Understanding early visual coding from information theory

By Li Zhaoping

EU-India-China summer school, ISI, June 2007

Contact: z.li@ucl.ac.uk

Facts: neurons in early visual stages: retina, V1, have particular receptive fields. E.g., retinal ganglion cells have center surround structure, V1 cells are orientation selective, color sensitive cells have, e.g., red-center-green-surround receptive fields, some V1 cells are binocular and others monocular, etc.

Question: Can one understand, or derive, these receptive field structures from some first principles, e.g., information theory?

Example: visual input, 1000x1000 pixels, 20 images per second --- many megabytes of raw data per second.

Information bottle neck at optic nerve.

Solution (Infomax): recode data into a new format such that data rate is reduced without losing much information.

Redundancy between pixels.

1 byte per pixel at receptors $\longrightarrow$ 0.1 byte per pixel at retinal ganglion cells?

Consider redundancy and encoding of stereo signals

S^L

S^R

Redundancy is seen at correlation matrix (between two eyes)

$$R^S \equiv \begin{pmatrix} < S_L^2 > & < S_L S_R > \\ < S_R S_L > & < S_R^2 > \end{pmatrix} = < S_L^2 > \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

0<= r <= 1.

Assume signal ($S_{L,}$ S_R) is gaussian, it then has probability distribution:

$$P(\mathbf{S}) \propto exp(-\sum_{ij} S_i \tilde{S_j} (R^S)^{-1}_{ij}/2))$$

An encoding:

$$O_+ = S_+ \equiv (S_L + S_R)/\sqrt{2}, \qquad O_- = S_- \equiv (S_L - S_R)/\sqrt{2}$$

Gives zero correlation $\langle O_+ O_- \rangle$ in output signal (O_+, O_-), leaving output Probability

$$P(O_+, O_-) = P(O_+)\, P(O_-)$$ factorized.

The transform S to O is linear.

O_+ is binocular, O_- is more monocular-like.

Note: S_+ and S_- are eigenvectors or principal components of the correlation matrix $R^{S,}$ with eigenvalues $\langle S^2_\pm \rangle = (1 \pm r) \langle S_L^2 \rangle$

In reality, there is input noise $N_{L,R}$ and output noise $N_{o,\pm}$, hence:

$$O_{\pm} = [(S_L + N_L) \pm (S_R + N_R)]/\sqrt{2} + N_{o,\pm},$$

Effective output noise:

$$N_{\pm} = (N_L \pm N_R)/\sqrt{2} + N_{o,\pm}.$$

Let:

$$< N^2 > \equiv < N_L^2 > = < N_R^2 >, \text{ and } < N_o^2 > \equiv < N_{o,+}^2 > = < N_{o,-}^2 >.$$

Input $S_{L,R}$+ $N_{L,R}$ has

$$I_{L,R} = \frac{1}{2}\log_2 \frac{< S_{L,R}^2 > + < N^2 >}{< N^2 >}$$

Bits of information about signal $S_{L,R}$

Input $S_{L,R}$+ $N_{L,R}$ has

$$I_{L,R} = \frac{1}{2}\log_2 \frac{< S_{L,R}^2 > + < N^2 >}{< N^2 >}$$

bits of information about signal $S_{L,R}$

Whereas outputs $O_{+,-}$ has

$$I_{\pm} = \frac{1}{2}\log_2 \frac{< O_{\pm}^2 >}{< N_{\pm}^2 >} = \frac{1}{2}\log_2 \frac{< S_{\pm}^2 > + < N^2 > + < N_o^2 >}{< N^2 > + < N_o^2 >}$$

bits of information about signal $S_{L,R}$

Note: redundancy between S_L and S_R cause higher and lower signal powers $<O_+^2>$ and $<O_-^2>$ in O_+ and O_- respectively, leading to higher and lower information rate I_+ and I_-

If cost ~ $<O_{\pm}^2>$

$$I_{\pm} = \tfrac{1}{2}\log_2(< O_{\pm}^2 >) + \text{constant} = \tfrac{1}{2}\log_2(\text{cost}) + \text{constant},$$

Gain in information per unit cost $(\Delta I/\Delta \tilde{\text{cost}})$

smaller in O_+ than in O_- channel.

