# Theoretical understanding of the early visual processes by data compression and data selection

©Li Zhaoping
Department of Psychology, University College London, UK

This is the expansion from the original paper appearing in *Network: computation in neural systems* 2006. Much more detailed materials are added in this manuscript for teaching purposes. This is still a very rough draft — I hope to update the draft continuously and make it available to students. Feedbacks are welcome. If you like better explanations or more details in any parts of this manuscript, or if you think certain parts of the text are not clear or confusing, or anything else, please do not hesitate to contact me at z.li@ucl.ac.uk .

**Abstract:**
Early vision is best understood in terms of two key information bottlenecks along the visual pathway — the optic nerve and, more severely, attention. Two effective strategies for sampling and representing visual inputs in the light of the bottlenecks are (1) data compression with minimum information loss and (2) data deletion. This paper reviews two lines of theoretical work which understand processes in retina and primary visual cortex (V1) in this framework. The first is an efficient coding principle which argues that early visual processes compress input into a more efficient form to transmit as much information as possible through channels of limited capacity. It can explain the properties of visual sampling and the nature of the receptive fields of retina and V1. It has also been argued to reveal the independent causes of the inputs. The second theoretical tack is the hypothesis that neural activities in V1 represent the bottom up saliencies of visual inputs, such that information can be selected for, or discarded from, detailed or attentive processing. This theory links V1 physiology with pre-attentive visual selection behavior. By making experimentally testable predictions, the potentials and limitations of both sets of theories can be explored.

# Contents

# Chapter 1

# Introduction and scope

Vision is the most intensively studied aspect of the brain, physiologically, anatomically, and behaviorally (Zigmond et al 1999). Theoretical studies of vision suggest computational principles or hypotheses to understand why physiology and anatomy are as they are from behavior, and vice versa. The retina and V1, since they are better known physiologically and anatomically, afford greater opportunities for developing theories of their functional roles, since theoretical predictions can be more easily verified in existing data or tested in new experiments. This paper reviews some such theoretical studies. Focusing on the *why* of the physiology, it excludes descriptive models concerning *what* and *how*, e.g., models of the center-surround receptive fields of the retinal ganglion cells, or mechanistic models of how orientation tuning in V1 develops. Useful reviews of early vision with related or different emphases and opinions can be found in, e.g., Atick (1992), Meister and Berry 1999, Simoncelli and Olshausen (2001), Lennie (2003), Lee (2003), and Olshausen and Field (2005).

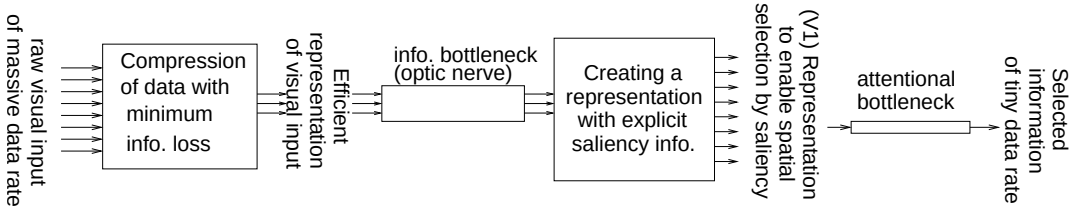



Figure 1.1: Process flow diagram illustrating two bottom-up strategies proposed for early vision to reduce data rate through information bottlenecks — (1) data compression with minimum information loss, and, (2) creating a saliency map to enable lossy selection of information.

Early vision creates representations at successive stages along the visual pathway, from retina to lateral geniculate nucleus (LGN) to V1. Its role is perhaps best understood in terms of how these representations overcome critical information bottlenecks along the visual pathway. This review focuses on the studies that developed these theories.

Retinal receptors could receive information at an estimated rate of $10^9$ bits per second (Kelly 1962), i.e., roughly 25 frames per second of images of 2000x2000 pixels at one byte per pixel. Along the visual pathway, the first obvious bottleneck is the optic nerve from retina to LGN en route to V1. One million ganglion cells in humans, each transmitting information at about one bit/second (Nirenberg, et al 2001) give a transmission capacity of only $10^6$ bits/second in the optic nerve, a reduction of 3 orders of magnitude. The second bottleneck is more subtle, but much more devastating. Visual attention is estimated as having the capacity of only 40 bits/second for humans (Sziklai 1956).

Data compression without information loss can reduce the data rate very effectively, and should thus be a goal for early vision. Engineering image compression methods, for instance the JPEG algorithm, can compress natural image data 20 fold without noticable information loss. However, the reduction from $10^9$ to 40 bits/second is heavily lossy, as demonstrated by our blindness to unattended visual inputs even when they are salient, the phenomenon known as inattentional blindness (Simons and Chabris 1999). Therefore, data deletion by information selection must occur along the visual pathway. An effective method of selection is to process only a limited portion of visual space at the center of vision (which has a higher spatial resolution). Then, selection should be such that the selected (rather than the ignored) location is more likely important or relevant to the animal. While attentional selection is often goal-directed, such as during reading when gaze is directed to the text locations, carrying out much of the selection quickly and by bottom-up (or autonomous) mechanisms is computationally efficient, and indeed essential to respond to unexpected events. Bottom up selection is more potent (Jonides 1981) and quicker (Nakayama and Mackeben 1989) than top-down selection, which could be based on features, or objects, as well as location (Pashler 1998). Early visual processes could facilitate bottom up selection by explicitly computing and representing bottom up saliency to guide selection of salient locations. Meanwhile, any data reduction before the selection should be as information lossless as possible, for any lost information could never be selected to be perceived. This suggests a process flow diagram in Fig. (1.1) for early vision to incorporate sequentially two data reduction strategies: (1) data compression with minimum information loss and (2) creating a representation with explicit saliency information to facilitate selection by saliency.

First I review studies motivated by the first data reduction strategy. It has been argued that early visual processes should take advantage of the statistical regularities or redundancies of visual inputs to represent as much input information as possible given limited neural resources (Barlow 1961). Limits may lie in the number of neurons, power consumption by neural activities, and noise, leading to information or attentional bottlenecks. Hence, input sampling by the cones, and activity transforms by the receptive fields (RFs), should be optimally designed to encode the raw inputs in an efficient form, i.e., data compression with minimal information loss — an efficient coding principle. As efficient coding often involves removing redundant representations of information, it could also have the cognitive role of revealing the underlying independent components of the inputs, e.g., individual objects. This principle has been shown to explain, to various extents, the color sensitivities of cones, distributions of receptors on the retina, properties of RFs of retinal ganglion cells and V1 cells, and their behavioral manifestations in psychophysical performance. As efficiency depends on the statistics of input, neural properties should adapt to prevailing visual scenes, providing testable predictions about the effects of visual adaptation and development conditions.

An important question is the stage along the visual pathway at which massively lossy information selection should occur. Postponing lossy selection could postpone the irreversible information deletion, and unfortunately also the completion of cognitive processing. While it is reasonable to assume that data compression with minimum information loss may continue till little more efficiency can be gained, efficient coding should encounter difficulties in explaining major ongoing processing at the stage serving the goal of lossy selection. I will review the difficulties in using efficient coding to understand certain V1 properties such as the over-complete representation of visual inputs, and the influence on a V1 neuron's response of contextual inputs outside its RF. These properties will be shown to be consistent with the goal of information selection, the second data reduction strategy. Specifically, V1 is hypothesized (Li 1999ab, 2002, Zhaoping 2005) to create a bottom up saliency map of visual space, such that a location with a higher scalar value in this map is more likely selected. The saliency values are proposed to be represented by the firing rates of V1 neurons, such that the RF location of the most active V1 cell is most likely selected, regardless of its feature tuning. This hypothesis additionally links V1 physiology with the visual behavior of pre-attentive selection and segmentation, again providing testable predictions and motivating new experimental investigations.

This paper presents a particular, rather than an all-inclusive, view. While the efficient coding theory and the V1 saliency map theory involve different theoretical concepts and methodologies,

they both concern the understanding of early vision in terms of its role of overcoming information bottlenecks in the visual and cognitive pathways. The very same experimental data shaped the development of both theories, indicating that data exposing limitations in one theory can drive the development of another as we move from one visual stage to the next. The many gaps in our understanding of early vision, and in the coverage of previous work, will hopefully motivate stimulating discussions and future studies.

# Chapter 2

# The efficient coding principle

This section will review the formulation of this principle and its application to understand retina and V1 processes. Response properties of large monopolar cells (LMC) in blowfly's eye and the cone densities on human retina will illustrate optimal input sampling given a finite number of sensors or neural response levels. The RF transforms (in space, time, color, stereo) of the retinal ganglion cells and V1 cells will illustrate how input redundancy should be more or less reduced in low or high noise conditions respectively. Knowledge of the information theory should help the understanding of the analytical formulation of the effcient coding principle. A brief introduction to the information theory is provided below for this purpose.

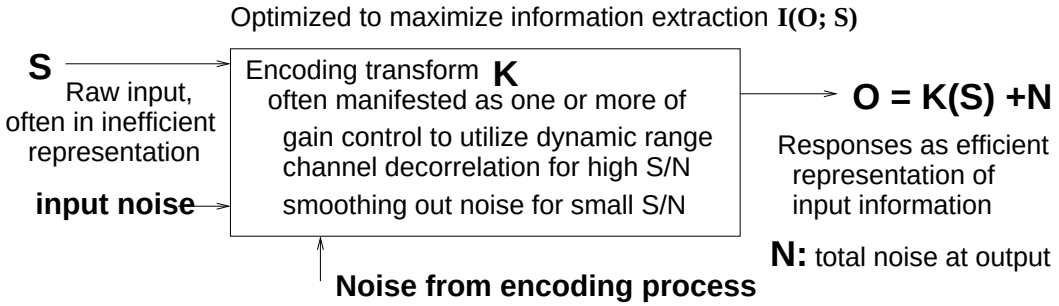



Figure 2.1: Efficient coding K transforms the signal $\mathbf{S}$ to neural responses $\mathbf{O}$ to extract maximum amount of information $I(\mathbf{O};\mathbf{S})$ about signal $\mathbf{S}$, given limited resources, e.g., capacity (dynamic range) or energy consumption of the output channels. Often, gain control accommodates the signal within the dynamic range. With high signal-to-noise (S/N), removing correlations between input channels makes information transmitted by different output channels non-redundant. With low S/N, averaging between channels helps smoothing out noise and recover inputs from correlated responses.

### 2.0.1 A brief introduction on information theory — skip if not needed

This brief introduction to information theory (Shannon and Weaver 1949) is for the purpose of getting sufficient intuition in order to adequately apply it to understand sensory information coding and transmission.

#### Measuring information amount

One is presumably familiar with the computer terminoloyg "bits". For instance, an integer between 0-255 needs 8 bits to represent or convey it. Before you know anything about that integer, you may

know that its can be equally likely any one integer from 0 up to 255, i.e., it has a probability of $P(n) = 1/256$ to be any $n \in 0 - 255$. However, once someone told you the exact number, say $n = 10$, this integer has a probability $P(n) = 1$ for $n = 10$ and $P(n) = 0$ otherwise, and you need no more bits of information to know more about this integer.

Note that $\log_2 256 = 8$, and $\log_2 1 = 0$. That is, before you know which one among the 256 possibilities $n$ is, it has

$$-\log_2 P(n) = \log_2 256 = 8 \text{ bits} \tag{2.1}$$

of information missing from you. Once you know $n = 10$, you miss no bits of information since $-\log_2 P(n = 10) = 0$.

Similarly, if you flip a coin, and each flip gives head or tail with equal probability, then there is only one bit of information regarding the outcome of the coin flipping, since there is a probability $P = 1/2$ for either head or tail, giving $-\log_2 P(\text{head}) = -\log_2 P(\text{tail}) = 1$ bit. Suppose that you can get information about integer $n \in [0, 255]$ by coin flipping. Say the first coin flip says by head or tail whether $n \in 0 - 127$ or $n \in 128 - 255$. After this coin flip, let us say that it says $n \in 0 - 127$. Then you flip the coin again, and this time to determine whether $n \in 0 - 63$ or $n \in 64 - 127$, and then you flip again to see whether the number is in the first or second 32 integers of either interval, and so on. And you will find that you need exactly 8 coin flips to determine the number exactly. Thus, an integer between 0-255 needs 8 bits of information. Here, one bit of information means an answer to one "yes-no" question, and $n$ bits of information means answers to n "yes-no" questions.

Now let us say that we are flipping a special coin with $P(\text{head}) = 9/10$ and $P(\text{tail}) = 1/10$. So before the coin is even flipped, you can already guess that the outcome is most likely to be "head". So the coin flipping actually tells you less information than you would need if the outcomes are equally likely. For instance, if the outcome if "head", then you would say, well, that is what I guessed, and this little information from the coin flip is almost useless except to confirm your guess, or useful to a smaller extent. If the coin flip gives "tail", it surprises you, and hence this information is more useful. More explicitly,

$$\begin{aligned} -\log_2 P(\text{head}) &= -\log_2 9/10 \approx 0.152 \text{ bit} \\ -\log_2 P(\text{tail}) &= -\log_2 1/10 \approx 3.3219 \text{ bit} \end{aligned}$$

So, a outcome of "head" gives you only 0.152 bit of information, but a "tail" gives 3.3219 bits. If you do many coin flips, on average each flip gives you

$$P(\text{head})(-\log_2 P(\text{head})) + P(\text{tail})(-\log_2 P(\text{tail})) = 0.9 \cdot 0.152 + 0.1 \cdot 3.3219 = 0.469 \text{ bit} \tag{2.2}$$

of information, less than the one bit of information if head and tail are are equally likely. More generally, the average amount of information for probability distribution $P(n)$ for variable $n$ is

$$I = -\sum_n P(n) \log_2 P(n) \text{ bits} \tag{2.3}$$

which is more when the distribution $P(n)$ is more evenly distributed, and most in amount when $P(n) =$ constant, i.e., exactly evenly distributed. So if variable $n$ can take $N$ possibilities, the most amount of information is $I = \log_2 N$ bits, hence 8 bits for an integer $n \in [0, 256]$. The formula for information is the same as that for entropy, which we denote by $H(n)$ as the entropy on variable $n$. Hence, a more evenly distributed $P(n)$ means more varibility in $n$, or more randomness, or more ignorance about $n$ before one knows its exact value.

**Information transmission and information channels**

Let a signal $S$ be transmitted via some channel to a destination giving output $O$. The channel can have some noise, and let us assume

$$O = S + N \tag{2.4}$$

So for instance, $S$ can be the input at the sensory receptor, and $O$ can be the output when it is transmitted to a destination neuron. Before you receive $O$, all you have is the expectation that $S$ has a probability distribution $P_S(S)$. So you have

$$H(S) = -\sum_S P_S(S) \log_2 P_S(S) \text{ bits} \tag{2.5}$$

of ignorance or missing information about $S$. Let us say that you also the channel well enough to know the probability distribution $P_N(N)$ for the noise $N$. Then you receive a signal $O$, and you can have a better guess on $S$, as following a probability distribution $P(S|O)$, which is the conditional probability of $S$ given $O$. As you can imagine, $P(S|O)$ must have a narrower distribution than $P_S(S)$. For instance, if you know originally that $S$ is between $-10$ to $10$, and you know that the noise is mostly $N \in -1, 1$, and if you received an $O = 5$, then you can guess that $S \in 4 - 6$. So your guess on $S$ has narrowed down from $(-10, 10)$ to $(4, 6)$. If $S$ can only take on integer values (for instance), you originally had about $\log_2 21$ bits of information missing, now given $O$, you have only about $\log_2 3$ bits of information missing from you. So given output $O$, you can guess what $S$ is to some extent, though not as good as if you received $S$ directly. The amount of information still missing is the conditional entropy

$$H(S|O) \equiv -\sum_S P(S|O) \log_2 P(S|O) \tag{2.6}$$

which should be much small than $-\sum_S P_S(S) \log_2 P_S(S)$. So the amount of information $O$ tells you about $S$ is then

$$H(S) - H(S|O) = (-\sum_S P_S(S) \log_2 P_S(S)) - (-\sum_S P(S|O) \log_2 P(S|O)) \tag{2.7}$$

The first and second terms are the amount of information missing about $S$ before and after, respectively, knowing $O$.

In different trials, you will receive many different output signals $O$. Let signal $O$ to have probability distribution $P_O(O) = \sum_S P_S(S) P_N(O - S)$, assuming that $N$ is independent of $S$. So on average, the information that $O$ contains about $S$ is

$$\begin{aligned} I(O;S) &\equiv \sum_O P_O(O)(H(S) - H(S|O)) \\ &= (-\sum_S P_S(S) \log_2 P_S(S)) - (-\sum_{O,S} P(O,S) \log_2 P(S|O)) \end{aligned}$$

Here $P(O, S)$ is the joint probability distribution of $O$ and $S$. If an information channel transmits $I(O; S)$ bits of information from source $S$ to output $O$ per unit time, then this channel is said to have a capacity of at least $I(O; S)$ bits per unit time.

A particular useful example is when $S$ and $N$ are both gaussian,

$$P(S) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-S^2/(2\sigma_s^2)} \quad P(N) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-N^2/(2\sigma_n^2)} \tag{2.8}$$

with zero means and variances $\sigma_s^2$ and $\sigma_n^2$ respectively. Then, $O$ is also gaussian with zero mean and variance $\sigma_s^2 + \sigma_n^2$. Then $H(S) = \int dS P(S) \log_2 P(S) = \log_2 \sigma_s$ + constant, and the information in $O$ about $S$ is

$$I(O;S) = H(O) - H(S|O) = H(O) - H(N) = \frac{1}{2} \log_2(1 + \frac{\sigma_s^2}{\sigma_n^2}), \tag{2.9}$$

which depends on the signal-to-noise ratio $\sigma_s^2/\sigma_n^2$.

**Mutual information, and information redundancy**

One can say that once $O$ is known, one knows something about $S$. This means $O$ and $S$ share some information, whose amount is exactly $I(O;S)$, which is called mutual information. The difference between $O$ and $S$ is caused by noise, and that is the information not shared between $S$ and $O$. Hence, this mutual information is symmetric between $O$ and $S$,

$$I(O;S) = I(S;O) = \sum_{O,S} P(O,S) \log_2 \frac{P(O,S)}{P(S)P(O)} \tag{2.10}$$

We can use this result in another situation where information is shared between nearby pixels in images. Let $S_1$ and $S_2$ be the image intensities in two nearby pixels of an image. Normally, these two intensities are likely to be similar in most natural images. Hence, if you know $S_1$, you can already guess something about $S_2$. Or, $P(S_2|S_1) \neq P(S_2)$, so $S_1$ and $S_2$ are not independent variables. $P(S_2|S_1)$ usually has a narrower distribution than $P(S_2)$. So we say that information provided by $S_1$ and $S_2$ are somewhat redundant, although information provided by $S_2$ is not exactly the same as that by $S_1$. When there is information redundancy, we have $H(S_1) + H(S_2) > H(S_1, S_2)$, i.e., the summation of the amount of information provided by $S_1$ and $S_2$ separately is larger than the information contained by the two signals together. Then the amount of mutual information between $S_1$ and $S_2$ is

$$I(S_1;S_2) = H(S_1) + H(S_2) - H(S_1,S_2) \tag{2.11}$$

In general, given $N$ signals $S_1$, $S_2$, ..., $S_N$,

$$\sum_i H(S_i) \geq H(S_1, S_2, ..., S_N) \tag{2.12}$$

with equality when all $S$'s are independent or when there is no redundancy.

