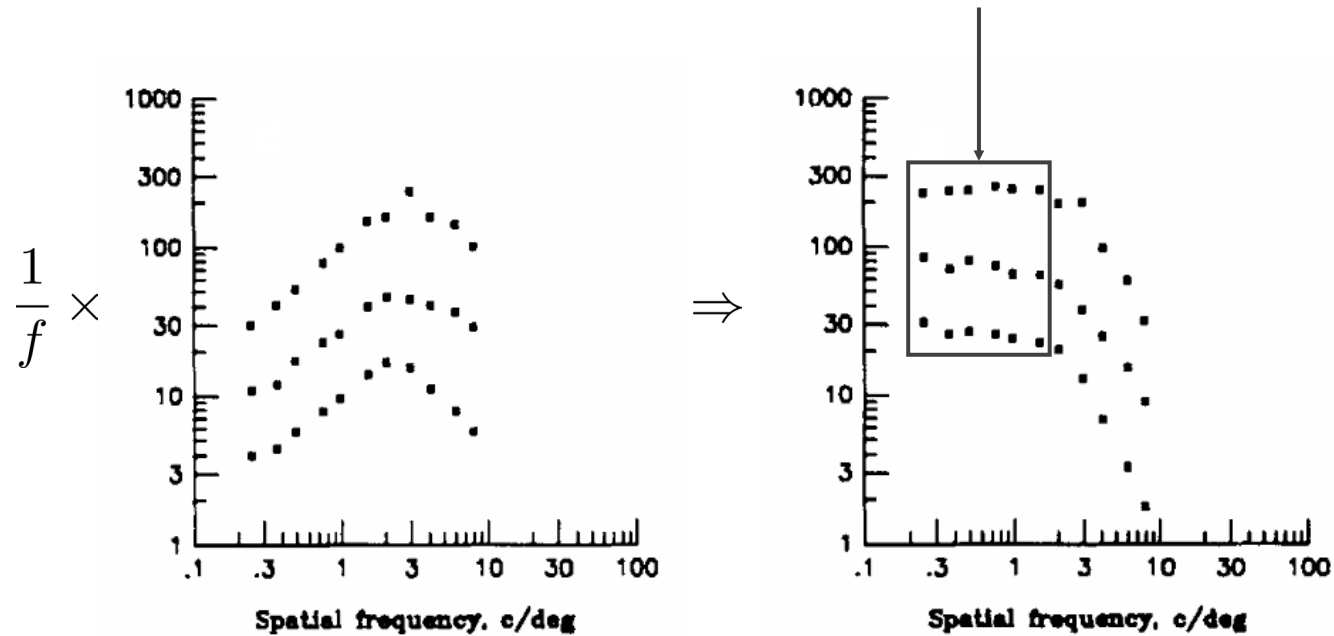
Based on the convolution theorem,

$$\begin{aligned} \hat{R}(u, v) &= \hat{I}(u, v) \hat{D}_s(u, v) \\ |\hat{R}(u, v)| &= |\hat{I}(u, v)| \cdot |\hat{D}_s(u, v)| \end{aligned}$$

Atick (1992) was the first to point out

$$\sqrt{|\hat{I}(u, v)|} \sqrt{|\hat{D}_s(u, v)|} \propto \frac{1}{f} \sqrt{|\hat{D}_s(u, v)|}$$

= constant up to certain cut-off frequency



A flat amplitude spectrum would mean that with natural images as input, the retinal output is *decorrelated* in space (up to cut-off frequency).

A spatially decorrelated retinal output signifies that the redundancy of information carried by neighboring units is reduced.

Thus in the retina, experimental evidence does support the redundancy reduction or efficient coding hypothesis.

For high frequency features, this relationship does not hold, instead the retina acts like a low pass filter.

A natural next question to ask: is it possible to derive the RF structures, assuming some postulated coding principles?