If cost ~ $<O_\pm^2>$

$$I_\pm = \frac{1}{2}\log_2(< O_\pm^2 >) + \text{constant} = \frac{1}{2}\log_2(\text{cost}) + \text{constant},$$

Gain in information per unit cost $(\Delta I/\Delta \tilde{\text{cost}})$

smaller in O_+ than in O_- channel.

Hence, gain control on $O_\pm$ is motivated.

$$O_\pm \longrightarrow g_\pm O_\pm$$

To balance the cost and information extraction, optimize by finding the gain g_+ such that

$$E(V_\pm) \equiv \sum_a (< O_a^2 >) - \lambda \sum_a (I_a) = \text{cost} - \lambda \cdot \text{Information}$$

Is minimized. This gives, for k = + or -

$$g_k^2 \quad \propto \quad \text{Max}\left\{ [\frac{1}{2}\frac{\langle S_k^2\rangle}{\langle S_k^2\rangle + \langle N^2\rangle}(1 + \sqrt{1 + \frac{4\lambda}{(\ln 2)\langle N_o^2\rangle}\frac{\langle N^2\rangle}{\langle S_k^2\rangle}}) - 1], 0 \right\}$$

$$g_k^2 \quad \propto \quad \text{Max}\left\{[\frac{1}{2}\frac{\langle S_k^2\rangle}{\langle S_k^2\rangle + \langle N^2\rangle}(1+\sqrt{1+\frac{4\lambda}{(\ln 2)\langle N_o^2\rangle}\frac{\langle N^2\rangle}{\langle S_k^2\rangle}})-1],0\right\}$$

In the zero noise limit when $\frac{<S_\pm^2>}{<N^2>} \gg 1$,

$$g^2 \propto <S^2>^{-1}$$

This equalizes the output power $<O_+^2> \approx <O_-^2>$ --- whitening

When output noise N_o is negligible, output O and input S+N convey similar amount of information about signal S, but uses much less output power with small gain $g_\pm$

$\langle O_+^2\rangle \sim \langle O_-^2\rangle$ --- whitening also means that output correlation matrix

$R^o_{ab} = \langle O_a O_b\rangle$

Is proportional to identity matrix, (since $\langle O_+ O_-\rangle = 0$).

Any rotation (unitary or ortho-normal transform):

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} O_+ \\ O_- \end{pmatrix} = \begin{pmatrix} \cos(\theta)O_+ + \sin(\theta)O_- \\ -\sin(\theta)O_+ + \cos(\theta)O_- \end{pmatrix}.$$

Preserves de-correlation $\langle O_1 O_2\rangle = 0$

Leaves output cost Tr (R^o) unchanged

Leaves amount of information extracted I = $\frac{1}{2}\log\frac{\det R^o}{\det R^N}$ unchanged

Tr, det, denote trace and determinant of matrix.

Both encoding schemes:

$$S_{L,R} \to O_{\pm} \text{ and } S_{L,R} \to O_{1,2},$$

With former a special case of latter, are optimal in making output decorrelated (non-redundant), in extracting information from signal S, and in reducing cost.

In general, the two different outputs:

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \begin{pmatrix} S_L(\cos(\theta)V_+ + \sin(\theta)V_-) + S_R(\cos(\theta)V_+ - \sin(\theta)V_-) \\ S_L(-\sin(\theta)V_+ + \cos(\theta)V_-) + S_R(-\sin(\theta)V_+ - \cos(\theta)V_-) \end{pmatrix}$$

prefer different eyes. In particular, $\theta = 45^o$ gives

$$O_{1,2} \sim S_L(g_+ \mp g_-) + S_R(g_+ \pm g_-)$$

The visual cortex indeed has a whole spectrum of neural ocularity.