Redundancy exists in many natural information representation such as natural images or natural languages (represented as a string of letters, and the nearby letters are correlated). When information is represented redundantly, we say that the representation is not efficient. In our example, if $\sum_i H(S_i)$ = 100 bits $> H(S_1, S_2, ..., S_N)$ = 50 bits, it is not efficient to use 100 bits to represent 50 bits of information. Sending the signals $\mathbf{S} \equiv (S_1, S_2, ..., S_N)$ (per unit time) through an information channel in this form would require a channel capacity of at least 100 bits per unit time. Shannon and Weaver (1949) showed that theoretically, all the information (of amount $H(S_1, S_2, ..., S_N)$) about $\mathbf{S} \equiv (S_1, S_2, ..., S_N)$ could be faithfully transmitted through a channel of a capacity of only $H(S_1, S_2, ..., S_N)$ (e.g., 50 bits) per unit time, by encoding $\mathbf{S}$ into some other form $\mathbf{S}' = f(\mathbf{S})$, where $f(.)$ is an (invertable) encoding transform. In such a case, $\mathbf{S}'$ would be a more efficient representation of the original information in $\mathbf{S}$, and the information channel would be more efficiently used.

## 2.1 Formulation of the efficient coding principle

The formulation of the efficient coding principle for early vision goes as follows (Atick 1992). Let sensory input signal $\mathbf{S} \equiv (S_1, S_2, ..., S_M)$ occur with probability $P(\mathbf{S})$. Due to input sampling noise $\mathbf{N}$, the actually received signals in the sensory receptors are

$$\mathbf{S}' = \mathbf{S} + \mathbf{N} \tag{2.13}$$

The amount of sensory information received is thus $I(\mathbf{S}';\mathbf{S})$. For this, the data rate in each channel $i$ is $I(S_i';S_i)$, giving a total input data rate of $\sum_i I(S_i';S_i)$ which is no less than the received information $I(\mathbf{S}';\mathbf{S})$ due to likely information redundancy between different channels $S_i$ and $S_j$.

Let there be an encoding process $\mathsf{K}$ that transforms the input to neural responses (see Fig. (2.1))

$$\mathbf{O} = \mathsf{K}(\mathbf{S}') + \mathbf{N}_o \tag{2.14}$$

where $\mathbf{N}_o$ is the intrinsic output noise not attributable to input noise, and $\mathsf{K}$ can be a linear (kernel) or nonlinear function. For instance, in a blowfly's compound eye, $\mathbf{S}$ is the input contrast, $\mathsf{K}(\mathbf{S})$

describes the sigmoid-like gain control of $\mathbf{S}$ by large monopolar cells (LMC). For another example, $\mathbf{S} = (S_1, S_2, ..., S_M)$ could be a vector describing inputs to $M$ photoreceptors, $\mathbf{O}$ another vector of inputs to many retinal ganglion cells, the receptor-to-ganglion transform maybe approximated linearly as

$$O_i = [\sum_j \mathsf{K}_{ij}(S_j + N_j)] + N_{o,i}, \tag{2.15}$$

where $\mathsf{K}_{ij}$ is the effective neural connection from the $j^{th}$ receptor to the $i^{th}$ ganglion via the retinal interneurons.

This output channel $\mathbf{O}$ thus transmits

$$I(\mathbf{O}; \mathsf{K}(\mathbf{S}')) = H(\mathbf{O}) - H(\mathbf{O}|\mathsf{K}(\mathbf{S}')) = H(\mathbf{O}) - H(\mathbf{N}_o) \tag{2.16}$$

amount of information (where $H$ denotes entropy and $H(.|.)$ conditional entropy). This information is transmitted at a total data rate of $\sum_i I(O_i; (\mathsf{K}(\mathbf{S}')_i) = \sum_i [H(O_i) - H(N_{o,i})]$. However, this transmitted output information contains information about the input noise $\mathbf{N}$ transformed by $\mathsf{K}$ into $\mathsf{K}(\mathbf{S}') - \mathsf{K}(\mathbf{S})$. The total noise in the response $\mathbf{O}$ due to both the transmitted input noise and output intrinsic noise is

$$\text{total output noise} \quad \mathbf{N}^{(o)} \equiv \mathsf{K}(\mathbf{S}') - \mathsf{K}(\mathbf{S}) + \mathbf{N}_o \tag{2.17}$$

Of output information $I(\mathbf{O}; \mathsf{K}(\mathbf{S}'))$, the useful part about the sensory input $\mathbf{S}$ is

$$\text{Information} \quad I(\mathbf{O}; \mathbf{S}) = H(\mathbf{O}) - H(\mathbf{O}|\mathbf{S}) = H(\mathbf{S}) - H(\mathbf{S}|\mathbf{O}) \tag{2.18}$$

where, e.g.,

$$H(\mathbf{O}|\mathbf{S}) = -\int d\mathbf{O} d\mathbf{S} P(\mathbf{O}, \mathbf{S}) \log_2 P(\mathbf{O}|\mathbf{S}) \tag{2.19}$$

where $P(\mathbf{O}|\mathbf{S})$ is the conditional probability of $\mathbf{O}$ given $\mathbf{S}$, and $P(\mathbf{O}, \mathbf{S}) = P(\mathbf{O}|\mathbf{S})P(\mathbf{S})$ is the joint probability distribution. (Note that $P(\mathbf{O}|\mathbf{S})$ depends on the probability distribution $P_N(\mathbf{N})$ and $P_{N_o}(\mathbf{N}_o)$ of the input noise $\mathbf{N}$ and output intrinsic noise $\mathbf{N}_o$). The probability distribution of output $\mathbf{O}$ alone is thus $P(\mathbf{O}) = \int d\mathbf{S} P(\mathbf{O}|\mathbf{S})P(\mathbf{S})$.

The extracted information $I(\mathbf{O}; \mathbf{S})$ at the output $\mathbf{O}$ can not exceed the amount of information $I(\mathbf{S}'; \mathbf{S})$ received at the input stage, i.e., $I(\mathbf{O}; \mathbf{S}) \leq I(\mathbf{S}'; \mathbf{S})$. To make $I(\mathbf{O}; \mathbf{S}) \to I(\mathbf{S}'; \mathbf{S})$, one needs sufficient output channel capacity, or sufficient dynamic range in the output responses, such that each received input $\mathbf{S}'$ can be least ambiguously mapped to an output response level $\mathbf{O}$ (thus no information is lost). For instance, this could be achieved by

$$\mathbf{O} = \text{large scale factor } (\mathbf{S} + \mathbf{N}) + \mathbf{N}_o \tag{2.20}$$

such that the output noise, $\mathbf{N}^{(o)} = \text{large scale factor } \mathbf{N} + \mathbf{N}_o$, is dominated by transmitted input noise. However, this makes the output dynamic range very large, costing a total output channel capacity of $\sum_i [H(O_i) - H(N_{o,i})]$ (in which each $H(O_i)$ increases with the dynamic range for $O_i$). Significant cost can be saved by reducing the information redundancy between the output channels, which is inherited from the redundancy between the input channels. In particular, the amount of redundant information at input stage is

$$\sum_i I(S'_i; S_i) - I(\mathbf{S}'; \mathbf{S}). \tag{2.21}$$

In other words, the input stage uses much more input channel capacity $\sum_i I(S'_i; S_i)$, or receives more data rate, than the input information rate $I(\mathbf{S}'; \mathbf{S})$. For instance, the input information rate $I(\mathbf{S}'; \mathbf{S})$ maybe one megabyte/second, while using a data rate $\sum_i I(S'_i; S_i)$ of 10 megabye/second. Using a suitable encoding $\mathsf{K}$ to remove such redundancy could save the output channel capacity or dynamic range, thus saving neural cost, while still transmitting input as faithfully as possible, i.e., to have $I(\mathbf{O}; \mathbf{S}) \to I(\mathbf{S}'; \mathbf{S})$. In the example above, this means transmitting $I(\mathbf{O}; \mathbf{S})$ at a rate

of nearly one megabyte/second, but using a data rate or channel capacity $\sum_i [H(O_i) - H(N_{o,i})]$ of much less than 10 megabyle/second.

In general, though, removing input redundancy is not always the best strategy, the optimal encoding K should depend on the input statistics such as input signal-to-noise ratio (S/N). When the input has a high signal-to-noise ratio S/N, i.e., the variations in $S_i$ is much larger than that of the input noise, the input data rate $I(S_i'; S_i)$ in each channel is high. In such a case, an encoding K that reduces information redundancy between different input channels $S_i$ and $S_j$, or decorrelates $S_i$ and $S_j$, can reduce the output data rate so that output channels do not require high channel capacity or large dynamic range. In low input S/N regimes, the input data rate is low, input smoothing, which thus introduces or retains correlations, helps avoid unnecessary waste of output channel capacity in transmitting noise. In order words, $I(\mathbf{O}; \mathbf{S})$ should be maximized while minimizing the output cost. These points are elaborated throughout this section.

In general, output entropy

$$H(\mathbf{O}) = I(\mathbf{O}; \mathbf{S}) + H(\mathbf{O}|\mathbf{S}) \tag{2.22}$$

conveys information both about $\mathbf{S}$ by the amount $I(\mathbf{O}; \mathbf{S})$ and about noise by the amount $H(\mathbf{O}|\mathbf{S})$. When the input noise $\mathbf{N} \to 0$,

$$H(\mathbf{O}|\mathbf{S}) \to \text{entropy of the output intrinsic noise } \ H(\mathbf{N}_o) = \text{constant}, \tag{2.23}$$

then maximizing $I(\mathbf{O}; \mathbf{S})$ is equivalent to maximizing $H(\mathbf{O})$ — the maximum entropy encoding method. When the total output data rate, or output channel capacity, $\sum_i H(O_i)$, is fixed, the inequality $H(\mathbf{O}) \leq \sum_i H(O_i)$ implies that $H(\mathbf{O})$ is maximized when the equality $H(\mathbf{O}) = \sum_i H(O_i)$ is achieved. Mathematically, equality $H(\mathbf{O}) = \sum_i H(O_i)$ occurs when different output neurons convey different aspects of the information in $\mathbf{S}$. If one neuron always responds exactly the same as another, information from the second neuron's response is redundant, and the total information conveyed by one neuron is the same as that by both. Thus, $H(\mathbf{O})$ is maximized when neurons respond independently, i.e.,

$$P(O_1, O_2, ...O_N) = P(O_1)P(O_2)...P(O_N), \tag{2.24}$$

the joint probability factorizes into marginal probabilities. Such a coding scheme for $\mathbf{O}$ is said to be an independent component code (or factorial code) of input $\mathbf{S}$. This is why in the noiseless limit, $I(\mathbf{O}; \mathbf{S})$ is maximized when responses $O_i$ and $O_j$ are not correlated. If decorrelating different $O_i$'s does not give sufficient output data rate or entropy $H(\mathbf{O})$, then the individual entropy $H(O_i)$ for each channel could be increased by (1) equalizing the probabilities of different output response levels (sometimes known as the histogram equalization method), and (2) increasing the output dynamic range or number of distinguishable response levels. For instance, if neuron $i$ has only $n = 2$ possible response values $O_i$ (per second), it can transmit no more than $H(O_i) = \log_2 n = 1$ bit/second (when $n = 2$) of information when each response value is utilized equally often, in this case

$$P(O_i = \text{a particular response}) = 1/n \text{for each response values.} \tag{2.25}$$

So $M$ such (decorrelated) neurons can jointly transmit $M$ bits/second when $n = 2$. More information can be transmitted if $n$ is larger, i.e., if the neuron can a larger dynamic range or more response levels.

Typically, natural scene signals $\mathbf{S}$ obey statistical regularities in $P(\mathbf{S})$ with (1) different signal values not occurring equally often, and, (2) different input channels $S_i$ and $S_j$, e.g., responses from neighboring photoreceptors, conveying redundant information. For instance, if two responses from two photoreceptors respectively are very correlated, once one response is known, the second response is largely predictable, and only the difference between it and the first response (or, the non-predictable residual response) conveys additional, non-redundant, information. If $M$ such photoreceptors (input channels) contain $8$ bits/second of information in each channel $i$, $S/N \gg 1$ is good. If, say, 7 out of the 8 bits/second of information in each channel is redundant information already present in other channels, the total amount of joint information $H(\mathbf{S})$ is only about $M$ bits/second (for large $M$), much less than the apparent $8 \times M$ bits/second. Transmitting the

raw input directly to the brain using $\mathbf{O} = \mathbf{S}$ would be inefficient, or even impossible if, e.g., the $M$ output channels $\mathbf{O} = (O_1, O_2, ..., O_M)$ have a limited capacity of only $H(O_i) = 1$ bit/second each. The transform or coding $\mathbf{S} \rightarrow \mathbf{O} \approx \mathsf{K}(\mathbf{S})$ could maximize efficiency such that (1) neurons $O_i$ and $O_j$ respond independently, and (2) each response value of $\mathbf{O}$ is equally utilized. Then, all input information could be faithfully transmitted through responses $\mathbf{O}$ even though each output channel conveys only 1 bits/second. Accordingly, e.g., the connections from the photoreceptors to the retinal ganglion cells are such that, in bright illumination (i.e., when signal-to-noise is high), ganglion cells are tuned to response differences between nearby photoreceptors, making their responses more independent from each other. These ganglion cells are called feature detectors (Barlow 1961) for responding to informative (rather than redundant) image contrast features.

However, when the input $S/N \ll 1$ is so poor that each input channel has no more than, say, 0.1 bit/second of useful information. For instance, for zero mean gaussian signals $S_i' = S_i + N_i$, $I(S_i'; S_i) = 0.1$ bits/second implies a signal-to-noise ratio of $\langle S_i^2 \rangle / \langle N_i^2 \rangle = 0.149$ (where $\langle ... \rangle$ means ensemble average, e.g., $\langle S_i^2 \rangle = \int dS_i P(S_i) S_i^2$). $M$ such channels can transmit a data rate of only $\sum_i I(S_i'; S_i) = 0.1M$ bits/second, and much less in the information rate $I(\mathbf{S}'; \mathbf{S})$ when considering input redundancy. Such a small data rate is sufficient to fit into $M$ output channels of 1 bit/second even without encoding, i.e., even when $\mathbf{O} = \mathbf{S}' + \mathbf{N}_o$ (when output intrinsic noise $\mathbf{N}_o$ is not too large). The output channel capacity $H(\mathbf{O}) = I(\mathbf{O}; \mathbf{S}) + H(\mathbf{O}|\mathbf{S})$ wastes a significant or dominant fraction $H(\mathbf{O}|\mathbf{S})$ on transmitting input noise $\mathbf{N}$ which is typically less redundant between input channels. In fact, most or much of the output variabilities are caused by input noise rather than signal, costing metabolic energy to fire action potentials (Levy and Baxter 1996). To minimize this waste, a different transform $\mathsf{K}$ is desirable to average out input noise. For instance, if input has two channels, with very correlated inputs $S_1 \approx S_2$, but independent and identically distributed (i.i.d) noises $N_1$ and $N_2$. An output channel $O_1 = (S_1' + S_2')/2 \approx S_1 + (N_1 + N_2)/2$ would roughly double the signal-to-noise (of variance) in this output channel compared to that of the input channels. When all output channels carry out some sort of average of various input channels (which are themselves correlated), these different output channels $O_i$ and $O_j$ would be correlated or would carry redundant information. With low input data rate, the output channel capacity (e.g., of M bits/second) is often not fully utilized, and the different output response levels are not equally utilized. These output redundancy, both in correlation between channels and in unequal utilization of response levels of each channel, should help to recover the original signal $\mathbf{S}$.

Hence, efficient coding in different input signal-to-noise conditions require different strategies. It is de-correlation and/or output histogram equalization at high S/N case but smoothing or averaging out noise in the low S/N. Finding the most efficient $\mathsf{K}$ given any S/N level thus results in an optimization problem of minimizing the quantity

$$E(\mathsf{K}) = \text{neural cost } - \lambda \times I(\mathbf{O}; \mathbf{S}), \tag{2.26}$$

where the Lagrange multiplier $\lambda$ balances information extraction $I(\mathbf{O}; \mathbf{S})$ and cost. The optimal code $\mathsf{K}$ is the solution(s) to equation $\partial E(\mathsf{K})/\partial \mathsf{K} = 0$.

The above is an analytical formulation (Srinivasan, Laughlin, Dubs 1982, Linsker 1990, Atick and Redlich 1990, van Hateren 1992) of the efficient coding principle (Barlow 1961), which proposes that early visual processing, in particular the RF transformation, compresses the raw data with minimum loss, such that maximum information $I(\mathbf{O}; \mathbf{S})$ can be transmitted faithfully to higher visual areas despite information bottlenecks such as the optic nerve. The neural cost is often the required output channel capacity $\sum_i H(O_i)$ or the required output power (cf. Levy and Baxter 1996) $\sum_i \langle O_i^2 \rangle$. Importantly, in the noiseless limit, different output neurons of an efficient code carry different independent components in the data, promising cognitive advantages by revealing the underlying perceptual entities, e.g., even objects, responsible for the data. This efficient coding principle is sometimes also termed Infomax (i.e., maximizing $I(\mathbf{O}; \mathbf{S})$), sparse coding (i.e., minimizing $\sum_i H(O_i)$ or $\sum_i \langle O_i^2 \rangle$), independent component analysis, and (in low noise cases) redundancy reduction (Nadal and Parga 1993).

We now apply this principle to understand input sampling by the retinal cells and transformations by the RFs of the retinal and V1 cells. For better illustration, most examples below are

simplified to focus only on the relevant dimension(s), e.g., when focusing on input contrast levels to blowfly's eye, dimensions of space and time are ignored.

## 2.2 Efficient neural sampling in the retina

### 2.2.1 Contrast sampling in a fly's compound eye

In a fly's compound eye, we consider $\mathbf{S}$ to be a scalar value $S$ for the input contrast to the photoreceptor, the encoding transform $\mathsf{K}(.)$ is the contrast response function of the secondary neuron, the large monopolar cell (LMC), receiving inputs from the photoreceptor. The scalar response of LMC is

$$O = \mathsf{K}(S + N) + N_o, \tag{2.27}$$

with input (scalar) photoreceptor noise $N$ and output (scalar) intrinsic noise $N_o$ in the LMC. The encoding $\mathsf{K}(.)$ should be a monotonous function to map larger contrast inputs to larger responses. This monotonous function should be designed such that response $O$ extracts most amount of information $I(O;S)$ about the input while saving neural costs.