Our starting point is a linear (discrete) image synthesis model:

Given a zero-centered image set, for each image I we assume that it has arisen as a linear combination of a set of basis functions, i.e.

$$I = \Phi x,$$

where I has dimensionality that equals the number of pixels (e.g. $(M \cdot N) \times 1$ for an image with M row and N columns), columns of Φ are the basis functions (which have the same dimensionality as the image), while x is the column vector of coefficients (or weights) for each basis function.

For the images, our goal is to devise a filter set W (with each row as one filter), that uncovers the weights for the basis functions (analogous to having a neural representation):

$$u = WI,$$

where u is related to x by a permutation matrix P and scaling matrix S

$$u = W\Phi x = PSx.$$

If we demand that the output of filtering with W is spatially decorrelated (i.e. whitened), then we may seek W satisfying the constraint that

$$\text{cov}(u, u) = \text{a diagonal matrix.}$$

We can further constrain that

$$\text{cov}(u, u) = \mathbb{I},$$

where $\mathbb{I}$ is the identity matrix, given that the scaling of each basis is arbitrary.

The covariance matrix of $\mathbf{u}$ is related to that of $\mathbf{I}$ by

$$\begin{aligned}\text{cov}(\mathbf{u}, \mathbf{u}) &= \langle (\mathbf{W}\mathbf{I})(\mathbf{W}\mathbf{I})^T \rangle \\ &= \langle \mathbf{W}\mathbf{I}\mathbf{I}^T\mathbf{W}^T \rangle \\ &= \mathbf{W}\langle \mathbf{I}\mathbf{I}^T \rangle \mathbf{W}^T \\ &= \mathbf{W} \text{cov}(\mathbf{I}, \mathbf{I}) \mathbf{W}^T\end{aligned}$$

Since any symmetric matrix has a symmetric eigenvalue decomposition, there exists $\mathbf{U}$ and $\mathbf{\Lambda}$ such that

$$\text{cov}(\mathbf{I}, \mathbf{I}) = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,$$

where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues, while the columns of $\mathbf{U}$ are the corresponding eigenvectors (which are orthonormal, i.e. $\mathbf{U}^T\mathbf{U} = \mathbb{I}$).

Our problem thus becomes identifying W such that

$$\begin{aligned} W U \Lambda U^T W^T &= \mathbb{I} \\ [W U \Lambda^{\frac{1}{2}}][\Lambda^{\frac{1}{2}} U^T W^T] &= \mathbb{I} \end{aligned}$$

Among infinitely many solutions, two simple ones (making use of $U^T U = \mathbb{I}$) for performing the whitening are