Summary of the coding steps:

Input: S+N, with signal correlation (input statistics) R^s

get eigenvectors (principal components) S' of R^s

$$S + N \longrightarrow S' + N' = K_o(S+N)$$

rotation of coordinates

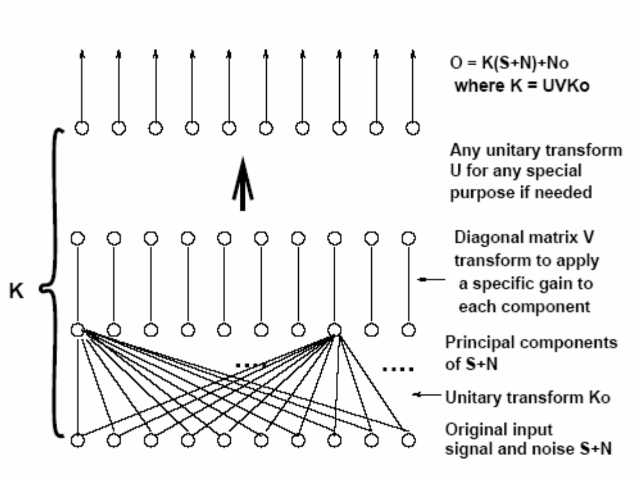
gain control V on each principal component

$$S' + N' \longrightarrow O = g(S'+N') + N_o$$

rotation U' (multiplexing) of O

$$O' \longrightarrow U'O = U' g K_o S + \text{noise}$$

Neural output = $\underbrace{U' g K_o}_{\text{Receptive field, encoding kernel}}$ sensory input + noise



Variations in optimal coding:

Factorial codes

Minimum entropy, or minimum description length codes

Independent components analysis

Redundancy reduction

Sparse coding

Maximum entropy code

Predictive codes

Minimum predictability codes, or least mutual information between output Channels.

They are all related!!!

Another example, visual space coding, i.e., spatial receptive fields

Signal at spatial location x is $S_x = S(x)$

Signal correlation is $R^S{}_{x,x'} = < S_x S_{x'}> = R^S(x-x')$ --- translation invariant

Principal components S_K are Fourier transform of S_x

$$S_x \to S_k \sim \sum_x K_o^{kx} S_x \sim \sum_x e^{-ikx} S_x$$

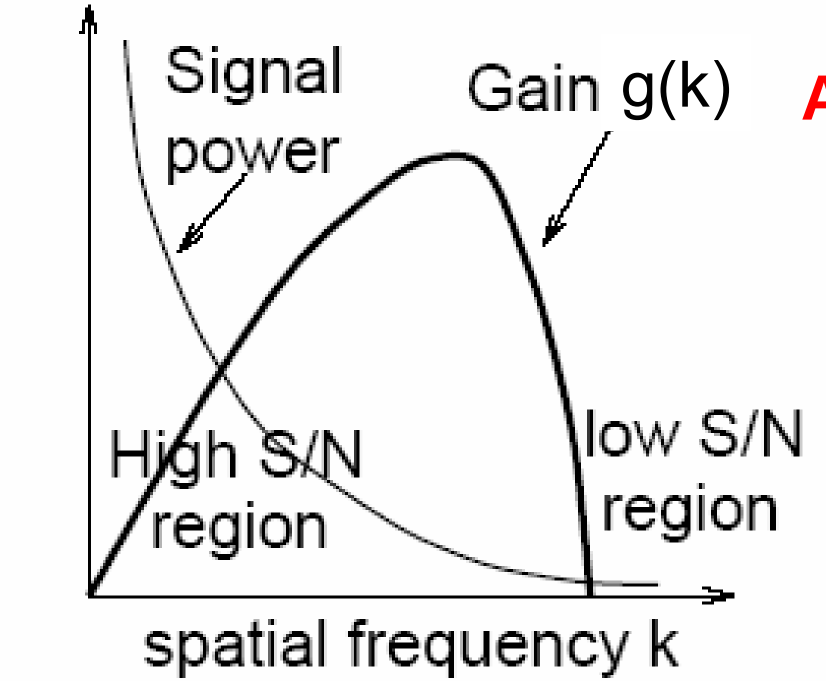
Eigenvalue spectrum (power spectrum): $< S_k^2 > \sim 1/k^2$

Assuming white noise power $<N_k^2>$ = constant, high S/N region is at low frequency, i.e., small k, region.