Let $S$, $N$, and $N_o$ have independent probability distributions $P(S)$, $P_N(N)$, and $P_{N_o}(N_o)$ respectively. The probability distribution of $S' = S + N$ is then a convolution of that of $P(S)$ and $P_N(N)$

$$P_{S'}(S') = \int dS P_N(S' - S)P(S), \tag{2.28}$$

and the probability distribution of the response $O$ is the result of another convolution

$$P(O) = \int dS' P_{N_o}(O - \mathsf{K}(S'))P_{S'}(S') \tag{2.29}$$

Hence, if $P(S)$ is a unimodal function centered around $\hat{S}$, and $N$ and $N_o$ are all zero mean and unimodally distributed, then $P(O)$ is a unimodal function centered around $\mathsf{K}(\hat{S})$. Let the output $O$ to be constrained within range $[O_{min}, O_{max}]$, the output data rate is

$$I_{out} \equiv I(O;S') = H(O) - H(N) = H(O) + \text{constant} \tag{2.30}$$

Meanwhile, if $S$ and $N$ have standard deviations $\sigma_S$ and $\sigma_N$ respectively, the input data rate is

$$I(S';S) = H(S') - H(N) \approx \frac{1}{2}\log_2[1 + \sigma_S^2/\sigma_N^2] \tag{2.31}$$

where the approximation, valid when the variables are roughly gaussian distributed, serves to give a rough idea that the input data rate depends on the input signal-to-noise $\sigma_S^2/\sigma_N^2$. Note that, the information $I(O;S)$ extracted in $O$ about $S$ satisfies

$$I(O;S) \le I(S';S), \qquad I(O;S) \le I(O;S') = I_{out} \tag{2.32}$$

As we saw in section (2.1), when the input signal-to-noise is large enough, in particular when $I(S';S) \gg I(O;S')$, $I(O;S)$ is maximized when output entropy $H(O)$ or output data rate is maximized. When we are not worried about the output cost, i.e., when $\lambda \to \infty$ in equation (2.26), maximizing $H(O)$ should be implemented by making $O$ should be uniformly distributed over $O \in [O_{min}, O_{max}]$. This can be achieved by a choice of $\mathsf{K}$ such that

$$d\mathsf{K}(S')/dS' \propto P_{S'}(S') \tag{2.33}$$

Then the probability of $\mathsf{K}(S')$ is $P_{\mathsf{K}}(\mathsf{K}(S')) = P(S')(d\mathsf{K}(S')/dS')^{-1} = \text{constant}$, giving

$$P(O) = \int d\mathsf{K}(S') P_{\mathsf{K}}(\mathsf{K}(S'))P_{N_o}(O - \mathsf{K}(S')) = \text{constant}. \tag{2.34}$$

In high input signal-to-noise, $S' \to S$, and $P_{S'}(S') \to P(S)$. Hence, the choice of encoding is thus

$$\begin{aligned} d\mathsf{K}/dS &\propto P(S), \quad \text{or,} \\ \mathsf{K}(S) &\propto \text{cummulative distribution of } P(S), \text{ when input S/N is high enough} \end{aligned} \tag{2.35}$$

The contrast response function in the LMC of the flies has indeed been found to be consistent with this strategy (Laughlin 1981). This strategy, illustrated in Fig. (2.2), makes the number of response levels $O$ allocated to each input interval, matches input density $P(S)$, i.e. $dO/dS \propto P(S)$, so that all output response levels are utilized equally. As long as the input signal-to-noise is high enough, the relationship between $\mathsf{K}(S)$ and input statistics $P(S)$ as expressed in equation (2.35) should not change with the background adaptation or light level — this was observed in Laughlin et al (1987).

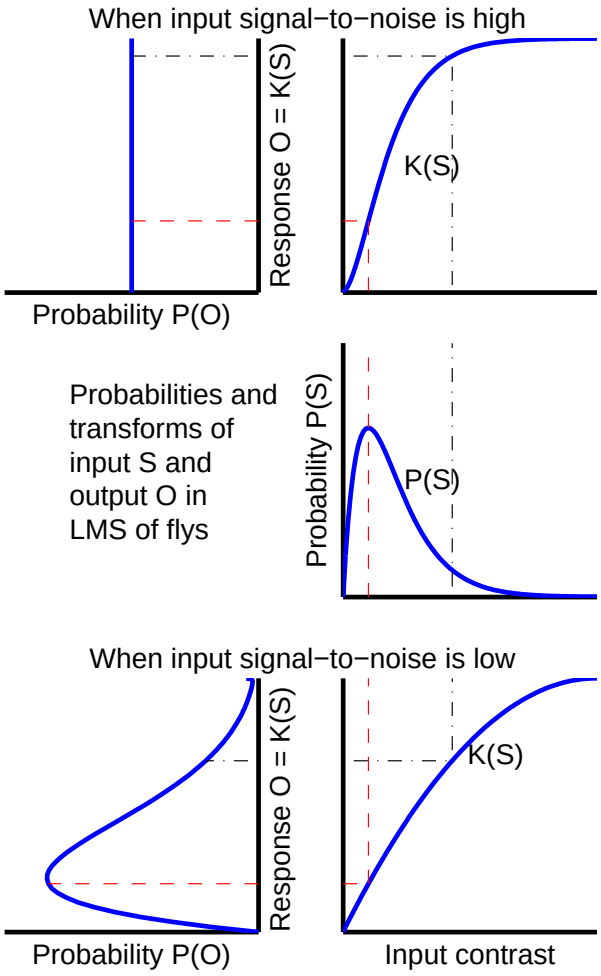



Figure 2.2: Schematic illustration of the probabilities of input contrast encoding in the fly's LMC. The input contrast $S$ has a unimodal distribution $P(S)$ (modelled as in some arbitrary contrast units) (in arbitrary contrast unit). When input S/N is high, the contrast transform $\mathsf{K}(S)$ is such that is closely follows the cummulative distribution of $P(S)$, such that the gain or slope on $\mathsf{K}(S)$ scales with input density $P(S)$, and the output distribution $P(O)$ is uniform, thus equally utilizing all the LMC response levels (after Laughlin 1981). When input S/N is too low, $\mathsf{K}(S)$ (modelled after the data in Laughlin et al 1987) no longer follows the cummulative distribution of $P(S)$, and the $P(O)$ is peaked near zero.

However, when input signal-to-noise is too low, the input data rate $I(S'; S)$ could be lower

than the maximum possible output data rate $I_{out}$ achieved when $O$ is uniformly distributed over the dynamic range. If $O$ continues to be uniformly distributed over its dynamic range, much or most of the response variabilities would convey input noise $N$ rather than the contrast signal $S$. In such a case, to save the output cost, $P(O)$ should not be uniformly distributed, but rather have a distribution peaked near the low response levels (for lower cost). This means that $\mathsf{K}(S)$ should have a lower gain near the typical input levels around $\hat{S}$, so as not to amplify too much input noise. Such a peaked, rather than uniform distribution of $O$ reduces $H(O)$. It is reasonable to expect that an output data rate roughly matching the low input data rate should be adequate

$$I_{out} = H(O) - H(N_o) \approx I(S'; S) \quad \text{when input S/N is low} \tag{2.36}$$

Different optimal contrast encoding transforms $\mathsf{K}$ for different background light levels suggest that the the photoreceptors and/or LMC cells are able to adapt to different statistics in the input ensemble. This is indeed observed experimentally (Laughlin et al 1987) — adaptation can occur within seconds. As we will see later, this is just one example of many in which sensory adaptations are explained or predicted by the efficient coding principle.

### 2.2.2 Spatial sampling by receptor distribution on the retina

When one considers the information about the locations $x$ of the visual objects, it needs to be sampled by photoreceptors receiving inputs at locations $x$.

Analogously, human cones are more densely packed in the fovea, so that their density matches the distribution of the images of relevant objects on the retina (Lewis, Garcia, and Zhaoping 2003), so that the limited resource of $10^7$ cones can be best utilized. Here, input $\mathbf{S}$ is the location of a relevant visual object, output $\mathbf{O}$ is the identity or index of the cone most excited by the object, and $I(\mathbf{O}; \mathbf{S})$ is the amount of information in cone responses about the object's location. Assuming (in a gross simplification) that only the objects of interest are relevant, and that they tend to be brought to the center of vision by eye movements, the distribution $P(\mathbf{S})$ of the object image locations on the retina will indeed peak at the fovea. This distribution, quantitatively derived from the characteristics of the human eye movements, indeed matches the retinal distribution of the human cones reasonably well, although this leaves open whether the eye movement characteristics is the cause or effect of the cone distribution or whether they co-adapt to each other. If the cones were uniformly distributed on the retina, the peripheral ones would be under-utilized, fewer cones would be at the fovea, on average giving less precise information about the object locations.

### 2.2.3 Color sampling by wavelength sensitivities of the cones

Equation (2.26) has also been applied to color sampling by cones at a single spatial location. Here, the input is the visual surface color reflectance $\mathbf{S} = S(l)$ as a function of light wavelength $l$, and the outputs $\mathbf{O} = (O_r, O_g, O_b)$ model responses from red (r), green (g), and blue (b) cones of wavelength sensitivity $R(l - l_i)$ for $i = r, g, b$ with peak sensitivity occurring at optimal wavelength $l_i$. Given sensory noise $N_i$ and illumination $E(l)$ from sun light, $O_i = \int dl R(l - l_i) S(l) E(l) + N_i$. Just as the contrast response function $\mathsf{K}(S)$ of the fly's LMC can be optimized to maximize information extraction, the color sensitivities can be similarly optimized by the choice of $l_i$, an operation that largely explains the cones' sensitivities in humans (Lewis and Zhaoping 2006). This makes responses from different cones (particularly red and green) suitably correlated with each other, to smooth out the often substantial noise in dim light and/or under fine spatial resolution.

## 2.3 Efficient coding by early visual receptive fields

The efficient coding principle has been much more extensively applied to understand the RF transforms of the receptor responses by retinal ganglion cells (or LGN cells) and V1 neurons. Now

we denote the receptor outputs by $\mathbf{S}+\mathbf{N}$, including both signal $\mathbf{S}$ and noise $\mathbf{N}$, and post-synaptic responses by $\mathbf{O}$. The problem is simplified by approximating the neural transforms as linear

$$\mathbf{O} = \mathsf{K}(\mathbf{S}+\mathbf{N}) + \mathbf{N}_o, \quad \text{or, in component form,} \quad O_i = \sum_j \mathsf{K}_{ij}(S_j+N_j) + N_{o,i} \tag{2.37}$$

where $\mathbf{N_o}$ is the neural noise introduced by the transform, so $\mathsf{K}\mathbf{N}+\mathbf{N}_o$ is the total noise (originally denoted by symbol $\mathbf{N}$). As discussed earlier, whether the optimal RF transform $\mathsf{K}$ decorrelates inputs or not depends on the input S/N level. To focus on such RF transforms as combining the original $\mathbf{S}$ channels, I omit nonlinear gain control processes such as those in the LMC of blowflies (Nadal and Parga 1994).

Optimizing $\mathsf{K}$ accurately requires precise information about $P(\mathbf{S})$, i.e., a joint probability distribution on M pixel values $(S_1, S_2, ..., S_M)$. Unfortunately, this is not available for large M. However, given the second order correlation $R^S_{ij} \equiv \langle S_i S_j \rangle$ between inputs, a maximum entropy approximation of $P(\mathbf{S})$ is a Gaussian distribution $P(\mathbf{S}) \propto \exp[-\sum_{ij} S_i S_j (R^S)^{-1}_{ij}/2]$, where $(R^S)^{-1}$ is the inverse matrix of matrix $R^S$ (with elements $R^S_{ij}$) and the signals are simplified as (or pre-transformed to) having zero mean. This approximation has the advantage of enabling analytical solutions of the optimal $\mathsf{K}$ (Linsker 1990, Atick and Redlich 1990, Atick et al 1992, Dong and Atick 1995, Li and Atick 1994ab, Li 1996), and captures well our ignorance of the higher order statistics.

Alternatively, one can sample natural scene statistics and obtain $\mathsf{K}$ by simulation algorithms, e.g., through gradient descent in the $\mathsf{K}$ space to minimize $E(\mathsf{K})$. Bell and Sejnowski (1997) did this, finding V1 RFs by maximizing $H(\mathbf{O})$ (corresponding to the noiseless limit $\mathbf{N} \to 0$ when $I(\mathbf{O};\mathbf{S}) = H(\mathbf{O})$ + constant), with neural cost constrained to a fixed output dynamic range. Note that once $\mathbf{O}$ is obtained, $\mathbf{S}$ can be reconstructed by $\mathbf{S} = \mathsf{K}^{-1}\mathbf{O}$ + noise when $\mathsf{K}$ is invertible (i.e., when $\mathbf{O}$ is a complete or over-complete representation). While input reconstruction is not the goal of efficient coding, it is worth noting the link between efficient coding and another line of works often referred to as sparse coding, also aimed to understand early visual processing (Olshausen and Field 1997, van Hateren and Ruderman 1998, Simoncelli and Olshausen 2001). These works proposed that visual input $\mathbf{S}$ with input distributions $P(\mathbf{S})$ can be generated as a weighted sum of a set of basis function, weighted by components $O_1, O_2, ...$ of $\mathbf{O}$ with sparse distributions $P(O_i)$ for all $i$. Thus, the column vectors of $\mathsf{K}^{-1}$ correspond to the basis functions. Since larger $I(\mathbf{O};\mathbf{S})$ enables better generation of $\mathbf{S}$ from $\mathbf{O}$, and since sparseness for $\mathbf{O}$ is equivalent to constraining the neural cost as entropies $\sum_i H(O_i)$, such sparse coding formulation is an alternative formulation of the efficient coding principle. Indeed, in practice, their typical algorithms find $\mathbf{O}$ and $\mathsf{K}^{-1}$ by minimizing an objective function $\mathcal{E} = \langle (\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2 \rangle + \lambda \sum_i \mathrm{Sp}(O_i)$ where $\mathrm{Sp}(O_i)$, e.g., $\mathrm{Sp}(O_i) = |O_i|$, describes a cost of non-sparseness (which encourages a sharply peaked distribution $P(O_i)$ and thus low $H(O_i)$), while the reconstruction error $\langle (\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2 \rangle$ should roughly scale with $2^{-2I(\mathbf{O};\mathbf{S})}$. It is thus not surprising that these algorithms (e.g., Olshausen and Field 1997), which were mostly simulated for low noise cases, produce results similar to those by simulation algorithms (e.g., Bell and Sejnowski 1997) for efficient coding to minimize $E(\mathsf{K})$ of equation (2.26), also in the noiseless limit. All these simulational algorithms have the advantage of being performed online while being exposed to individual natural images $\mathbf{S}$, thus all orders of statistics in $P(\mathbf{S})$ are absorbed by the algorithms without having to approximate $P(\mathbf{S})$. Importantly, their results (e.g., Bell and Sejnowski 1997, Olshausen and Field 1997, van Hateren and Ruderman 1998, Simoncelli and Olshausen 2001) confirmed the previous analytical results on $\mathsf{K}$, particularly of V1 RFs (Li and Atick 1994ab, Li 1996), obtained by approximating $P(\mathbf{S})$ by up to second order statistics only.

In general, inputs $\mathbf{S} = S(x, t, e, c)$ depend on space $x$, time $t$, eye origin $e$, and input cone type $c$. The RF transform for a V1 cell, for instance, can reflect selectivities to all these input dimensions, so that a cell can be tuned to orientation (involving only $x$), motion direction (involving $x, t$), spatial scale ($x$), eye origin ($e$), color ($c$), and depth ($x, e$) or combinations of them. I will review the findings in the efficient coding formulation as in equations (2.26) and (2.37) using Gaussian approximation for $P(\mathbf{S})$ (also with all noise assumed to be Gaussian and independent), to take advantage of the analytical convenience and insight, and of the flexibility to handle different signal-to-noise levels. The analytical approach also avoids tampering with translation and scale invariance in input statis-

tics (something which is hard to avoid in simulation studies when images of, say, 12x12 pixels are used) which can bias the scales and shapes of the RFs found.

### 2.3.1 The general solution to efficient codings of gaussian signals

Here I briefly summarize the analytical results on the optimal encoding $\mathsf{K}$ when the neural cost is $\sum_i \langle O_i^2 \rangle$ and when all signals and noises are assumed as gaussian. Readers not interested in mathematical details may skip this summary which is not essential for understanding when reading on. For simplicity, it is assumed that the input noise $N_i$ and the intrinsic output noise $N_{o,i}$, in different input/output channels, are independent and identically distributed with their respective noise powers $N^2 = \langle N_i^2 \rangle$ and $N_o^2 = \langle N_{o,i}^2 \rangle$. The neural cost is

$$\sum_i \langle O_i^2 \rangle = \mathrm{Tr}(R^O) = \mathrm{Tr}[\mathsf{K}(R^S + N^2)\mathsf{K}^T + N_o^2]$$

where $R^O$ is the output correlation matrix with elements $R_{ij}^O = \langle O_i O_j \rangle$ and Tr(.) denotes the trace of a matrix. The extracted information at the output is

$$I(\mathbf{O};\mathbf{S}) = \frac{1}{2}\log_2 \frac{\det R^O}{\det R^{No}}$$

where $R^{No}$ is the correlation matrix of the output noise $\mathsf{K}\mathbf{N} + \mathbf{N}_o$ which is composed of the intrinsic output noise $\mathbf{N}_o$ and the input noise relayed through $\mathsf{K}$. Thus, $R_{ij}^{No} = \langle (\mathsf{K}\mathbf{N}+\mathbf{N}_o)_i (\mathsf{K}\mathbf{N}+\mathbf{N}_o)_j \rangle$. It is now clear that $\mathrm{Tr}(R^O)$, $\det(R^O)$, and $\det(R^{No})$ are all invariant to a change of the encoding matrix $\mathsf{K} \to \mathsf{U}\mathsf{K}$ by any unitary matrix $\mathsf{U}$, which satisfies $\mathsf{U}\mathsf{U}^\dagger = 1$. In other words, the optimal encoding solutions $\mathsf{K}$ to minimize

$$E(\mathsf{K}) = \text{cost} - \lambda I(\mathbf{O};\mathbf{S}) = \mathrm{Tr}(R^O) - \frac{\lambda}{2}\log_2 \frac{\det R^O}{\det R^{No}}$$

are degenerate by the $\mathsf{U}$ transform symmetry. Let $R^S$ have eigenvectors $V^1, V^2, ..., V^k, ...$, and let the projection of $\mathbf{S}$ on these vectors be $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_k, ....$ A special optimal coding solution $\mathsf{K}$ within the degenerate class of solutions is such that $\mathsf{K}R^S\mathsf{K}^T$ is diagonal, and

$\mathsf{K}$ = $\mathsf{g}\mathsf{K}_o$, such that

- $\mathsf{K}_o$ is an unitary matrix whose row vectors are (the complex conjugate of the) eigenvectors $V^1, V^2, ..., V^k, ...$ $(\mathsf{K}_o R^S \mathsf{K}_o^T)_{ij} = \lambda_i \delta_{ij}$, where $\lambda_i$ for $i = 1, 2, ...$ are the eigenvalues of $R^S$,
- $\mathsf{g}$ is a diagonal matrix whose diagonal elements are gain factors $\mathsf{g}_{kk} = g_k$,

Then, the output channels are $O_k = g_k(\mathcal{S}_k + \mathcal{N}_k) + N_{o,k}$, where $\mathcal{N}_k$ is the projection of input noise $\mathbf{N}$ on the eigenvector $V^k$. Note that $R_{ij}^O = \langle O_i O_j \rangle \propto \delta_{ij}$. The objective of minimization is then

$$E(\mathsf{K}) = \sum_k E(g_k), \quad \text{where} \quad E(g_k) = \langle O_k^2 \rangle - \lambda I(O_k; S_k) \tag{2.38}$$

$$\langle O_k^2 \rangle = g_k^2(\langle \mathcal{S}_k^2 \rangle + N^2) + N_o^2 \tag{2.39}$$

$$I(O_k; S_k) = \frac{1}{2}\log_2 \frac{g_k^2(\langle \mathcal{S}_k^2 \rangle + \langle N^2 \rangle) + \langle N_o^2 \rangle}{g_k^2 \langle N^2 \rangle + \langle N_o^2 \rangle} \tag{2.40}$$

Minimizing $E(\mathsf{K})$ is then minimizing each individual $E(g_k)$ by finding the optimal gain $g_k$,

$$g_k^2 \propto \mathrm{Max}\left\{ \left[ \frac{1}{2(1 + \langle N^2 \rangle / \langle S_k^2 \rangle)} \left( 1 + \sqrt{1 + \frac{4\lambda}{(\ln 2)\langle N_o^2 \rangle} \frac{\langle N^2 \rangle}{\langle S_k^2 \rangle}} \right) - 1 \right], 0 \right\} \tag{2.41}$$

which, given $\langle N_o^2 \rangle$, depends only on the signal-to-noise (S/N) ratio $\langle \mathcal{S}_k^2 \rangle / \langle N^2 \rangle$. Hence, in the gaussian approximation of the signals, the optimal encoding transform in general $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, under

neural cost $\sum_i \langle O_i^2 \rangle$, can be decomposed into three conceptual components: (1) principal component decomposition of inputs by the unitary matrix $\mathsf{K}_o$ that diagonalizes $R^S$, (2) gain control $g_k$ of each principal component $\mathcal{S}_k$ according to its S/N, and (3) multiplexing the resulting components by another unitary matrix U. This is illustrated in Fig (2.3). Coding in space, stereo, time, color, at different S/N levels simply differ by input statistics $P(\mathbf{S})$ (i.e., differ by pair-wise signal correlations $R^S$ in the Gaussian approximation) and S/N, but will lead to a diversity of transforms K like the RFs observed physiologically.