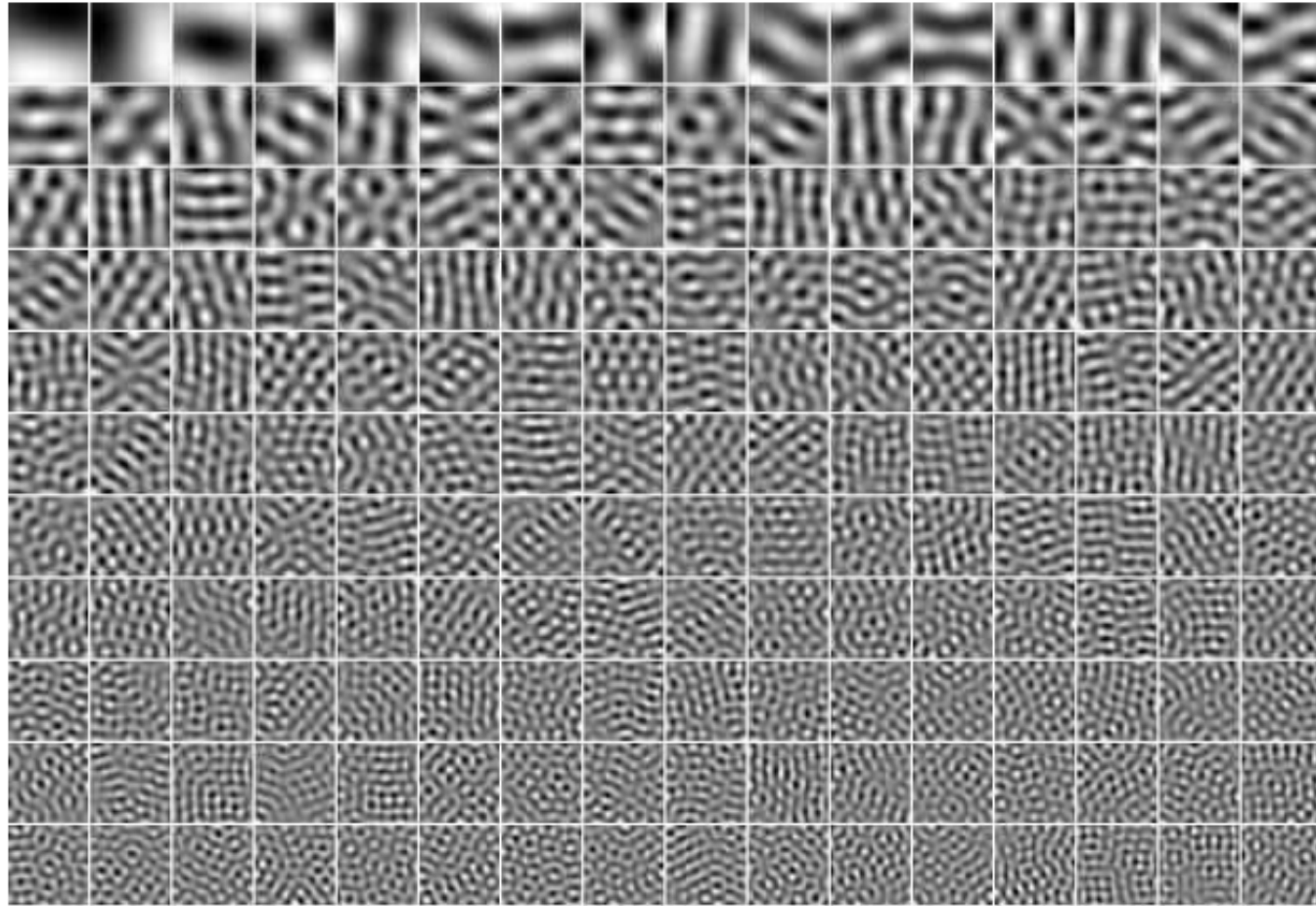
$$W = \Lambda^{-\frac{1}{2}} U^T,$$

and

$$W = U \Lambda^{-\frac{1}{2}} U^T,$$

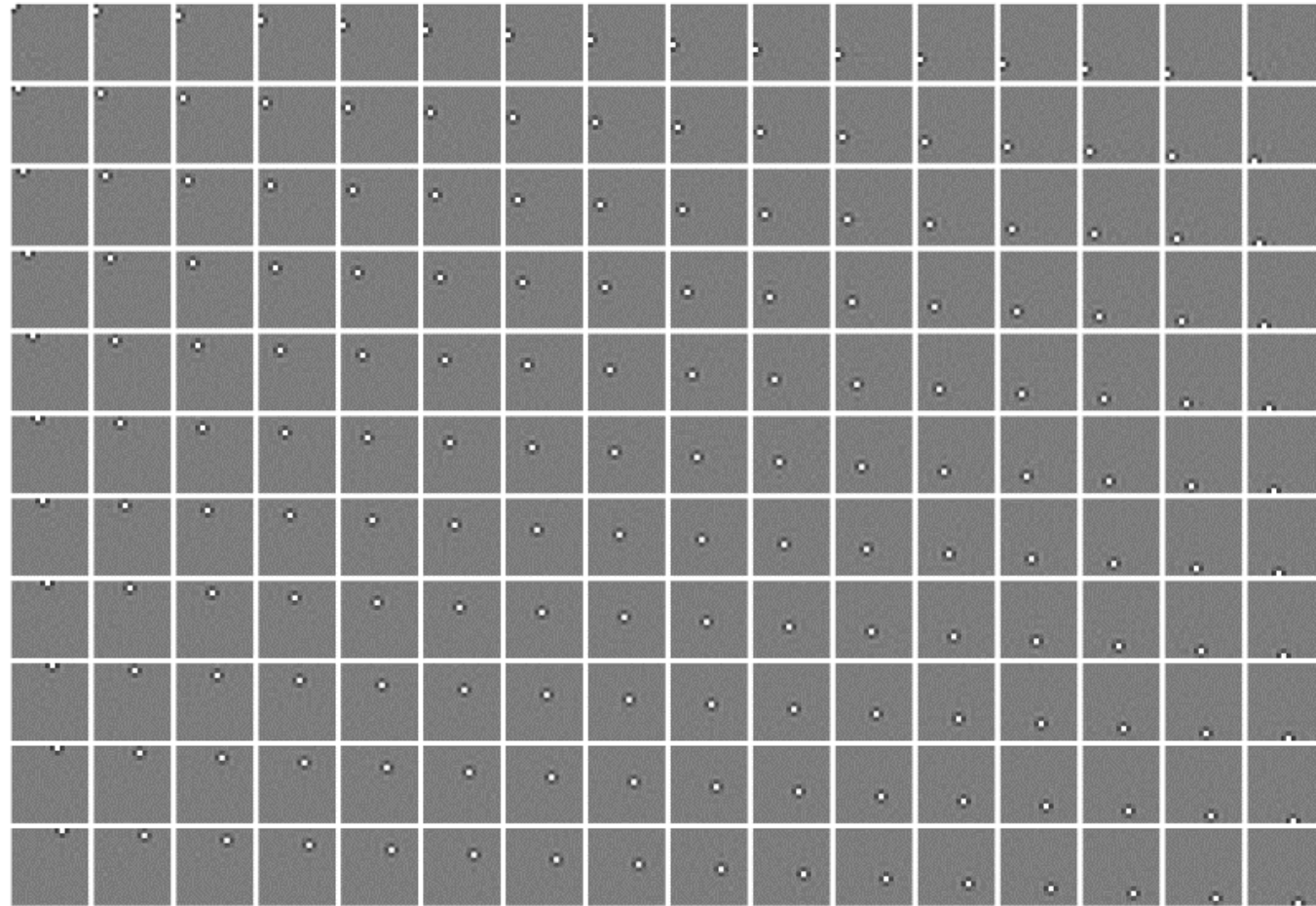
which are the classical principle component analysis and the symmetric solutions for W , respectively.

Obtaining $W = \Lambda^{-\frac{1}{2}} U^T$ from a set of natural image patches, one would obtain filters as these:



Hyvarinen et al (2007)

Obtaining $W = U\Lambda^{-\frac{1}{2}}U^T$ from a set of natural image patches, one would obtain filters that resemble center-surround receptive fields:



Hyvarinen et al (2007)

We have therefore shown that with natural images as input, by assuming the need to have spatially decorrelated output from the retina, we can obtain filters resembling the center-surround RF structures seen in the retina and the LGN (as one of the whitening solutions).

How about the elongated RFs of V1?

Olshausen and Field (1996) did the seminal work showing that one can in fact obtain basis functions and filter sets resembling elongated RFs by adopting a sparse coding model of natural images.

The starting point is a similar linear image synthesis model

$$\mathbf{I} = \Phi \mathbf{x} + \mathbf{v},$$

but in a probabilistic context, where $\mathbf{x}$ is the column vector of linear coefficients for each basis function, and $\mathbf{v}$ is Gaussian additive noise.

The probability of $\mathbf{I}$ given $\mathbf{x}$ is modelled by

$$P(\mathbf{I}|\mathbf{x}, \Theta) = \frac{1}{Z_v} e^{-\frac{1}{2\sigma_v^2} |\mathbf{I} - \Phi \mathbf{x}|^2},$$

where σ_v^2 is the variance of the noise, Z_v is a normalizing constant, and Θ denotes all model parameters.