Gain control, $V(k) \sim <S^2_k>^{-1/2} \sim k$, --- whitening in space

At high k, where S/N is small, V(k) decays quickly with k to cut down noise according to

$$g_k^2 \propto \text{Max}\left\{[\frac{1}{2}\frac{\langle S_k^2\rangle}{\langle S_k^2\rangle + \langle N^2\rangle}(1+\sqrt{1+\frac{4\lambda}{(\ln 2)\langle N_o^2\rangle}\frac{\langle N^2\rangle}{\langle S_k^2\rangle}})-1], 0\right\}$$

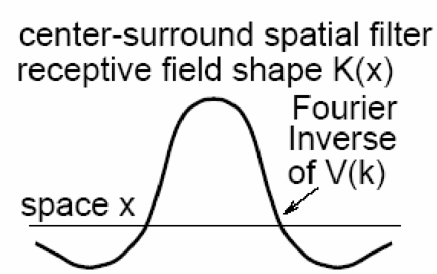


A band-pass filter

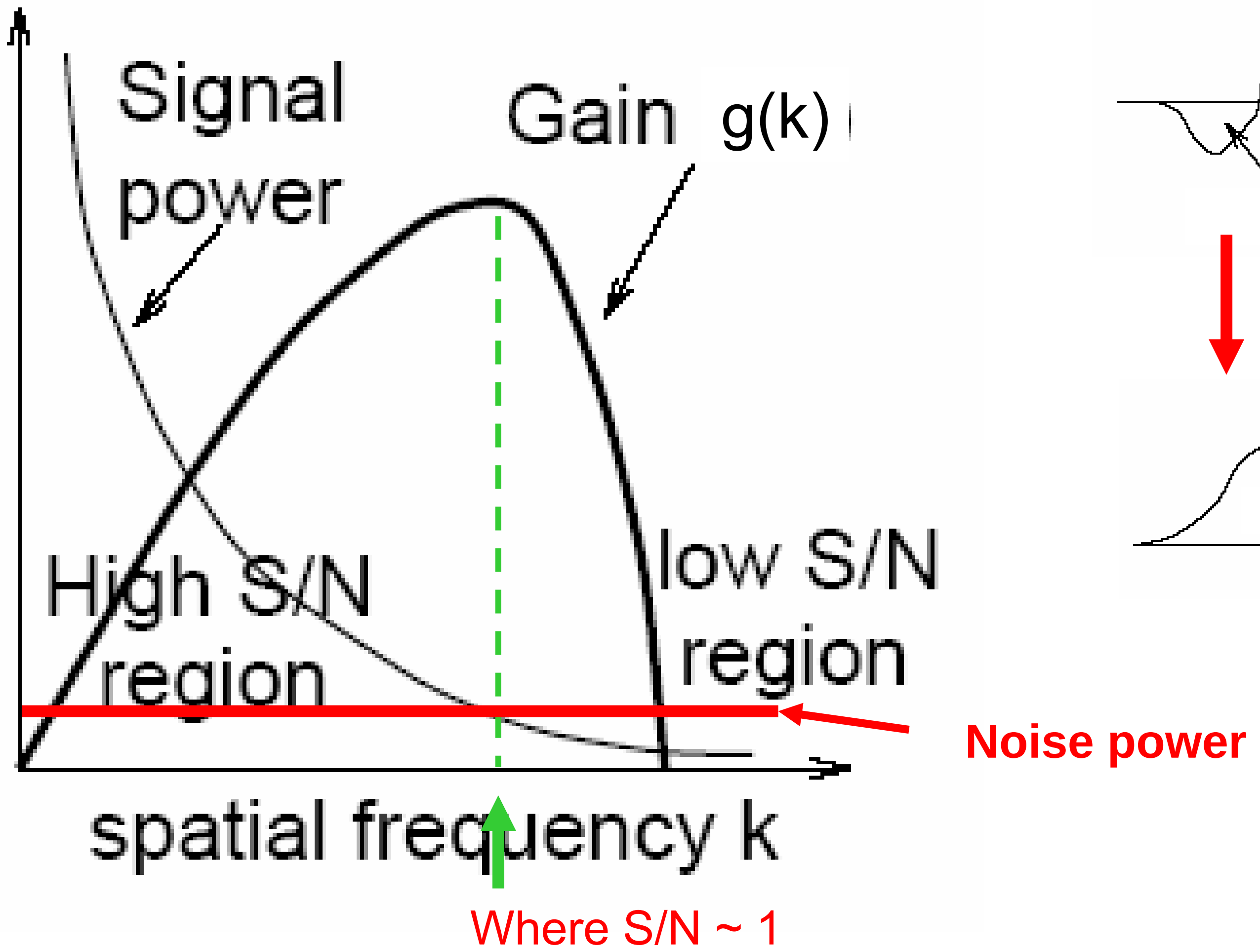
Let the multiplexing rotation be the inverse Fourier transform: $U^{x'k} \sim e^{ikx'}$