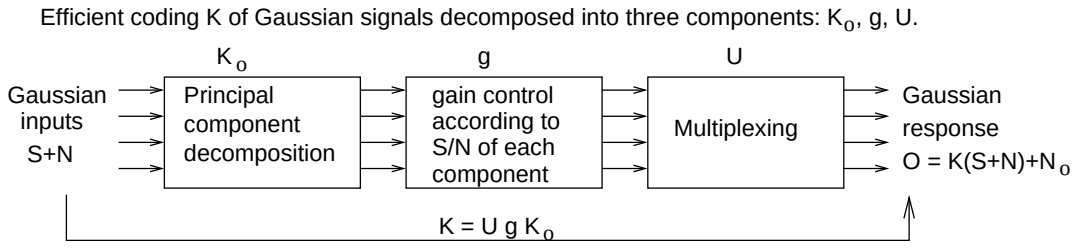



Figure 2.3: Three conceptual components, $\mathsf{K}_o$, g, and U, in the efficient coding $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$ of Gaussian signals.

## 2.4 Illustration: stereo coding in V1

For illustration (Fig. (2.4)), we focus first only on the input dimension of eye origin, $e = L, R$, for left and right eyes with 2-dimensional input signal $\mathbf{S} = (S_L, S_R)$. The coding K is then

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \begin{pmatrix} \mathsf{K}_{1L} & \mathsf{K}_{1R} \\ \mathsf{K}_{2L} & \mathsf{K}_{2R} \end{pmatrix} \begin{pmatrix} S_L + N_L \\ S_R + N_R \end{pmatrix} + \begin{pmatrix} N_{o,1} \\ N_{o,2} \end{pmatrix}$$

This coding transform is linear, approximating the effective transform by the receptive fields of the neurons in the primary visual cortex whose responses modelled as $(O_1, O_2)$. So one would expect that a cortical neuron $i = 1, 2$ in general responds to input from the left and right eyes by different sensitivities specified by $\mathsf{K}_{iL}$ and $\mathsf{K}_{iR}$. The single abstract step to find an optimal coding K by solving $\partial E/\partial \mathsf{K} = 0$ is decomposed into several conceptual steps here for didactic convenience. The signals $\mathbf{S} = (S_L, S_R)$ may be the pixel values at a particular location, average image luminances, or the Fourier components (at a particular frequency) of the images. For simplicity, assume that they have zero means and equal variance (or power) $\langle S_L^2 \rangle = \langle S_R^2 \rangle$. Binocular input redundancy is evident in the correlation matrix:

$$R^S \equiv \begin{pmatrix} \langle S_L^2 \rangle & \langle S_L S_R \rangle \\ \langle S_R S_L \rangle & \langle S_R^2 \rangle \end{pmatrix} \equiv \langle S_L^2 \rangle \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$$

where $0 \leq r \leq 1$ is the correlation coefficient. The correlation is manifested in the shape of the probability distribution of $(S_L, S_R)$ in the input space shown in Fig. (2.5). In that distribution, each sample data point $(S_L, S_R)$ is such that $S_L$ and $S_R$ tend to be similar, and the distribution is shaped like an elipse whose major and minor axes are not along the coordinate directions. The input distribution in the Gaussian approximation is then

$$P(\mathbf{S}) = P(S_L, S_R) \propto \exp[-(S_L^2 + S_R^2 - 2r S_L S_R)/(2\sigma^2)]$$

where $\sigma^2 = \langle S_L^2 \rangle (1 - r^2)$.

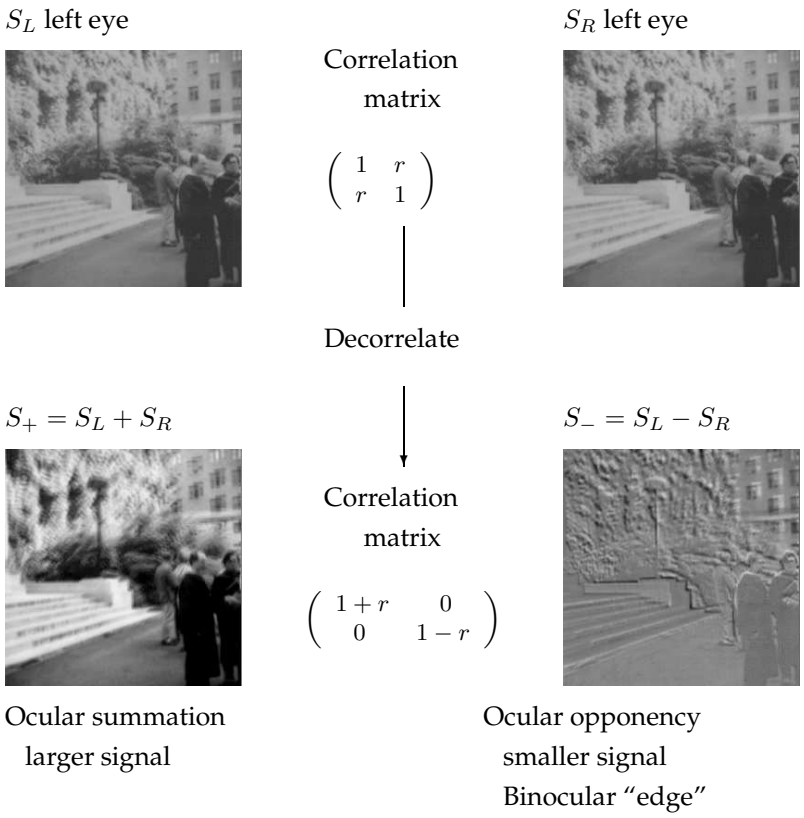



Figure 2.4: Efficient coding illustrated by stereo coding. Left: correlated inputs $(S_L, S_R)$ from the two eyes are transformed to two decorrelated (by second-order) signals $S_\pm \propto S_L \pm S_R$, ocular summation and opponency, of different powers $\langle S_+^2 \rangle > \langle S_-^2 \rangle$.

### 2.4.1 Principal component analysis

The eigenvectors of the correlation matrix $R^S$ are

$$\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

with eigenvalues $\langle S_L^2 \rangle (1 \pm r)$ respectively. In other words, the principal components of the original signal $(S_L, S_R)$ are

$$\begin{pmatrix} S_+ \\ S_- \end{pmatrix} \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} S_L \\ S_R \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} S_L + S_R \\ S_L - S_R \end{pmatrix}$$

and their signal powers are respectively $\langle S_\pm^2 \rangle = \langle S_L^2 \rangle (1 \pm r)$. The ocular summation signal $S_+$ is stronger and conveys information about the 2-dimensional images, whereas the weaker signal $S_-$ conveys ocular contrast ("edge") or depth information (Fig. (2.4). We note that these components $S_+$ and $S_-$ are not correlated:

$$\langle S_+ S_- \rangle \propto \langle (S_L + S_R)(S_L - S_R) \rangle = \langle S_L^2 \rangle - \langle S_R^2 \rangle - \langle S_L S_R \rangle + \langle S_R S_L \rangle = 0$$

Their probability distribution is now factorized into components

$$P(\mathbf{S}) = P(S_+) P(S_-) \propto \exp[-S_+^2 / (2\langle S_+^2 \rangle)] \exp[-S_-^2 / (2\langle S_-^2 \rangle)]$$

The transform $(S_L, S_R)^T \rightarrow (S_+, S_-)^T \equiv \mathsf{U}_0 (S_+, S_-)^T$ is merely a $45^o$ rotation of the coordinates by a rotational matrix $\mathsf{U}_0$ in the two-dimensional space of the input signal, as indicated in Fig. (2.5).

The directions for $S_+$ and $S_-$ in the input signal space are exactly the major and minor axes of probability distribution of input signals. As with any coordinate rotation, $\mathsf{U}_0$ preserves the total signal power $\sum_{i=L,R}\langle S_i^2\rangle = \sum_{i=+,-}\langle S_i^2\rangle$, as easily verified.

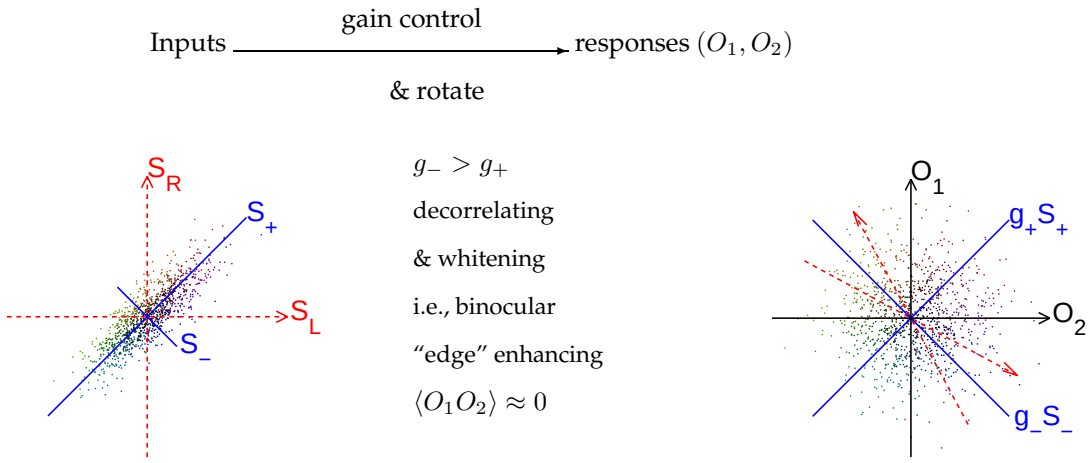



Figure 2.5: Schematics of data $\mathbf{S}$ (in the noiseless condition) and their transforms to responses $\mathbf{O}$ by efficient coding. Each dot is a sample data from distributions $P(\mathbf{S})$ or $P(\mathbf{O})$ in the two dimensional space of $\mathbf{S}$ or $\mathbf{O}$. Correlation $\langle S_L S_R\rangle > 0$ is manifested in the elliptical shape of the data distribution (particularly in the high S/N condition). Gain control, $S_\pm \to g_\pm S_\pm$, produces, under high or low input S/N, decorrelated or correlated responses $(O_1, O_2)$. When S/N$\to\infty$, the weaker signal $S_-$ is relatively amplified for (ocular) contrast or edge enhancement, $g_- > g_+$, leading to whitening or equal power responses $g_+^2\langle S_+^2\rangle \approx g_-^2\langle S_-^2\rangle$. Both $O_1$ and $O_2$ are excited by input from one eye (left and right respectively) and inhibited by input from another.

With sensory noise $\mathbf{N} = (N_L, N_R)$, the input signals become $O_{L,R} = S_{L,R} + N_{L,R}$. The rotational transform simply gives

$$O_\pm = S_\pm + N_\pm$$

where $N_\pm \equiv (N_L \pm N_R)/\sqrt{2}$. Assuming that $N_L$ and $N_R$ are independent and identically distributed (IID), then so are $N_+$ and $N_-$. So $O_+$ and $O_-$ are also decorrelated, $\langle O_+ O_-\rangle = 0$. with factorized probability distribution $P(\mathbf{O}) = P(O_+)P(O_-) \propto \exp[-\frac{1}{2}O_+^2/\langle O_+^2\rangle - \frac{1}{2}O_-^2/\langle O_-^2\rangle]$. Note that $\langle O_i^2\rangle = \langle S_i^2\rangle + \langle N_i^2\rangle$, and, denoting $\langle N^2\rangle \equiv \langle N_R^2\rangle = \langle N_L^2\rangle$, then $\langle N_i^2\rangle = \langle N^2\rangle$ for all $i = L, R, +, -$.

Since the transform $(O_L, O_R) \to (O_+, O_-)$ is merely a coordinate rotation, $(O_L, O_R)$ and $(O_+, O_-)$ consume the same amount of total output power $\langle O_+^2\rangle + \langle O_-^2\rangle = \langle O_L^2\rangle + \langle O_R^2\rangle$, and contain the same amount of information $I(\mathbf{O};\mathbf{S})$ about input signal $\mathbf{S}$. The cortical cell that receives $O_+$ is a binocular cell, summing inputs from both eyes, while the cell receiving $O_-$ is ocularly opponent or unbalanced. It is known that for Gaussian signals, the information in each channel $O_i = S_i + N_i$ about the original signal $S_i$, for $i = L$, $R$, $+$, or $-$, is

$$I(O_i; S_i) = \frac{1}{2}\log_2\frac{\langle O_i^2\rangle}{\langle N_i^2\rangle} = \frac{1}{2}\log_2\frac{\langle S_i^2\rangle + \langle N_i^2\rangle}{\langle N_i^2\rangle} = \frac{1}{2}\log_2[1 + \frac{\langle S_i^2\rangle}{\langle N_i^2\rangle}], \tag{2.42}$$

depends only on the signal-to-noise $\langle S_i^2\rangle/\langle N_i^2\rangle$. Non-zero correlation $\langle S_L S_R\rangle$ gives non-zero $\langle O_L O_R\rangle$, and means that some of the information in $O_L$ and $O_R$ (quantified in bits by $I(O_L; S_L)$ and $I(O_R; S_R)$) about the original signal $\mathbf{S}$ is redundant. In contrast, since $O_+$ and $O_-$ are independent, information in $O_+$ and $O_-$ (quantified by $I(O_+; S_+)$ and $I(O_-; S_-)$) is non-redundant. Hence the total information

$$I(\mathbf{O};\mathbf{S}) = I(O_+; S_+) + I(O_-; S_-) < I(O_L; S_L) + I(O_R; S_R). \tag{2.43}$$

For example, let $\langle S_L^2\rangle/\langle N^2\rangle = \langle S_R^2\rangle/\langle N^2\rangle = 10$, i.e., the original signal-to-noise power in input

channels $O_{L,R}$ is 10, and let the binocular correlation be $r = 0.9$. Then

$$\begin{aligned} I(O_\pm; S_\pm) &= \frac{1}{2}\log_2[1 + (1 \pm r)\langle S_L^2\rangle/\langle N^2\rangle] = 2.16 \text{ or } 0.5 \text{ bits for } O_+ \text{ or } O_- \text{ channels respectively;} \\ I(O_{L,R}; S_{L,R}) &= \frac{1}{2}\log_2[1 + \langle S_L^2\rangle/\langle N^2\rangle] = 1.73 \text{ bits for both } O_L \text{ and } O_R \text{ channels.} \end{aligned}$$

This means channels $O_\pm$ would require a total information channel capacity of $I(O_+; S_+)+I(O_-; S_-) = 2.66$ bits, less than that $I(O_L; S_L) + I(O_R; S_R) = 3.46$ bits required by channels $O_{L,R}$. Meanwhile, $O_\pm$ and $O_{L,R}$ transmit exactly the same information about the original signal $\mathbf{S}$, since knowing either $O_\pm$ or $O_{L,R}$ gives the same conditional probability distribution $P(\mathbf{S}|\mathbf{O})$ about $\mathbf{S}$, weather $\mathbf{O}$ is represented by $O_\pm$ or $O_{L,R}$. In other words, knowing $O_\pm$ or $O_{L,R}$ enables us to recover original signal $\mathbf{S}$ to exactly the same precision. Hence, we say that the coding $O_\pm$ is more efficient than $O_{L,R}$, since it requires less total information channel capacity.

The quantity

$$[\sum_{i=L,R} I(O_i; S_i)]/I(\mathbf{O}; \mathbf{S}) - 1$$

measures the degree of redundancy in the code $\mathbf{O} = (O_L, O_R)$. It is this redundancy that causes unequal signal powers $\langle O_+^2\rangle > \langle O_-^2\rangle$ and information rates $I(O_+; S_+) > I(O_-; S_-)$.

### 2.4.2 Gain control

If $\langle O_\pm^2\rangle$ is the coding cost, the information,

$$I_\pm \equiv I(O_\pm; S_\pm) = \frac{1}{2}\log_2(\langle O_\pm^2\rangle) + \text{constant},$$

increases logarithmically with the cost. Hence, spending any extra power budget gives a better return in the weaker $O_-$ than the stronger $O_+$ channel. This motivates awarding different gains $g_\pm$ to the two channels, $O_\pm \to g_\pm O_\pm$ with $g_+ < g_-$ to amplify the ocular "edge" channel $S_- + N_-$ relatively, provided that this does not amplify input noise $N_-$ too much. In reality, the coding transform $O_{L,R} \to O_\pm$ brings additional noise $\mathbf{N_o}$, hence

$$O_\pm = g_\pm(S_\pm + N_\pm) + N_{o,\pm} = g_\pm S_\pm + g_\pm N_\pm + N_{o,\pm}$$

This gives output signal $g_\pm S_\pm$, and the output noise becomes $N_\pm = g_\pm(N_L \pm N_R)/\sqrt{2} + N_{o,\pm}$. Assuming $\langle N_o^2\rangle \equiv \langle N_{o,+}^2\rangle = \langle N_{o,-}^2\rangle$, for simplicity, the output power is now

$$\langle O_\pm^2\rangle = g_\pm^2(\langle S_\pm^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle$$

and the information is

$$I(O_\pm; S_\pm) = I_\pm = \frac{1}{2}\log_2\frac{\langle O_\pm^2\rangle}{\langle N_\pm^2\rangle} = \frac{1}{2}\log_2\frac{g_\pm^2(\langle S_\pm^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_\pm^2\langle N^2\rangle + \langle N_o^2\rangle} < \frac{1}{2}\log_2[1 + \frac{\langle S_\pm^2\rangle}{\langle N^2\rangle}] \quad (2.44)$$

Hence, the extra coding noise $\mathbf{N}_o$ reduces the amount of information in the channel, as expected. Increasing the gain $g_\pm \to \infty$ makes the output noise $\mathbf{N}_o$ negligible and the output information $I_\pm \to \frac{1}{2}\log_2(1 + \langle S_\pm^2\rangle/\langle N^2\rangle)$ approach the original value. However, this would cost a lot of output power. Balancing need to reduce the output power against that for information preservation, the optimal encoding is thus to find the gains $g_\pm$ that minimize

$$E(g_+, g_-) = \sum_{k=+,-} E(g_k) \equiv \sum_{k=+,-} [\langle O_k^2\rangle - \lambda I_k] = \text{cost} - \lambda \cdot I(\mathbf{O}; \mathbf{S}) \quad (2.45)$$

Note that this objective quantity $E$ to be minimized, expressed as a function of the gains $g_\pm$, is the same as that in the equation (2.26) for the optimal efficient coding, where $E(\mathsf{K})$ was expressed as a

function of the encoding transform K. As will be clear later, the gains $g_\pm$ to the independent component channels $O_\pm$ are some essential parameters in characterizing the full encoding transform K. Since, for each $k = +, -$,

$$E(g_k) = g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle - \frac{\lambda}{2}\log_2 \frac{g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle}$$

The optimal gain $g_k$ can be obtained by $\partial E/\partial g_k = 0$, as (the same as in equation (2.41))

$$g_k^2 \propto \text{Max}\left\{\left[\frac{1}{2(1+\langle N^2\rangle/\langle S_k^2\rangle)}\left(1+\sqrt{1+\frac{4\lambda}{(\ln 2)\langle N_o^2\rangle}\frac{\langle N^2\rangle}{\langle S_k^2\rangle}}\right)-1\right], 0\right\} \tag{2.46}$$

Hence, this optimal gain depends on the signal-to-noise (S/N) ratio $\langle S_k^2\rangle/\langle N^2\rangle$. This dependence is qualitatively different for high and low S/N regions:

$$g_k^2 \propto \begin{cases} \langle S_k^2\rangle^{-1}, & \text{decrease with } \langle S_k^2\rangle \text{ if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \gg 1 \\ \text{Max}\{\alpha\langle S_k^2\rangle^{1/2} - 1, 0\}, & \text{increase with } \langle S_k^2\rangle \text{ if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \ll 1, (\alpha \text{ is a constant}) \end{cases} \tag{2.47}$$

such that $g_k$ decreases and increases with S/N at high and low S/N regions respectively.