It is assumed that the goal of the perceptual system is to infer the most probable $\mathbf{x}$ given the image and basis functions

$$P(\mathbf{x}|\mathbf{I}, \Theta) \propto P(\mathbf{I}|\mathbf{x}, \Theta)P(\mathbf{x}|\Theta)$$

$$\hat{\mathbf{x}} = \arg \max_{\mathbf{x}} P(\mathbf{x}|\mathbf{I}, \Theta)$$

The implementation of efficient and sparse coding is achieved by assuming that

$$P(\mathbf{x}|\Theta) = \prod_{i=1}^m P(x_i|\Theta),$$

where $P(x_i|\Theta)$ is a probability distribution that encourages sparseness (i.e. peaking at zero with a heavy tail), e.g.

$$P(x_i|\Theta) = \frac{1}{Z_\beta} e^{-\beta S(x_i)},$$

where S is a function that is used along with parameter β and normalizing constant Z_β to shape $P(x_i|\Theta)$ as required.

The problem of identifying the optimal basis functions can then be formulated as a two-step optimization problem:

$$\begin{aligned}
\hat{\Phi} &= \arg \max_{\Phi} \max_{\mathbf{x}} \ln P(\mathbf{I}|\Theta) \\
&\approx \arg \min_{\Phi} \min_{\mathbf{x}} -\ln(P(\mathbf{I}|\mathbf{x}, \Theta)P(\mathbf{x}|\Theta)) \\
&= \arg \min_{\Phi} \min_{\mathbf{x}} - [\ln(P(\mathbf{I}|\mathbf{x}, \Theta)) + \ln(P(\mathbf{x}|\Theta))] \\
&= \arg \min_{\Phi} \min_{\mathbf{x}} \left[\frac{1}{2\sigma_v^2} |\mathbf{I} - \Phi\mathbf{x}|^2 + \beta \sum_{i=1}^m S(x_i) \right] \\
&= \arg \min_{\Phi} \min_{\mathbf{x}} \left[|\mathbf{I} - \Phi\mathbf{x}|^2 + \lambda \sum_{i=1}^m S(x_i) \right]
\end{aligned}$$

where $\lambda = 2\sigma_v^2\beta$.

For a given Φ , we find $\hat{\mathbf{x}}$ by gradient descent, where the learning rule is given by

$$\Delta \mathbf{x} \propto 2\Phi^T (\mathbf{I} - \Phi \mathbf{x}) - \lambda S'(\mathbf{x})$$

where $S'(\mathbf{x}) = [S'(x_1), S'(x_2), \dots S'(x_m)]^T$.