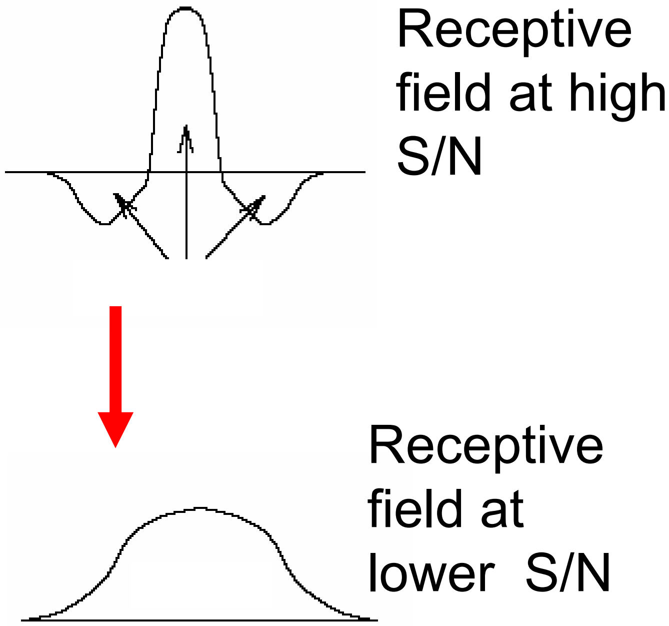
The full encoding transform is

$$O_{x'} = \Sigma_k U^{x'k} g(k) \Sigma_x e^{-kx} S_x = \Sigma_k \underbrace{g(k) \Sigma_x e^{-k(x'-x)}} S_x + \text{noise}$$



Understanding adaptation by input strength





When overall input strength is lowered, the peak of V(k) is lowered to lower spatial frequency k, a band-pass filter becomes a low pass (smoothing) filter.

Another example: optimal color coding

Analogous to stereo coding, but with 3 input channels, red, green, blue.

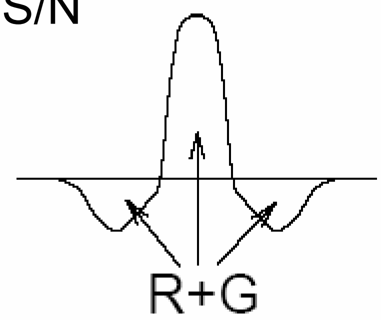
For simplicity, focus only on red and green

Input signal S_r, S_g Input correlation $R^S_{rg} > 0$

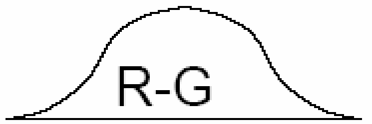
Eigenvectors: $S_r + S_g$ ← Luminance channel, higher S/N