### 2.4.3 Contrast enhancement, decorrelation, and whitening in the high S/N region

We first analyze the situation in the high S/N limit when $g_k^2 \propto \langle S_k^2\rangle^{-1}$. As expected, this suppresses the stronger ocular summation signal $S_+$ relative to the weaker ocular contrast signal $S_-$, to reduce cost. Since $g_- > g_+$, this encoding emphasizes the binocular difference, or contrast, or edge channel $S_-$ relative to the ocular summation channel $S_+$ which conveys the common aspects of inputs to the two eyes. Such a relationship in the relative gains is thus performing contrast enhancement. With negligible coding noise $\mathbf{N_o}$ (i.e., $\frac{\langle N_o^2\rangle}{g_\pm^2\langle N^2\rangle} \ll 1$), output $\mathbf{O}$ and the original input $\mathbf{S}+\mathbf{N}$ contain about the same amount of information about the true signal $\mathbf{S}$, but $\mathbf{O}$ consumes much less power with $g_+ \ll g_- \leq 1$, when input ocular correlation $r \sim 1$. For example, when $\langle S_-^2\rangle/\langle N^2\rangle = 7$, $\langle S_+^2\rangle/\langle N^2\rangle = 127$, $r = 60/77$, the total information at input is

$$I(\mathbf{S}+\mathbf{N};\mathbf{S}) = \frac{1}{2}\sum_{+,-}\log_2(1+\langle S_i^2\rangle/\langle N^2\rangle) = 5\text{bits}$$

Assuming that the encoding noise $\mathbf{N}_o$ is negligible compared to input noise $\mathbf{N}$, i.e., $\langle N_o^2\rangle \ll \langle N^2\rangle$, sending the signals $S_\pm + N_\pm$ directly to the brain without gain control would cost output power $\langle S_+^2\rangle + \langle S_-^2\rangle + 2\langle N^2\rangle = 136$ when $\langle N^2\rangle = 1$. When we have $g_- = 1$, and reduce the gain $g_+$ to the stronger channel $S_+$ according to the optimal prescription $g_+ = g_-\sqrt{\langle S_-^2\rangle/\langle S_+^2\rangle} = \sqrt{7/127}$, the total information at the output is (assuming $\langle N_o^2\rangle/\langle N^2\rangle = 0.1$)

$$\sum_{k=+,-} I_k = \frac{1}{2}\sum_{k=+,-}\log_2\frac{g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle} = 4.2 \text{ bits}$$

which is a 15% loss of information due to the intrinsic noise $\mathbf{N}_o$ during the encoding process. But at the total output power $\sum_{k=+,-} g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle = 15.3$, this is a reduction of 89% in output power cost.

This gain $g_\pm \propto \langle S_\pm^2\rangle^{-1/2}$ also equalizes output power $\langle O_+^2\rangle \approx \langle O_-^2\rangle$, since $\langle O_\pm^2\rangle = g_\pm^2\langle S_\pm^2\rangle$ + noise power. Since $\langle O_+O_-\rangle = 0$, the output correlation matrix $R^O$, with elements

$$R_{ab}^O = \langle O_aO_b\rangle = \delta_{ab}\cdot \text{ constant},$$

is now proportional to an identity matrix. Such a transform $\mathbf{S} \to \mathbf{O}$, which leaves output channels decorrelated and with equal power, is called whitening (i.e., the output signals are like white noise which has channels that are independent and identically distributed). Now the two output channels $O_+$ and $O_-$ are equally and non-redundantly utilized.

### 2.4.4 Degeneracy of optimal encoding

Any coordinate rotation $\mathbf{O} \rightarrow \mathsf{U}\mathbf{O}$ by angle $\theta$ in the two dimensional space $\mathbf{O}$, multiplexes the channels $O_+$ and $O_-$ to give two alternative channels

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \mathsf{U} \begin{pmatrix} O_+ \\ O_- \end{pmatrix} \equiv \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} O_+ \\ O_- \end{pmatrix} = \begin{pmatrix} \cos(\theta)O_+ + \sin(\theta)O_- \\ -\sin(\theta)O_+ + \cos(\theta)O_- \end{pmatrix}. \quad (2.48)$$

We can see that $O_1$ and $O_2$ are still decorrelated

$$\begin{aligned} \langle O_1 O_2 \rangle &= \langle (\cos(\theta)O_+ + \sin(\theta)O_-)(-\sin(\theta)O_+ + \cos(\theta)O_-) \rangle \\ &= -\cos(\theta)\sin(\theta)(\langle O_+^2 \rangle - \langle O_-^2 \rangle) \\ &= 0 \end{aligned} \quad (2.49)$$

since $\langle O_+^2 \rangle = \langle O_-^2 \rangle$. Since $\langle O_1^2 \rangle = \langle O_2^2 \rangle$, the whitening is still maintained. It can be intuitively seen in Fig. (2.5) that responses could be equivalently read out from any two orthogonal axes rotated from the two depicted ones $(O_1, O_2)$. Hence, both encoding schemes $S_{L,R} \rightarrow O_\pm$ and $S_{L,R} \rightarrow O_{1,2}$, with the former a special case of the latter, are equally optimal in making the output decorrelated (non-redundant), and in convey information about $S_{L,R}$, and in saving the coding cost $\sum_a \langle O_a^2 \rangle$. This is a particular manifestation of a degeneracy in the optimal encoding solutions discussed in section (2.3.1), that is the objective of the optimization $E$ = cost -$\lambda I(\mathbf{O}; \mathbf{S})$ is invariant to the rotation $O_\pm \rightarrow O_{1,2}$.

Omitting noise,

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

Hence the two neurons coding $O_1$ and $O_2$ in general are differentially sensitive to inputs from different eyes. In particular, $\theta = -45^o$ gives $O_{1,2} \propto S_L(g_+ \mp g_-) + S_R(g_+ \pm g_-)$ shown in Fig. (2.5). With $g_- > g_+$, both $O_1$ and $O_2$ are excited by input from one eye (right and left respectively) and inhibited by input from another, extracting the ocular contrast signal. Varying U leads to a whole spectrum of possible neural ocularities from very binocular to very monocular, as is indeed the case in V1.

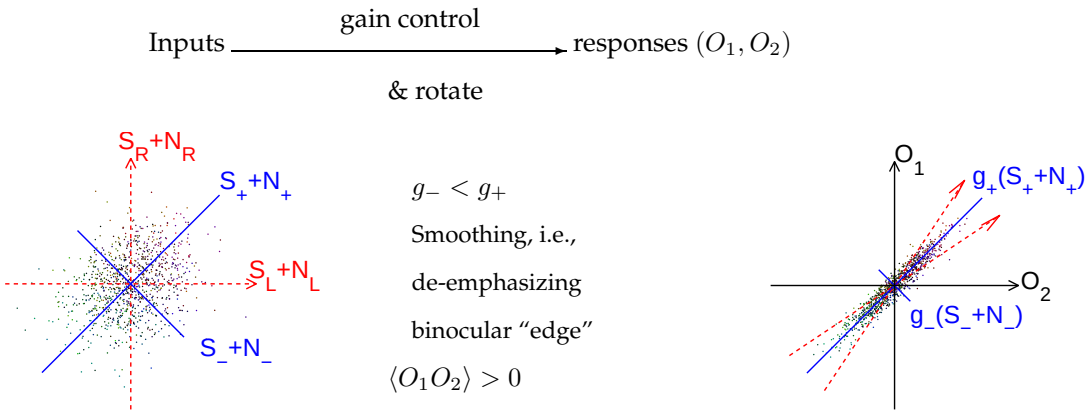



Figure 2.6: Stereo coding when S/N$\ll 1$, the weaker signal $S_-$ is de-emphasized or abandoned to avoid transmitting too much noise. Both $O_1$ and $O_2$ integrate the left and right inputs to smooth out noise, while preferring left and right eyes respectively.

### 2.4.5 Smoothing and output correlation in the low S/N region

When input S/N is too low, the binocular correlation is submerged by the independent noise in the two input channels, as seen by the less discernable elipse shape of the probability distribution of inputs in Fig. (2.6). Equation (2.46) indicates that the gain $g_k$ should decrease with signal strength $\langle S_k^2\rangle$, i.e., $g_- < g_+$. This is to avoid wasting the output channel by transmitting too much noise $g_- N_-$. Then, the weaker binocular contrast signal is de-emphasized or totally abandoned, as illustrated in Fig. (2.6). This is called smoothing, i.e., smoothing out the differences (or noises) between the inputs from different channels (in this case, inputs from different eyes), by averaging over the channels though the gain $g_+$ which is now stronger than $g_-$. The smoothing is thus the opposite to contrast enhancement.

The output channels $O_+$ and $O_-$ are still decorrelated, although they are no longer equally powered, with $\langle O_+^2\rangle \gg \langle O_-^2\rangle$, since the gain $g_+$ and input power $\langle S_+^2\rangle$ for the binocular summation channel are both larger than their counter parts $g_-$ and $\langle S_-^2\rangle$ in the opponent channel, giving

$$\langle O_+^2\rangle \gg \langle O_-^2\rangle$$

However, when $O_+$ and $O_-$ are multiplexed by a rotation matrix U to give a general $O_1$ and $O_2$ output channels, both $O_1$ and $O_2$ will be dominated by inputs from the $S_+$ channel when $g_+$ is sufficiently larger than $g_-$. In particular, when $g_+ \gg g_-$, $O_{1,2} \propto g_+(S_L + S_R)$ + noise, both output channels are integrating the correlated inputs $S_L$ and $S_R$ to smooth out noise, and are consequently correlated with each other. Indeed, equation (2.49) indicates that

$$\langle O_1 O_2\rangle \propto \langle O_+^2\rangle \neq 0 \quad \text{when } \langle O_+^2\rangle \gg \langle O_-^2\rangle.$$

Consider the example from (2.4.3), in which $r = 60/77$, the input noise power $\langle N^2\rangle = 1$, and the output noise power is $\langle N_o^2\rangle = 0.1$. et us reduce the input signal power such that $\langle S_+^2\rangle/\langle N^2\rangle = 1$ and $\langle S_-^2\rangle/\langle N^2\rangle = 7/127$. The total input information rate is now

$$I(\mathbf{S}+\mathbf{N};\mathbf{S}) = \frac{1}{2}\sum_{+,-}\log_2(1+\langle S_i^2\rangle/\langle N^2\rangle) = 0.54\text{bits}$$

mostly from the $S_+$ channel which supplies $0.5$ bits. This amount is much less than the $5$ bits in section (2.4.3) when the S/N is much higher. If we give $g_+ = 0.5$ and $g_- = 0$, the output information is

$$\sum_{k=+,-} I_k = \frac{1}{2}\sum_{k=+,-}\log_2\frac{g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle} = 0.39 \text{ bits}$$

exclusively from the $S_+$ channel (since the $S_-$ is ignored by the zero gain $g_- = 0$). This is a reduction of $25\%$ from the amount of input information. Meanwhile, the total output power is $0.7$, while sending the inputs with non-optimal gain (without the gain control) $g_\pm = 1$ would cost $3.26$ output power.

Multiplexing $O_+$ and $O_-$ by a $-45^o$ rotation, as in section (2.4.4) would spread this power cost in two channels $O_1$ and $O_2$, without changing the total power cost or the total information extracted. In each channel, the gain to $S_+$ signal is now $g'_+ = g_+/\sqrt{2}$, extracting information in the amount of $I(O_1;S_+) = I(O_2;S_+) = \frac{1}{2}\log_2\frac{(g'_+)^2(\langle S_+^2\rangle+\langle N^2\rangle)+\langle N_o^2\rangle}{(g'_+)^2\langle N^2\rangle+\langle N_o^2\rangle} = 0.3181$ bits. Each output channel is extracting more than half of the total output information of $0.39$ bits, giving a redundancy of $2.*0.3181/0.39-1 = 0.63$. This is expected since the two output channels are correlated, and the redundancy should help the input signal recovery. In any case, the low power cost and small amount of information extraction means that, at low S/N, the dynamic range and information channel capacity of the output channels (which should be determined by the maximum amount needed in high S/N conditions) are not fully utilized.

### 2.4.6 Adaptation of the optimal code to the statistics of the input environment

Changing the input statistics, i.e., the correlation matrix $R^S$, changes the optimal coding $\mathbf{S} \to \mathbf{O}$. These changes can be manifested as changes in signal-to-noise, in ocular correlation $r$, or in the

balance or symmetry between the two eyes. The differences in the input statistics can be caused by short term environmental adaptation, such as going from day time to night vision when the S/N changes, or long term differences such as in different visual development conditions (Li 1995). These changes lead to the changes in the eigenvectors or principal components of $R^S$, and to changes in S/N of the principal components, and thus the resulting stereo encoding.

For instance, in V1, different neurons have different sized receptive fields, and can thus integrate inputs spatially to different degrees. Thus the cells with smaller receptive fields receive inputs of smaller S/N. These cells should thus de-emphasize the $S_-$ channel, and are thus more binocular (unless the RF is so small that correlation $r \to 0$, leading to monocular cells (Li and Atick 1994b, Li 1995)). (This coupling between spatial coding and stereo coding is an example of coupling between various other input dimensions discussed later.) In dimmer environments, S/N is lowered for cells of all RF sizes. More V1 cells are then binocular, causing weaker sensitivity to depth information which derives from the $S_-$ channel.

In strabismus, the ocular correlation $r$ is weaker, giving stronger $S_-$ channel which is nevertheless weaker than the $S_+$ channel. The sensitivity $g_-$ to the $S_-$ channel is thus preserved for neurons of all or most RF sizes. As a result, fewer cells are binocular. The ocular dominance columns would be stronger, as indeed observed in animals whose eyes are misaligned surgically or optically during development.[28]

In animals with short inter-ocular distances, such as squirrel monkeys, the binocular correlation $r$ can be larger than that of other primates like humans. This is the opposite situation from that of the strabismus, and the $S_-$ channel has weaker signals. Consequently, more cells are binocular, and the ocular dominance columns should be weaker, as is indeed the case for squirrel monkeys.[26] Such developmental situations can also be simulated by artifically synchronous inputs to the two eyes, leading to similar consequences.[89]

Input statistics can also change with input characteristics. For instance, since the two eyes are displaced from each other horizontally, visual input oriented horizontally have stronger ocular correlation $r$ than input oriented vertically, as has been measured in natural scenes (Li and Atick 1994b). This is because any object in the scene can have their images on the two retinas at two different locations, when the object is not on the zero-disparity depth plane, i.e., at a distance shorter or longer than where the two eyes converge or focus on. This difference in image positions is called disparity. Usually, horizontal disparities are larger than vertical disparities due to the horizontal displacements of the eyes, causing smaller ocular correlation $r$. A vertical or horizontal input bar creates horizontal or vertical disparities mainly, and thus their ocular correlation differences. Therefore, inputs oriented vertically or horizontally create ocular correlations that are more towards or away from strabismus, respectively. Consequently, V1 neurons oriented horizontally are predicted to be more likely monocular (Li and Atick 1994a).

In monocular deprivation of the developmental conditions, inputs to one eye is deprived, leading to the asymmetry $R^S_{LL} = \langle S^2_L \rangle \neq R^S_{RR} = \langle S^2_R \rangle$. Consequently (Li 1995), the eigenvectors and eigenvalues of $R^S$ change: $S^+$ is strong-eye-dominant, and $S^-$, the binocular edge, is weak-eye-dominant and easily overcome by noise. In fact, $S^-$ has a negligible signal power for most scales under severe monocular deprivation when $a \ll 1$. This gives a majority of the strong-eye-dominant cells and a thicker corresponding ODC, as observed in physiology (e.g.,[29]).

## 2.5 Applying efficient coding to understand coding in space, color, time, and scale in retina and V1

Stereo coding illustrates a general recipe, as in Fig (2.3), for optimally efficient linear coding transformation $\mathbf{O} = \mathsf{K}(\mathbf{S} + \mathbf{N}) + \mathbf{N}_o$ of Gaussian signals $\mathbf{S}$ with correlation matrix $R^S$, given independent Gaussian input noise $\mathbf{N}$ and additional coding noise $\mathbf{N_o}$. The recipe contains three conceptual (though not neural) components: $\mathsf{K}_o$, $\mathsf{g}$, and $\mathsf{U}$, as follows:

$\mathbf{S} \to \mathcal{S} = \mathsf{K}_o\mathbf{S}$ — find principal components (PCA) $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, ...\mathcal{S}_k...)$ by transform $\mathsf{K}_o$

$\mathcal{S}_k \to O_k = g_k\mathcal{S}_k$ — gain control $g_k$ (a function of $\mathcal{S}_k/\mathcal{N}_k$) to each PCA $\mathcal{S}_k$ by equation (2.41)

$\mathbf{O} \rightarrow \mathsf{U}\mathbf{O}$ — freedom by any unitary transform $\mathsf{U}$ to suit any additional purpose[1].

The overall effective transform is $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, where $\mathsf{g}$ is a diagonal matrix with elements $\mathsf{g}_{kk} = g_k$. When $\mathsf{U} = 1$, the optimal coding transform is $\mathsf{K} = \mathsf{g}\mathsf{K}_o$. The resulting $\mathbf{O} = (O_1, O_2, ...)$ has decorrelated components and retains the maximum information about $\mathbf{S}$ for a given output cost $\sum_k \langle O_k^2 \rangle$. Using any other unitary transform $\mathsf{U}$ gives equally optimal coding , since it leaves the outputs $\mathbf{O}$ with the same information $I(\mathbf{O}, \mathbf{S})$ and cost, and, in the zero noise limit, the same decorrelation. The three conceptual steps above are equivalent to the single mathematical operation of finding the solution $\mathsf{K}$ of $\partial E / \partial \mathsf{K} = 0$ where $E(\mathsf{K}) =$ cost - $\lambda I(\mathbf{O}; \mathbf{S})$. The solution is degenerate, i.e., there are many equally good solutions corresponding to arbitrary choices of unitary transforms (or rotations) $\mathsf{U}$. The input statistics, manifested in the correlation matrix $R^S$, determine the optimal coding $\mathsf{K}$ through at least the first two conceptual steps. In particular, S/N levels control $g_k$, giving contrast enhancement and decorrelation in high S/N, and input smoothing and response correlation in low S/N.