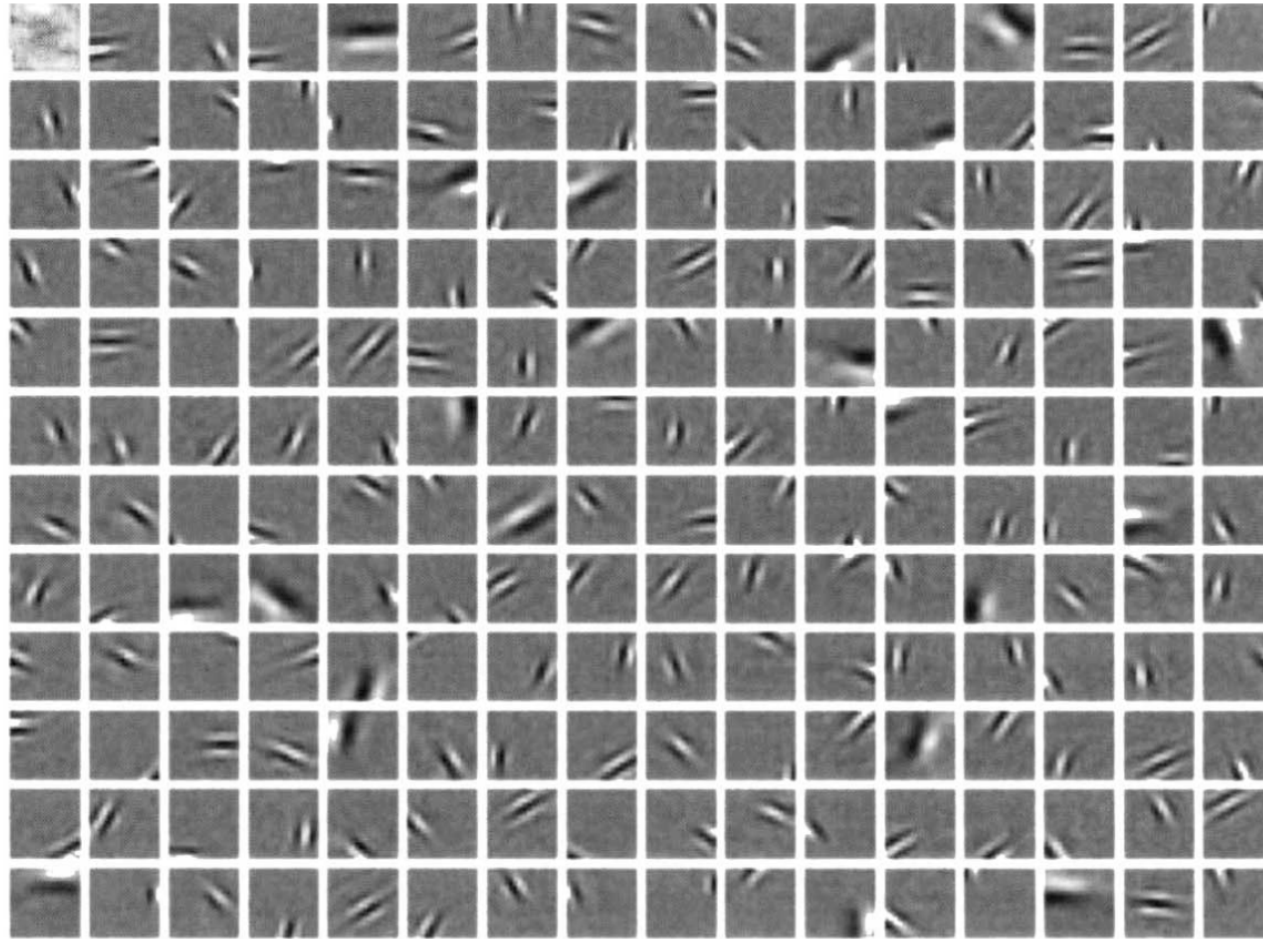
Then with fixed $\hat{\mathbf{x}}$, we seek to find

$$\begin{aligned} \hat{\Phi} &= \arg \min_{\Phi} \left[|\mathbf{I} - \Phi \hat{\mathbf{x}}|^2 + \lambda \sum_{i=1}^m S(\hat{x}_i) \right] \\ &= \arg \min_{\Phi} |\mathbf{I} - \Phi \hat{\mathbf{x}}|^2 \end{aligned}$$

which is also carried out by gradient descent, with the learning rule as

$$\Delta \Phi \propto (\mathbf{I} - \Phi \hat{\mathbf{x}}) \hat{\mathbf{x}}^T$$

With this optimization procedure run over whitened natural image patches, Olshausen and Field (1996) obtained basis functions as follows:



Olshausen et al (1996)

Concluding remarks

Modelling stimulus-spike relationship (i.e. neural encoding) in probabilistic frameworks has gained us insights into coding principles adopted by the early visual pathway.

In the retina, with main constraint being the need to send signal out compactly via the optic nerves, RF structures can be explained by the need for spatial decorrelation of output.

In V1, basis functions or filter set resembling elongated RFs can be obtained by different theoretical approaches, one of which being by assuming an efficient, sparse code.

Further topics

- Methodology of estimating RF using natural image stimuli
- Cells with nonlinear RFs (e.g. complex cells in V1)
- Direction selectivity / spatiotemporal inseparable RFs
- Sparse coding of natural scene videos
- Coding of colors
- Coding in higher visual areas

Main references

- Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. Olshausen and Field, 1996.
- The Independent Components of Natural Scenes are Edge Filters. Bell and Sejnowski, 1997.
- Theoretical Neuroscience, Dayan and Abbott, 2000.
- Natural Image Statistics - A Probabilistic Approach to Early Computational Vision. Hyvarinen, Hurri and Hoyer, 2009.