$S_r - S_g$ ← Chromatic channel, lower S/N

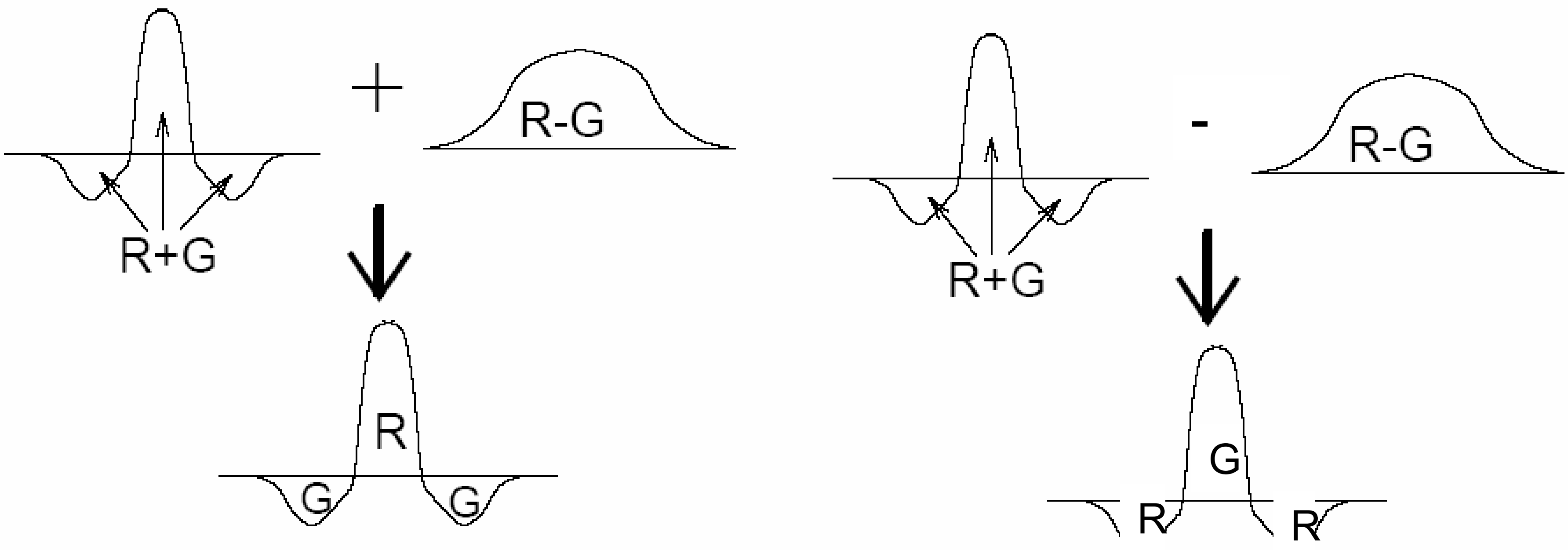
Gain control on $S_r + S_g$ --- lower gain until at higher spatial k



Gain control on $S_r - S_g$ --- higher gain then decay at higher spatial k



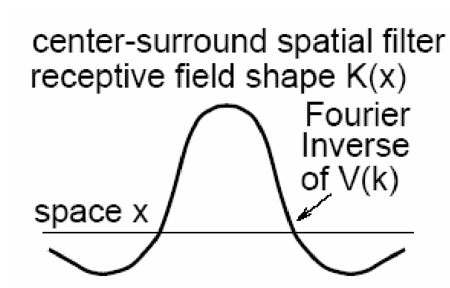
Multiplexing in the color space:



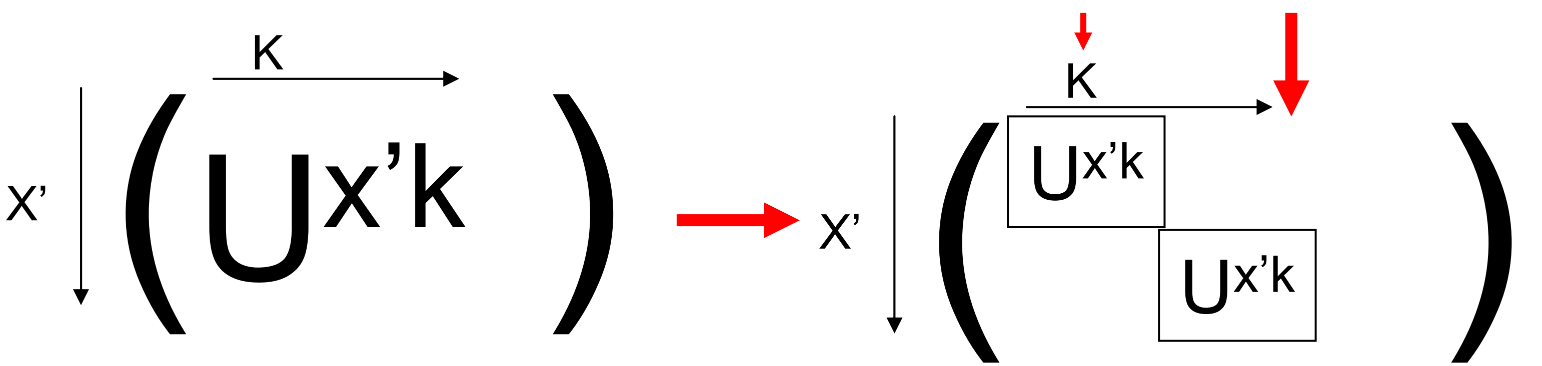
How can one understand the orientation selective receptive fields in V1?

Recall the retinal encoding transform:

$$O_{x'} = \Sigma_k U^{x'k} g(k) \Sigma_x e^{-kx} S_x = \Sigma_k g(k) \underbrace{\Sigma_x e^{-k(x'-x)}}\, S_x + \text{noise}$$

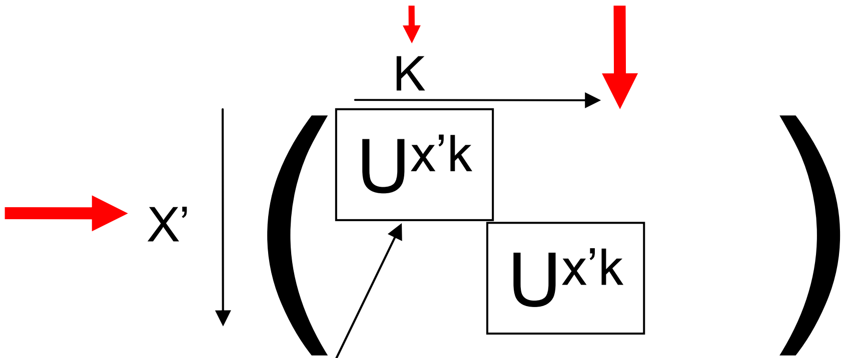


If one changes the multiplexing filter $U^{x'k}$, such that it is block diagonal, and for each output cell x', it is limited in frequency band in frequency magnitude and orientation --- V1 receptive fields.



V1 Cortical color coding

Different frequency bands



Lower frequency k bands, for chromatic channels

Higher frequency k bands, for luminance channel only

Orientation tuned cells

In V1, color tuned cells have larger receptive fields, have double opponency

Question: if retinal ganglion cell have already done a good job in optimal coding by the center-surround receptive fields, why do we need change of such coding to orientation selective? As we know such change of coding does not improve significantly the coding efficiency or sparseness.

Answer? Ref: (Olshausen, Field, Simoncelli, etc)

Why is there a large expansion in the number of cells in V1?
This leads to increase in redundancy, response in V1 from different cells are highly correlated.

What is the functional role of V1? It should be beyond encoding for information efficiency, some cognitive function beyond economy of information bits should be attributed to V1 to understand it.