While the inputs are correlated as described by $R^S$, the output correlation caused by inputs is

$$\langle O_i O_j \rangle = \mathsf{K}_{ia} \mathsf{K}_{jb} \langle S_a S_b \rangle = (\mathsf{K} R^S \mathsf{K}^\dagger)_{ij} \tag{2.50}$$

where the superscript † denote the conjugate transpose of a matrix, e.g., $\mathsf{K}_{ij} = (\mathsf{K}^\dagger)^*_{ji}$, with $*$ indicating complex conjugate. As $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, we have

$$\langle O_i O_j \rangle = [\mathsf{U}(\mathsf{g}\mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g})\mathsf{U}^\dagger]_{ij} \tag{2.51}$$

Since $\mathsf{U}\mathsf{U}^\dagger = 1$, i.e., $(\mathsf{U}\mathsf{U}^\dagger)_{ij} = \delta_{ij}$, $\langle O_i O_j \rangle \propto \delta_{ij}$ when $\mathsf{g}\mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g}$ is proportional to an identity matrix. The definition of $\mathsf{K}_o$ means that

$$(\mathsf{K}_o R^S \mathsf{K}_o^\dagger)_{ij} = \delta_{kk'} \lambda_k \equiv \delta_{kk'} \langle |\mathcal{S}_k|^2 \rangle \tag{2.52}$$

where $\lambda_k$ is the $k^{th}$ eigenvalue of $R^S$. Thus $\mathsf{g}\mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g}$ is proportional to identity when $g_k^2 \propto 1/\langle |\mathcal{S}_k|^2 \rangle$, which is the same when S/N is sufficiently high for all principal components. Hence, as expected, output channels are decorrelated

$$\langle O_i O_j \rangle \propto \delta_{ij} \quad \text{when S/N is sufficiently high for all input components} \tag{2.53}$$

In contrast, output channels are correlated

$$\langle O_i O_j \rangle \not\propto \delta_{ij} \quad \text{when S/N is low for some input components} \tag{2.54}$$

We can now apply our understanding to visual coding in space, time, and color, always approximating signals as Gaussian.

### 2.5.1 Efficient spatial coding for retina

In spatial coding (Srinivasan et al 1982, Linsker 1990, Atick and Redlich 1990), a signal at visual location $x$ is $S_x$. The input correlation is

$$R^S_{xx'} = \langle S_x S_{x'} \rangle.$$

As one can see in Fig. (2.7)A, nearby image pixels tend to have similar input intensity, just like inputs in two eyes tend to be similar in stereo vision. Furthermore, this similarity decreases with increasing distance between the two pixels (Fig. (2.7)D). This means $R^S_{xx'}$ decreases with increasing distance $|x - x'|$, and one can expect that $R^S_{xx'}$ is translation invariant, depending only on $x - x'$.

[1]The $\mathsf{U}$ symmetry holds when the cost is $\sum_i \langle O_i^2 \rangle$ or $H(\mathbf{O})$, but not $\sum_i H(O_i)$ except in the noiseless case. Given finite noise, the cost of $\sum_i H(O_i)$ would break the $\mathsf{U}$ symmetry to a preferred $\mathsf{U}$ as the identity matrix, giving zero second order correlation between output channels . The fact that early vision does not usually have the identity $\mathsf{U}$ suggests that the cost is more likely output power $\sum_i \langle O_i^2 \rangle$ than $\sum_i H(O_i)$. For instance, the retinal coding maximizes second order output correlation given $\sum_i \langle O_i^2 \rangle$ and $I(\mathbf{O}; \mathbf{S})$ in Gaussian approximation, perhaps aiding signal recovery.

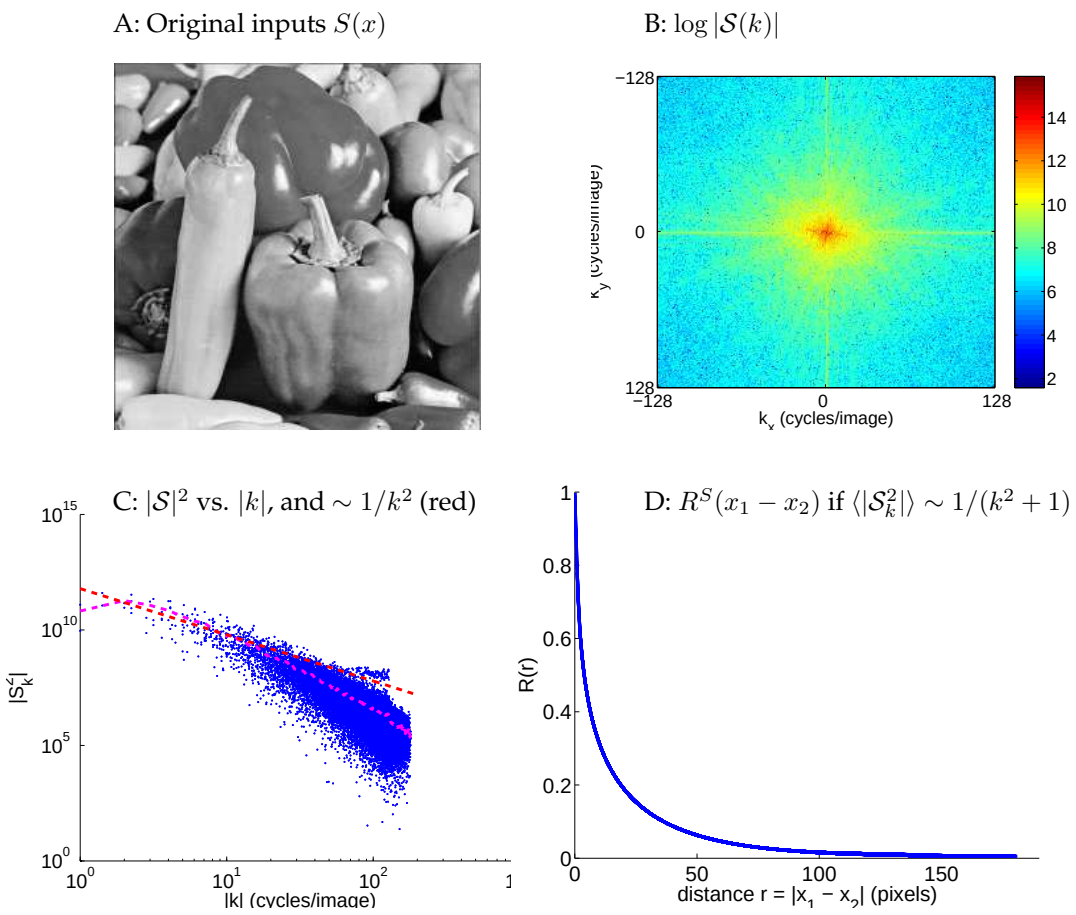



Figure 2.7: A: an example of visual input $S_x = S(x)$ in space $x$. The image has 256×256 pixels. B: Fourier transform $\mathcal{S}_k$, visualized as $\log|\mathcal{S}_k|$, as a function of spatial frequency $k = (k_x, k_y)$. C: Power $|\mathcal{S}_k|^2$ vs. $|k|$ for input $S(x)$ in A; Shown for comparison, in red color, is also $\sim 1/|k|^2$. In magenta color is the average of $|\mathcal{S}_k|^2$ over a local $k$ range for input in A. D: Spatial correlation $R^S(x_1 - x_2)$ assuming $\langle|\mathcal{S}_k|^2\rangle = 1/(k^2 + k_o^2)$ for low cutoff frequency $k_o = 1$ cycles/image.

Hence, we denote $R^S(x - x') = R^S_{xx'}$ as the auto-correlation function of the spatial inputs. Thus the principal components of $R^S$ are the Fourier components, as a function of space

$$S_x \propto e^{ikx},$$

where $k$ is the spatial frequency of the Fourier wave, and is used here as the index for the principal component. This principal component can be verified as

$$\sum_{x'} R^S(x - x')e^{ikx'} = e^{ikx} \sum_{x-x'} R^S(x - x')e^{-ik(x-x')} = \text{constant } e^{ikx}$$

while the eigenvalue is the Fourier transform $\sum_x R^S(x)e^{-ikx}$ of $R^S(x)$. Hence, the principal component transform matrix $\mathsf{K}_o$ has elements $(\mathsf{K}_o)_{kx} \propto e^{-ikx}$. When $k = 0$, the Fourier wave has zero frequency. The signal in this mode is thus analogous to the $S_+$ mode in stereo vision, signalling the average inputs in different input channels or locations. When $k \neq 0$, the input mode signal the input differences or contrast between different locations $x$, and is analogous to the mode $S_-$ in stereo vision. Having more input channels (locations) than just two channels (eyes) in stereo vision, spatial inputs can have many different ways of input changes with space, hence different frequency $k$ for different Fourier modes.

The amplitudes of the Fourier modes or principal components are

$$\mathcal{S}_k = \sum_x \mathsf{K}_o^{kx} S_x \sim \sum_x e^{-ikx} S_x.$$

Figure (2.7)AB give an example input $S_x$ and its Fourier amplitudes. It is clear that there is a trend of higher signal power in modes of lower spatial frequencies. This is again analogous to stereo vision, correlations between input channels make signal power higher in input modes that smoothes inputs.

The average powers of the Fourier modes are the eigenvalues $\lambda_k^S$ of $R^s$

$$\langle|\mathcal{S}_k|^2\rangle \propto \int dxdx' e^{-ik(x-x')}\langle S_x S'_x\rangle = \int dxdx' e^{-ik(x-x')} R^S(x-x') = \int dxdx' e^{-ikx'} R^S(x') \propto \lambda_k^S$$

as expected. We denote these eignvalues $\lambda_k^S$ here as $\mathcal{R}^S(k)$, the Fourier transform of $R^S(x)$.

Field (1987) measured the power spectrum as $\langle\mathcal{S}_k^2\rangle \sim 1/k^2$. Meanwhile, the general variation of signal power with frequency $|k|$ in any specific example such as Figure (2.7)AB can be similar but not identical to $1/k^2$. The measurements of $\langle\mathcal{S}_k^2\rangle$ also indirectly measured $R^S(x) \propto \int dk\langle\mathcal{S}_k^2\rangle e^{ikx}$ as the inverse Fourier transform of $\langle\mathcal{S}_k^2\rangle$. Figure (2.7)D shows that this correlation $R^S(x)$ can exist for long distances $x$ with $\langle\mathcal{S}_k^2\rangle \sim 1/(k^2+k_o^2)$ for a low cutoff frequency of $k_o = 1$ cycle/image.

When considering input noise, as shown superposed on an image, in Fig. (2.10), the noise at different locations are not assumed uncorrelated, thus

$$\langle N_x N_{x'}\rangle \equiv \langle N^2\rangle\delta_{xx'}$$

Hence, the power spectrum of the noise is constant, i.e., the noise is the white noise

$$\langle|\mathcal{N}_k|^2\rangle = \langle N^2\rangle$$

Let $k_p$ denote the spatial frequency when $\langle|\mathcal{S}_k|^2\rangle = \langle N^2\rangle$. Then, in the low frequency region when $k < k_p$, the signal-to-noise $\mathcal{S}^2/\mathcal{N}^2$ is high; in the high frequency region when $k > k_p$, the signal-to-noise $\mathcal{S}^2/\mathcal{N}^2$ is low. Therefore, when $k < k_p$, the gain $g_k$ or $g(k) \propto \langle\mathcal{S}_k^2\rangle^{-1/2} \sim k$ approximates whitening. This coding region thus emphasizes higher spatial frequencies and extracts image contrast. However, when frequency $k > k_p$, $\mathcal{S}^2/\mathcal{N}^2 \ll 1$ is low, $g(k)$ quickly decays with increasing $k$ according to equation (2.41) in order not to amplify image contrast noise. Hence, $g(k)$ as a function of $k$ peaks at $k_p$ where $\mathcal{S}^2(k)/\mathcal{N}^2(k) \sim 1$ (Fig (2.8)).

If $\mathsf{U}$ is the inverse Fourier transform $\mathsf{U}_{x'k} \sim e^{ikx'}$, the whole transform $\mathsf{K}_{x'x} = (\mathsf{U}\mathsf{g}\mathsf{K}_o)_{x'x}$ gives band-pass filters

$$K(x'-x) \equiv \mathsf{K}_{x'x} \sim \sum_k g(k) e^{ik(x'-x)}$$

with frequency sensitivities $g(k)$. This filter gives response

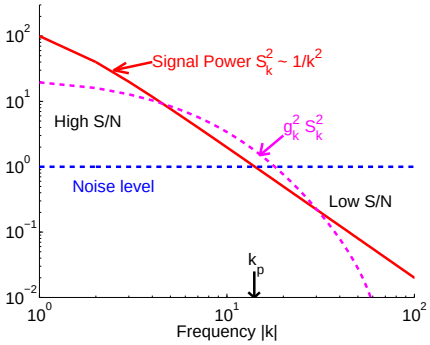
$$O_{x'} = \sum_x K(x'-x) S_x + \text{noise},$$

for an output neuron with RF centered at $x'$. This is what retinal output (ganglion) cells do, achieving a center-surround transform on the input and emphasizing the intermediate frequency band for which S/N is of order 1. That is, they enhance image contrasts up to an appropriate spatial detail without amplifying contrast noise. Note that this choice of $\mathsf{U}$ as the inverse of $K_o$ makes receptive fields for all neurons the same shape except for a translation of their center location $x'$. It also makes the RF shape small or localized. Since $\mathsf{K}(x)$ is a band-pass filter with optimal spatial frequency $k_p$, the spatial extent of the receptive field is of order $1/k_p$. The filter $K(x'-x)$ is radially symmetric since the statistics $\langle\mathcal{S}(k)^2\rangle$, and thus $g(k)$, depends only on the magnitude $|k|$. The contrast sensitivity function to image gratings is the behavioral manifestation of $g(k)$.

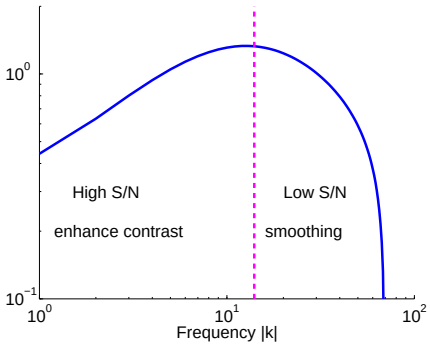
The output Fourier Amplitude is $\mathcal{O}(k) = g(k)\mathcal{S}(k)$, and thus the mean output power is

$$\langle|\mathcal{O}(k)|^2\rangle = g^2(k)\langle|\mathcal{S}(k)|^2\rangle \approx \text{constant for small } k \text{ up to } k < k_p, \tag{2.55}$$

A: Power of input signal $\langle \mathcal{S}_k^2 \rangle$, output signal $g^2(k)\langle \mathcal{S}_k^2 \rangle$, and noise $\langle \mathcal{N}_k^2 \rangle$ vs. $k$.




B: The optimal gain $g(k)$




C: The Spatial receptive field.

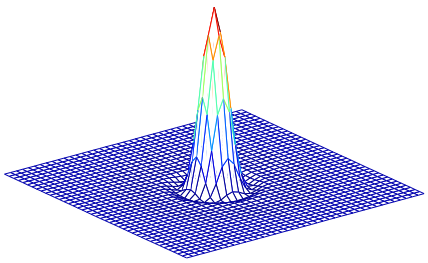


Figure 2.8: Illustration of input statistics and optimal encoding in space. A: the power spectra of input signal $\langle \mathcal{S}_k^2 \rangle = 200./(|k|^2+1)$, output signal $g^2(k)\langle \mathcal{S}_k^2 \rangle$, and white noise $\langle \mathcal{N}_k^2 \rangle = 1$. Note that $\langle \mathcal{S}_k^2 \rangle = \langle \mathcal{N}_k^2 \rangle$ at $k = k_p$ as indicated. B: the optimal gain $g(k)$ by equation (2.41), given input $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}_k^2 \rangle$ in A, when $\frac{4\lambda}{(\ln 2)\langle N_o^2 \rangle} = 100$. Note that $g(k)$ peaks around $k = k_p$. C: the shape of the receptive fields $\mathsf{K}(x) \sim \int \tilde{g}(k) e^{ikx}$, the inverse Fourier transform of $\tilde{g}(k) = g(k) e^{-(|k|/50)^4}$ where $e^{-(|k|/50)^4}$ is the extra low pass filter (modeling the optical transfer function of the eye) which together with the optimal filter $g(k)$ makes the effective receptive field. All $k$ are in units of cycles/image. Note that $1/k_p$ should roughly be the size of the receptive field.

since $g^2(k) \propto 1/\langle |\mathcal{S}(k)|^2 \rangle$ in lower $k < k_p$, as illustrated in Fig. (2.8 ). This means the output is like spatial white noise up to spatial frequency $k_p$. This can also be seen in the output correlation between two neurons $O_x$ and $O_{x'}$ at locations $x$ and $x'$, as

$$\langle O_x O_{x'} \rangle = \sum_{ab} \mathsf{K}_{xa} \mathsf{K}_{x'b} \langle S_a S_b \rangle = (\mathsf{K} R^S \mathsf{K}^\dagger)_{xx'} \tag{2.56}$$

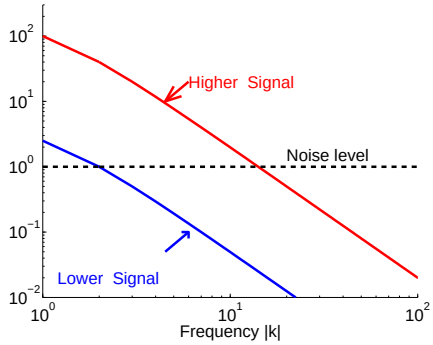
$$= (\mathsf{U} \mathsf{g} \mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g} \mathsf{U}^\dagger)_{xx'} = \int dk e^{ik(x-x')} g^2(k) \mathcal{R}(k) \tag{2.57}$$

Note that, since U is a unitary matrix composed of orthonormal vectors as its row or column vectors, in particular, $\int dk e^{ik(x-x'} \propto \delta(x-x')$. Then when $g^2(k)\mathcal{R}(k)$ =constant, which happens with whitening, the above equation becomes

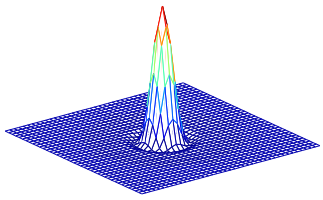
$$\langle O_x O_{x'} \rangle \propto \delta(x-x'), \text{with whitening or } g^2(k)\mathcal{R}(k) = \text{constant}.$$

When in general $g(k)\mathcal{R}(k)$ is invariant with $k$ only in the range $k < k_p$, and decreases with $k$ for $k > k_p$, different outputs $O_x$ and $O_{x'}$ are correlated when $x - x' \sim\leq 1/k_p$. Thus output correlation

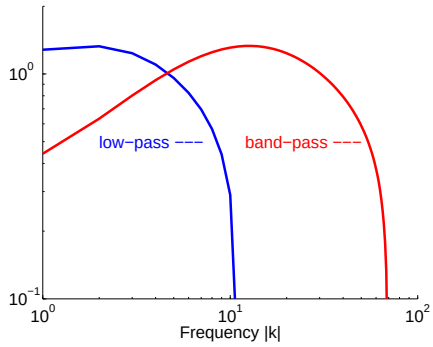
A: Inputs of high and low signal-to-noise




C: Center-Surround RF at higher S/N



B: Band and low pass filters $g(k)$ them




D: Smoothing RF at low S/N

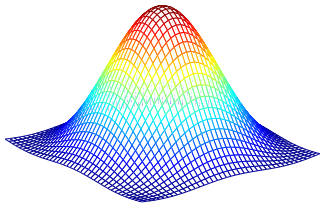


Figure 2.9: Illustration of how receptive fields adapt to changes in input signal-to-noise. A: the input power spectrum $\langle \mathcal{S}_k^2 \rangle = \hat{S}^2/(|k|^2+1)$ for high and low S/N, $\hat{S}^2 = 200$ and 5 respectively. B: for inputs in A, the resulting optimal gain $g(k)$ is band and low pass filters respectively. C: the shapes of the receptive fields $\mathsf{K}(x)$ for high and low S/N conditions. One is center-surround shaped with small size, and the other is gaussian smoothing shaped with a larger size. All other parameters are the same as in Fig. (2.8).

is particularly significant when S/N is low, when $g(k)\mathcal{R}(k)$ decays with $k$ for a larger range of $k$. Large output correlations indeed occur physiologically (Puchalla et al 2005).

In a dimmer environment, inputs are weakened, say from $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 200/k^2$ to $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 5/k^2$, the peak sensitivity of $g(k)$ occurs at a lower frequency $k_p \to k/\sqrt{40}$, effectively making $g(k)$ a low pass, as shown in Fig. (2.9). Accordingly, $K(x)$ integrates over space for image smoothing rather than contrast enhancing, to boost signal-to-noise while sacrificing spatial resolution, as illustrated in Fig. (2.9). This explains the dark adaptation of the RFs of retinal ganglion cells or LGN cells,[11,37] from center-surround contrast enhancing (band-pass) filter to Gaussian-like smoothing (low-pass) filter, to integrate signals and smooth out contrast noise. The smoothing filters naturally lead to highly correlated responses between output neurons, especially when the filter diameters are larger than the distances between the RFs.

### 2.5.2 Efficient coding in time

Coding in time $O_t = \sum_{t'} \mathsf{K}_{tt'} S_{t'}$ + noise is analogous to coding in space, when input $S_x$ indexed by space $x$ is now changed to input $S_t$ indexed in time $t$. However, the temporal filter $\mathsf{K}_{tt'}$ has to be such that it should be temporally translation invariant and causal, i.e., $\mathsf{K}_{tt'} = \mathsf{K}(t-t')$ and $\mathsf{K}(t) = 0$ when $t < 0$. It is also called an impulse response function of a neuron. Just like in space, the temporal correlation function $R^S_{tt'} = R^S(t-t')$ is expected to be temporally translation invariant,

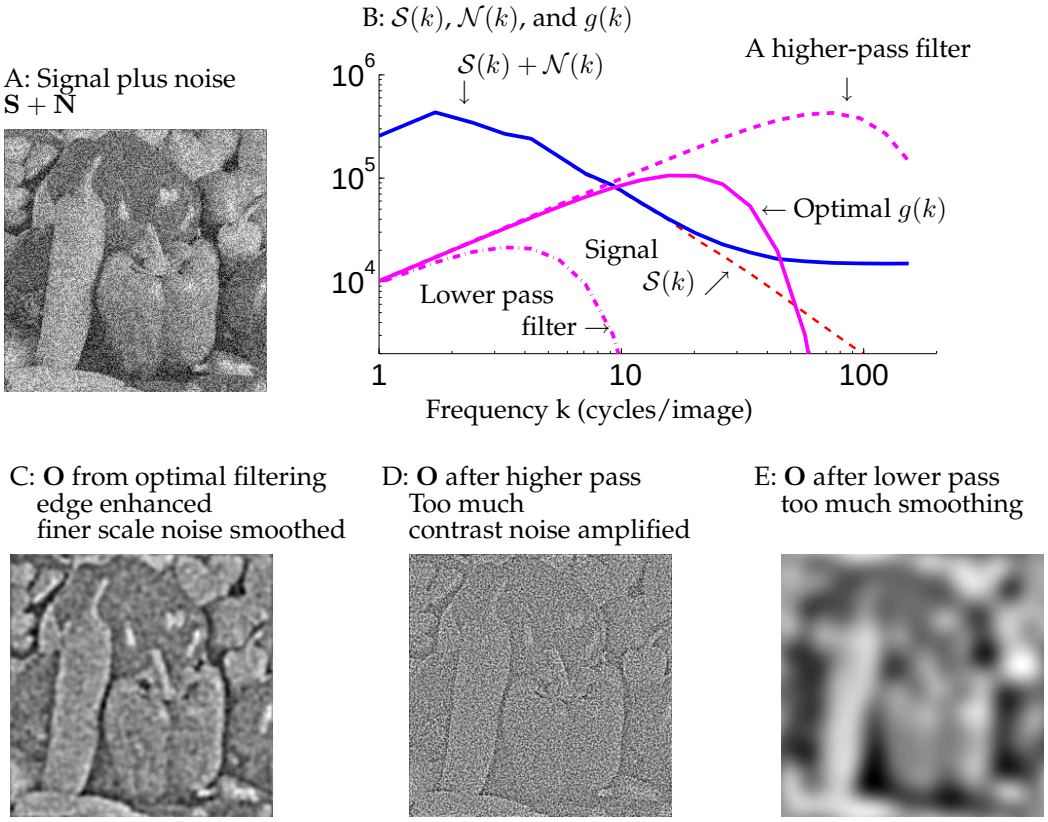



Figure 2.10: Signal transform by optimal and non-optimal coding of visual input in space. A: image $\mathbf{S}$ with white noise $\mathbf{N}$. B: amplitude spectrums $\mathcal{S}(k)$ (red-dashed), total inputs (blue) $\mathcal{S}(k) + \mathcal{N}(k)$, and the filter sensitivity functions $g(k)$ (magenta) as functions of frequency $k$. (The vertical axis has an arbitrary scale). The optimal curve $g(k)$ is solid magenta, it peaks near $k_p$ where $\mathcal{S}(k) + \mathcal{N}(k)$ starts to depart from $\mathcal{S}(k)$ significantly. For comparison, filters with higher and lower pass sensitivities are in magenta-dashed, and magenta-dash-dotted, respectively. D: response $\mathbf{O} = \mathsf{K}(\mathbf{S} + \mathbf{N})$ after optimal filtering $\mathsf{K}$ with the optimal sensitivity curve $g(k)$. Thus, image contrast (edge) is enhanced at low $k$ where $g(k)$ increases with $k$ but smoothed at high $k$ where $g(k)$ decreases with $k$ to avoid transmitting too much noise at finer spatial scale. D and E: outputs $\mathbf{O}$ when the filters are higher or lower pass as depicted in B. Gray scale values shown in A, C, D, E are normalized to the same range.

and thus can be characterized by the power spectrum in time

$$\mathcal{R}^S(\omega) \sim \int dt R^S(t) e^{-i\omega t} = \langle |\mathcal{S}_\omega|^2 \rangle$$

where $\mathcal{S}_\omega = \sum_t (\mathsf{K}_o)_{\omega,t} S_t \sim \int dt e^{-i\omega t} S(t)$ is the temporal Fourier transform of input $S(t)$ at temporal frequency $\omega$, representing the amplitude of the principal component of the inputs. One can expect that $R^S(t)$ should decay monotonically and smoothly with $t$, and thus $\mathcal{R}^S(\omega)$ is also expected to decay with $\omega$, as is measured experimentally.[17]

Given $\mathcal{R}^S(\omega)$ and noise spectrum, the temporal frequency sensitivity (often called temporal contrast sensitivity function experimentally)

$$g(\omega) \sim |\int dt \mathsf{K}(t) e^{-i\omega t}|$$

is determined by equation (2.41) according to the S/N value $\langle |\mathcal{S}_\omega|^2 \rangle / \langle |\mathcal{N}_\omega|^2 \rangle$ at this frequency $\omega$. Since $\mathcal{R}^S(\omega)$ decays with $\omega$, then, just as in spatial coding, $g(\omega)$ should increase with $\omega$ till at $\omega = \omega_p$ when the S/N is of order 1. In the example of Fig. (2.11AE), $\omega_p/(2\pi) \sim 5$ Hz.

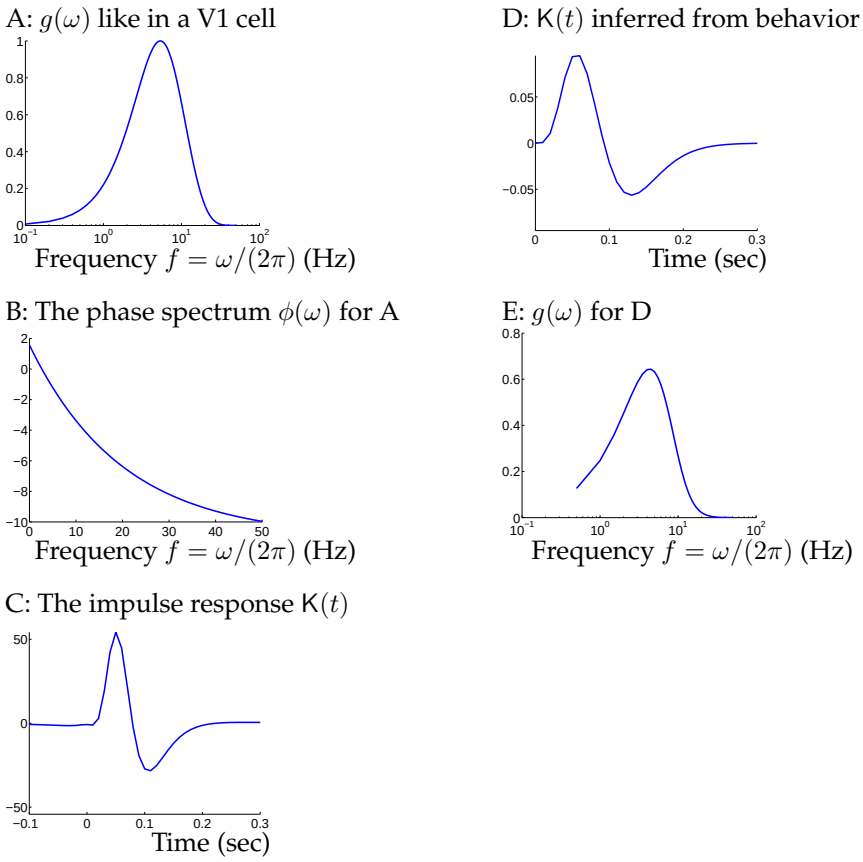



Figure 2.11: Examples of temporal contrast sensitivity functions $g(\omega)$, the impulse response functions $\mathsf{K}(t)$ and its phase spectrum $\phi(\omega)$. A: $g(f) = \exp(-(0.1f)^{1.5}) - \exp(-(0.3f)^{1.5}$ (where $f = \omega/(2\pi)$ is the frequency), as typically observed by Holub and Morton-Gibson (1981)[25] in neurons from cat visual area 17 by their responses to drifting gratings. B: a particular choise of phase spectrum, $\phi(f) = -2\pi\tau_p \cdot [20(1 - e^{-f/20})] + \phi_o$ for $\tau_p = 0.1$ second and $\phi_o = \pi/2$, to make $\mathsf{K}(t)$ causal and of small temporal spread. Note that $\phi(\omega)$ varies approximately linearly with $\omega$ except for very large $\omega$. C: the impulse response function $\mathsf{K}(t) = \int df g(f) e^{i(2\pi f t + \phi(f))}$ from A and B. D: an impulse response function $\mathsf{K}(t) = e^{-\alpha t}[(\alpha t)^5/5! - (\alpha t)^7/7!]$, inferred by Adelson and Bergen[1] from human visual behavior. Here, $\alpha = 70$/second. E: $g(f)$, the amplitude of Fourier transform of the $\mathsf{K}(t)$ in D.

In implementation, it is desirable that the causal temporal filter $\mathsf{K}(t)$ should also be of minimum temporal spread and have short latency, i.e., $\mathsf{K}(t)$ is significantly non-zero for only a short temporal window and for short times $t$. This can be done by appropriate (Dong and Atick 1995, Li 1996) choice of $\mathsf{U}_{t,\omega} \propto e^{i\omega t + i\phi(\omega)}$, i.e., the appropriate choice of $\phi(\omega)$, to make the temporal filter

$$\mathsf{K}(t) \sim \int d\omega g(\omega) e^{i(\omega t + \phi(\omega))}.$$

A minimal temporal spread for $\mathsf{K}(t)$ means that the individual waves $g(\omega)\cos(\omega t + \phi(\omega))$ for various frequencies $\omega$ that makes up $\mathsf{K}(t)$ are superposed constructively around a particular time $\tau_p$ when $\mathsf{K}(t)$ is large or significantly non-zero, and destructively (i.e., cancelling out) at other times when $\mathsf{K}(t) \approx 0$. Meanwhile, causality means that $\tau_p > 0$. The constructive superposition can be achieved when all waves $g(\omega)\cos(\omega t + \phi(\omega))$ of various frequencies $\omega$ have similar phases, i.e., temporal coherence, at $t \approx \tau_p$, thus

$$\omega\tau_p + \phi(\omega) \quad \text{is almost independent of } \omega.$$

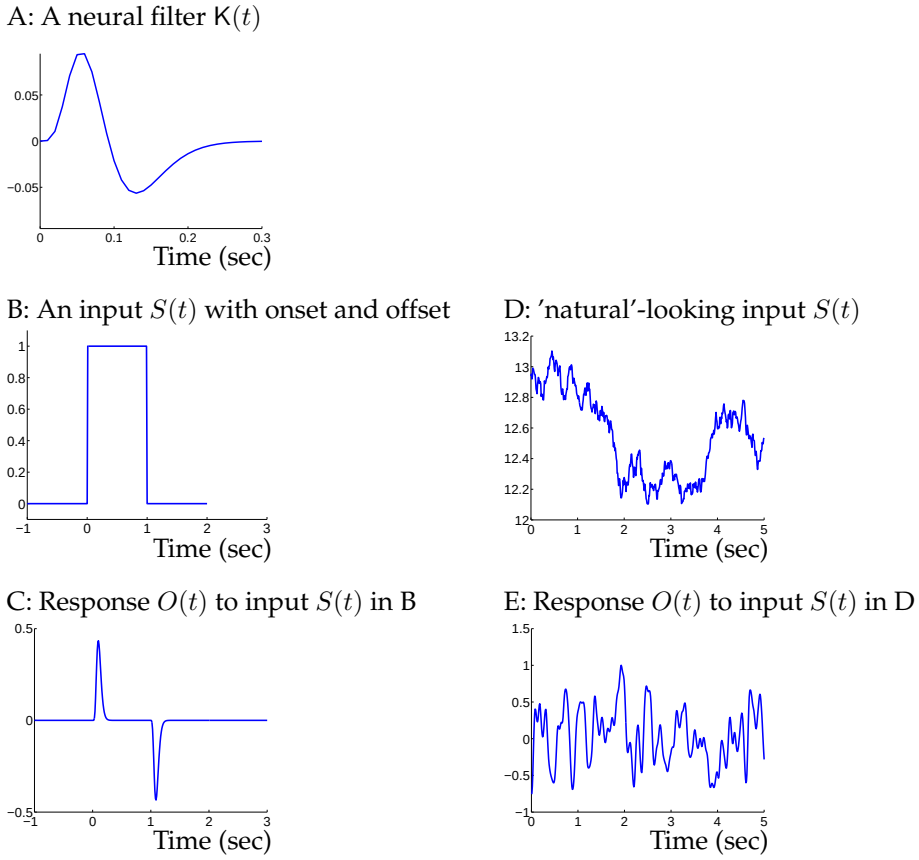



Figure 2.12: Predictive filtering. A: A neural filter $\mathsf{K}(t)$ as in Fig (2.11)D. B: An input $S(t)$. C: Response $O(t)$ to input $S(t)$ in B by filter $\mathsf{K}(t)$ in A. Note that responses are mainly to changes in input. D: An input $S(t)$ which has Fourier amplitude $\sim 1/(f+f_o)$ with $f_o = 0.0017$ Hz and random phases. Note that signal $S(t_1)$ and $S(t_2)$ are correlated even for long interval $t_2 - t_1$. E: Response $O(t)$ to input $S(t)$ in D by filter in A (the scale on vertical axis is arbitrary). Note that the response $O(t_1)$ and $O(t_2)$ are not correlated for $t_2 - t_1$ much longer than the width of $\mathsf{K}(t)$ in A, while the short time scale input flutuations (within 0.1 seconds) are smoothed out in the response.

Therefore,

$$\begin{aligned} \phi(\omega) \quad &\approx \quad -\omega\tau_p + \text{sign}(\omega)\phi_o, \\ &\text{where, sign}(x) = 1 \text{ or } -1 \text{ for positive and negative } x, \\ &\text{and } \phi_o \text{ is a constant that determine the shape of the temporal filtelr } \mathsf{K}(t) \end{aligned}$$

Fig. (2.11) illustrates such an example. At $t = \tau_p$, $\mathsf{K}(t = \tau_p) \sim \cos(\phi_o) \int d\omega g(\omega)$. Since $\tau_p > 0$ is at or near the time when the temporal filter $\mathsf{K}(t = \tau_p)$ is at its peak, and $\tau_p$ is thus effectively the latency of the impulse response function. Physiologically, neurons' temporal filters indeed have such phase spectrum[24] in which phase $\phi(\omega)$ is roughly a linear function of $\omega$.

A typical band-pass temporal filter $\mathsf{K}(t)$, as shown in Fig. (2.11) CD, is such that, at short time $t$, $\mathsf{K}(t)$ integrates in time to average out the noise, and at longer $t$, $\mathsf{K}(t)$ is opponent to its own value from earlier time $t$ to enhance temporal changes. This is because typically $g(\omega) \approx 0$ for small frequencies $\omega \approx 0$. This insensitivity to non-changing inputs makes $\mathsf{K}(t)$ often be called a predictive filter (or predictive coding), so that the optimal predictive filter would make the response minimal except when inputs change significantly. This is illustrated in Fig. (2.12BC). Input correlation $R^S(t)$ can be used to predict input $S(t_1)$ as $S(t_1) \approx \hat{S}(t_1)$ from input history $S(t < t_1)$. The difference $S(t_o) - \hat{S}(t_o)$ between the actual and predicted input is the non-predictable part of the input. The

predictability is used to constructed the optimal filter $\mathsf{K}(t)$ such that responses $O(t)$ is mainly caused by the unpredictable inputs, thus minimizing the response amplitudes.

Input from natural scenes have long range temporal correlations, with their power spectrum[17] $\mathcal{R}^S(\omega) \propto \omega^{-2}$. The filter responses to such inputs should have a white power spectrum $\langle O^2(\omega)\rangle =$ constant up to an $\omega_p$. This means that the output looks like white noise up to frequency $\omega_p$, and outputs are temporally decorrelated for time differences larger than $1/\omega_o$. This is illustrated in Fig. (2.12DE), and and confirmed physiologically for (LGN) neurons which receive inputs from retinal ganglion cells (Dan et al 1996).

Given a sustained input $S(t)$ over time $t$, the output $O(t) = \int \mathsf{K}(t-t')S(t')dt'$ may be more sustained or transient depending on whether the filter $g(\omega)$ is more low pass (performing temporal smoothing) or band pass (enhancing temporal contrast) (Srinivasan et al 1982, Li, 1992, Dong and Atick 1995, Li 1996, van Hateren and Ruderman 1998). As in spatial coding, dark adaptation makes the temporal filter of the neurons more low-pass and the responses more sustained.[37]

### 2.5.3 Efficient coding in color

Visual color coding (Buchsbaum and Gottschalk 1983, Atick et al 1992) is analogous to stereo coding, especially if we simplify by assuming only two cone types, red and green, of comparable input power $\langle S_r^2\rangle \approx \langle S_g^2\rangle$ and correlation coefficient $r \propto \langle S_r S_g\rangle$. Then, the luminance channel, $S_+ \sim S_r + S_g$, like the ocular summation channel, has a higher S/N than the chromatic channel $S_- \sim S_r - S_g$ which is like the ocular opponent channel. Optimal coding awards appropriate gains to them. In dim light, the diminished gain $g_-$ to the cone opponent channel is manifested behaviorally as loss of color vision, with the luminance channel $S_+$ dominating perception.

In the non-simplified version when three cone types are considered, the inputs $\mathbf{S}$ is now

$$\mathbf{S} = (S_r, S_g, S_b) \tag{2.58}$$

for inputs in red, green, and blue cones. Each input $S_i$ is the result of spectrum input $S(\lambda)$ as a function of light wavelength $\lambda$ (not to be confused with our Lagrange multiplier in the optimization) and the cone sensitivity function $R_i(\lambda)$

$$S_i = \int d\lambda S(\lambda) R_i(\lambda) \tag{2.59}$$

Thus the correlation

$$\langle S_i S_j\rangle = \int d\lambda_1 d\lambda_2 R_i(\lambda_1) R_j(\lambda_2) \langle S(\lambda_1) S(\lambda_2)\rangle. \tag{2.60}$$

Hence, the statistics of $S(\lambda)$ and the functions $R_i$ and $R_j$ determine the pair-wise correlation between input signals in different color cones. The resulting input correlation matrix $R^S$ is a $3 \times 3$ matrix

$$R^S = \begin{pmatrix} R^S_{rr} & R^S_{rg} & R^S_{rb} \\ R^S_{gr} & R^S_{gg} & R^S_{gb} \\ R^S_{br} & R^S_{bg} & R^S_{bb} \end{pmatrix} \tag{2.61}$$

which gives three principal components $\mathcal{S}_k$. Assuming (over-simply) that $\langle S(\lambda_1)S(\lambda_2)\rangle = \delta(\lambda_1 - \lambda_2)$, Buchsbaum and Gottschalk (1983) obtained the $R^S$ to give the three components as

$$\begin{pmatrix} \mathcal{S}_1 \\ \mathcal{S}_2 \\ \mathcal{S}_3 \end{pmatrix} = \mathsf{K}_o \begin{pmatrix} S_r \\ S_g \\ S_b \end{pmatrix} = \begin{pmatrix} 0.887 & 0.461 & 0.0009 \\ -0.46 & 0.88 & 0.01 \\ 0.004 & -0.01 & 0.99 \end{pmatrix} \begin{pmatrix} S_r \\ S_g \\ S_b \end{pmatrix} \tag{2.62}$$

The first component is roughly the achromatic gray scale input, the second for red-green opponent channel and the third roughly for blue-yellow opponency. For explicit notations, we also denote the components $\mathcal{S}_k$ by index $k =$ (Lum, RG, BY).

The signal powers in the three channels have the following ratio,

$$\langle \mathcal{S}^2_{Lum}\rangle : \langle \mathcal{S}^2_{RG}\rangle : \langle \mathcal{S}^2_{BY}\rangle = 97 : 2.8 : 0.015. \tag{2.63}$$

The simplifying assumption $\langle S(\lambda_1)S(\lambda_2)\rangle = \delta(\lambda_1 - \lambda_2)$ is likely to cause the distortions in both the composition of the components $\mathcal{S}_k$ and their relative signal powers $\langle \mathcal{S}_k^2 \rangle$.

Meanwhile, these three components are not unlike the YIQ color transmission scheme used in color TV transmission:

$$\begin{pmatrix} Y \\ I \\ Q \end{pmatrix} = \begin{pmatrix} +0.299 & +0.587 & +0.144 \\ +0.596 & -0.274 & -0.322 \\ +0.211 & -0.523 & +0.312 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.64}$$

where R, G, B used in the color TV for red, green, blue colors by the camera maybe roughly identified with our cones input $S_r, S_g, S_b$. The first component Y is the achromatic channel corresponding to gray scale in the black-and-white TV. The second I and third Q components are the chromatic channels. In color TV, a typical distribution of a given image is that Y contains 93% of the signal energy, I contains about 5% and Q about 2%. These values, obtained from TV images, can be seem as manifesting the input statistics $\langle S_i S_j \rangle$, and suggest that the signal power in the chromatic channels are not as weak as suggested in equation (2.63).

Perceptual color distortions after color adaptation can also be understood from the coding changes, in both the compositions and gains $g_\pm$ of the luminance and chromatic channels, induced by changes in input statistics (specifically in correlations, e.g., $\langle S_r S_g \rangle$, Atick et al 1993).

### 2.5.4 Coupling space and color coding in retina

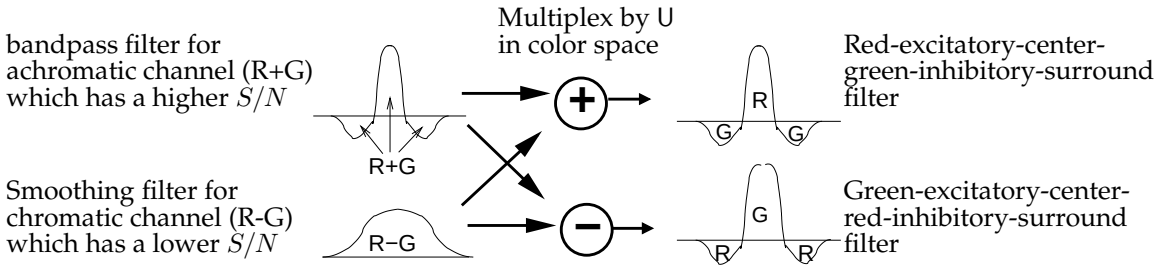



Figure 2.13: Coupling coding in space and color (only red (R) and green (G) for simplicity). Multiplexing the center-surround, contrast enhancing, achromatic (R+G) filter with the input smoothing chromatic (R-G) filter gives, e.g., a red-center-green-surround double (in space and in color) opponency RF observed in retina. All filters are shown in its 1-d profile $\mathsf{K}(x)$, with horizontal axis masking 1-dimensional space $x$ the and vertical axis masking the magnitude $\mathsf{K}(x)$ at location $x$. The markings $R$, $G$, $R+G$, and $R-G$ indicates the cone selectivity of the filter $\mathsf{K}(x)$ at particular spatial locations $x$.

Physiologically, color and space codings are coupled in, e.g., the red-center-green-surround double opponent RFs (Fig. (2.13)) of the retinal ganglion cells. This can be understood as follows.[4] Ignoring the temporal and stereo input dimension, visual inputs $\mathbf{S}$ depends both on space $x$ and the input sensory cone type $c = (r, g, b)$. Hence,

$$\mathbf{S} = (S_r(x), S_g(x), S_b(x))^T.$$

Meanwhile, the output responses $\mathbf{O}$ should be

$$\begin{pmatrix} O_1(x) \\ O_2(x) \\ O_3(x) \end{pmatrix} = \sum_{x'} \begin{pmatrix} \mathsf{K}_{1r}(x,x') & \mathsf{K}_{1g}(x,x') & \mathsf{K}_{1b}(x,x') \\ \mathsf{K}_{2r}(x,x') & \mathsf{K}_{2g}(x,x') & \mathsf{K}_{2b}(x,x') \\ \mathsf{K}_{3r}(x,x') & \mathsf{K}_{3g}(x,x') & \mathsf{K}_{3b}(x,x') \end{pmatrix} \begin{pmatrix} S_r(x') \\ S_g(x') \\ S_b(x') \end{pmatrix} \tag{2.65}$$

The input correlation matrix $R^S$ is

$$R^S = \begin{pmatrix} R^S_{rr}(x_1,x_2) & R^S_{rg}(x_1,x_2) & R^S_{rb}(x_1,x_2) \\ R^S_{gr}(x_1,x_2) & R^S_{gg}(x_1,x_2) & R^S_{gb}(x_1,x_2) \\ R^S_{br}(x_1,x_2) & R^S_{bg}(x_1,x_2) & R^S_{bb}(x_1,x_2) \end{pmatrix} \tag{2.66}$$

where

$$R^{cc'}(x_1,x_2) = \langle S^c(x_1) S^{c'}(x_1) \rangle$$

for cone types $c, c' = r, g, b$. As before, we expect translation invariance in space, thus $R^S_{cc'}(x_1,x_2) = R^S_{cc'}(x_1 - x_2)$. A simple assumption, confirmed by measurements,[80] is that the correlation $R^S$ is separable into a cross product of correlations in spatial and chromatic dimensions:

$$R^S == R^S(x) \otimes R^S(c) \equiv R^S(x) \begin{pmatrix} R^S_{rr} & R^S_{rg} & R^S_{rb} \\ R^S_{gr} & R^S_{gg} & R^S_{gb} \\ R^S_{br} & R^S_{bg} & R^S_{bb} \end{pmatrix} \tag{2.67}$$

Here $R^S(x)$ describes the spatial correlation as in section (2.5.1), while $R^S(c)$, the $3 \times 3$ matrix $R^S_{cc'}$ describes the cone correlations as in section (2.5.3).

Consequently, we may think of input signal $\mathbf{S}$ as composed of three parallel channels of spatial inputs

$$\mathcal{S}_{lum}(x), \mathcal{S}_{RG}(x), \mathcal{S}_{BY}(x)$$

for three decorrelated channels, Lum, RG, and BY, in the color dimension. Each of these channels of spatial inputs can have its efficient spatial coding as described in section (2.5.1). From what we learned for the spatial coding, the stronger luminance channel $\mathcal{S}_{lum}$ requires a center-surround or band pass spatial filter $\mathsf{K}_{Lum}(x)$ to enhance image contrast, while the weaker chromatic channels $\mathcal{S}_{RG}$ and $\mathcal{S}_{BY}$ requires spatial smoothing filters $\mathsf{K}_{RG}(x)$ and $\mathsf{K}_{BY}(x)$ to average out noise (thus color vision has a lower spatial resolution). Multiplexing the luminance channel with the RG channel for instance by rotation $\mathsf{U}$ in the color space, analogous to eq. (2.48) for stereo vision,

$$\begin{pmatrix} \mathsf{K}_1(x) \\ \mathsf{K}_2(x) \\ \mathsf{K}_3(x) \end{pmatrix} = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathsf{K}_{lum}(x) \\ \mathsf{K}_{RG}(x) \\ \mathsf{K}_{BY}(x) \end{pmatrix} \tag{2.68}$$

leads to addition or subtraction of the two filters, $\mathsf{K}_{lum}(x)$ and $\mathsf{K}_{RG}(x)$, as illustrated in Fig. (2.13), giving the red-center-green-surround or green-center-red-surround RFs.

The intuitive solution above can be more formally obtained as follows. The eigenvectors of the full input correlation $R^S$ in equation (2.67), as the cross product of $R^S(x)$ and $R^S(c)$, are also cross products of those in the respective dimensions:

$$(k,\kappa)^{th} \text{ eigenvector of } R^S \text{ is } \propto e^{ikx} \begin{pmatrix} S^\kappa_r \\ S^\kappa_g \\ S^\kappa_b \end{pmatrix}$$

where $k$ index the eigenvector $e^{ikx}$ in the space, and $\kappa$ = LUM, RG, BY index the eigenvector $(S^\kappa_r, S^\kappa_g, S^\kappa_b)^T$ in the chromatic dimension. The $\mathsf{K}_o$ for principal component transform is also a cross product of those, $\mathsf{K}_o(x)$ and $\mathsf{K}_o(c)$, in the respective dimensions:

$$\mathsf{K}_o = \mathsf{K}_o(x) \otimes \mathsf{K}_o(c), \quad \text{such that} \quad [\mathsf{K}_o]_{k,\kappa,x,c} \sim e^{-ikx} S^\kappa_c. \tag{2.69}$$

The mean power of the principal component $(k,\kappa)$ is also a product

$$\langle |\mathcal{S}_{k,\kappa}|^2 \rangle = \lambda^S_k \lambda^S_\kappa \sim \frac{1}{k^2} \cdot \langle |\mathcal{S}_\kappa|^2 \rangle$$

where $\lambda^S_k \sim 1/k^2$ and $\lambda^S_\kappa = \langle |\mathcal{S}_\kappa|^2 \rangle$ are the eigenvalues of $R^S(x)$ and $R^S(c)$ respectively. From this signal power (and thus a signal-to-noise ratio given noise power), the gain $g_{k,\kappa}$ can be obtained from equation (2.41). If we choose the $\mathsf{U}$ as

$$\mathsf{U} = \mathsf{U}(c) \otimes \mathsf{U}(x), \quad \text{such that,} \quad \mathsf{U}_{x,a,k,\kappa} = [\mathsf{U}(x)]_{xk} [\mathsf{U}(c)]_{a,\kappa} \sim e^{ikx} [\mathsf{U}(c)]_{a,\kappa}, \tag{2.70}$$

where $\mathsf{U}(c)$ is an unitary transform in color dimension, then,

$$\mathsf{K} = \mathsf{U}(c) \otimes \mathsf{U}(x) \times \mathsf{g} \times \mathsf{K}_o(x) \otimes \mathsf{K}_o(c). \tag{2.71}$$

In the format of

$$O_i(x) = \sum_{j=r,g,b} \sum_{x'} \mathsf{K}_{ij}(x - x')\mathbf{S}_j(x') + \text{noise}, \tag{2.72}$$

$$\mathsf{K} = \begin{pmatrix} \mathsf{K}_{1r}(x) & \mathsf{K}_{1g}(x) & \mathsf{K}_{1b}(x) \\ \mathsf{K}_{2r}(x) & \mathsf{K}_{2g}(x) & \mathsf{K}_{2b}(x) \\ \mathsf{K}_{3r}(x) & \mathsf{K}_{3g}(x) & \mathsf{K}_{3b}(x) \end{pmatrix} = \mathsf{U}(c) \times \begin{pmatrix} \mathsf{K}_{lum}(x) & 0 & 0 \\ 0 & \mathsf{K}_{RG}(x) & 0 \\ 0 & 0 & \mathsf{K}_{BY}(x) \end{pmatrix} \times \mathsf{K}_o(c) \tag{2.73}$$

where

$$\mathsf{K}_\kappa = \mathsf{U}(x) \times \mathsf{g}^\kappa \times \mathsf{K}_o(x)$$

in which $\mathsf{g}^\kappa$ is a diagonal matrix with diagonal elements $\mathsf{g}^\kappa_{kk} = g_{k,\kappa}$.

### 2.5.5 Efficient Spatial Coding in V1

Primary visual cortex receives the retinal outputs via LGN. V1 RFs are orientation selective and shaped like small (Gabor) bars or edges. Different RFs have different orientations and sizes (or are tuned to different spatial frequencies), in a multiscale fashion (also called wavelet coding (Daubechies 1992)) such that RFs of different sizes are roughly scaled versions of each other. Fig (2.14) show examples of RFs preferring a vertically oriented bar, a vertical edge, a right tilted bar, and a smaller, left tilted edge.

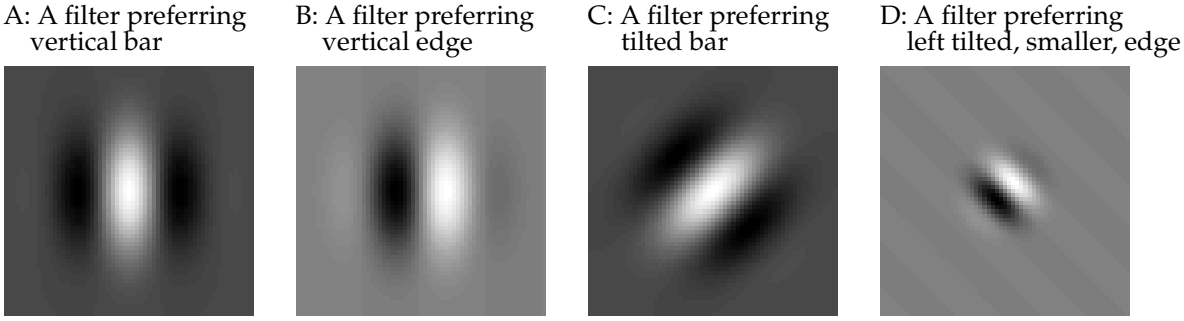



Figure 2.14: Illustration of oriented, multiscale, spatial filters like that in V1 neurons.

We postpone until later (see chapter (3)) our discussion regarding why V1 may choose different RFs from those in the retina. Here, we show how these V1 RFs can be understood in our efficient coding formulation. These RFs can again be seen as components of an optimal code using a particular form of rotation (unitary) matrix $\mathsf{U}$ in the coding transform $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$. As we can imagine, different $\mathsf{U}$ should lead to different $\mathsf{K}$ and thus different RFs. This can be shown in the following two examples using two different $\mathsf{U}$'s.

One is the case $\mathsf{U} = \mathbb{I}$, an identity matrix, such that $\mathsf{U}_{ij} = \delta_{ij}$. Then, $\mathsf{K} = \mathsf{g}\mathsf{K}_o$, and the $k^{th}$ RF, as the $k^{th}$ row vector of $\mathsf{K}$ to give output $O_k = \sum_x \mathsf{K}_{kx} S_x$ for the $k^{th}$ output neuron, is

$$\mathsf{K}_{kx} = (\mathsf{g}\mathsf{K}_o)_{kx} = \frac{1}{\sqrt{N}} g_k e^{-ikx} \tag{2.74}$$

where $N$ is the total number of input (spatial) nodes. This RF is spatially global since its value is non-zero at all spatial locations $x$. It is shaped like a (infinitely large) Fourier wave. Indeed, the response of a neuron with such a receptive field is

$$O_k = \sum_x \mathsf{K}_{kx} S_x = g_k \frac{1}{\sqrt{N}} \sum_x e^{-ikx} S_x = g(k)\mathcal{S}_k \tag{2.75}$$

which is proportional to the Fourier component of input $\mathbf{s}$. So the encoding $\mathsf{K}$ Fourier transforms the inputs, and adds gain control $g_k$ to each Fourier component. Each output neuron can be indexed by $k$ for the unique input frequency $k$ to which this neuron is sensitive to. This neuron does not respond to inputs of any other frequency, no matter how close the frequency is to its preferred value. In other words, the neuron is infinitelly tuned to frequency. Meanwhile, such a coding has no spatial selectivity to inputs, would require very long and massive neural connections to connect each output neuron to inputs from all input locations $x$, and the receptive fields for different neurons $k$ have different shapes (i.e., frequencies). Apparently, our visual system did not choose such a coding, as there is no evidence for global Fourier wave receptive fields with zero frequency tuning width.

Another example is $\mathsf{U} = \mathsf{K}_o^{-1}$ used in section (2.5.1) to construct the RFs for the retinal filter. In detail, this $\mathsf{U}$ takes the form,

$$\mathsf{U} = \frac{1}{\sqrt{N}} \begin{pmatrix} e^{ik_1x_1} & e^{ik_2x_1} & ... & e^{ik_nx_1} & ... \\ e^{ik_1x_2} & e^{ik_2x_2} & ... & e^{ik_nx_2} & ... \\ ... & & & ... & ... \\ e^{ik_1x_m} & e^{ik_2x_m} & ... & e^{ik_nx_m} & ... \\ ... & & & ... & ... \end{pmatrix} \tag{2.76}$$

to give a $\mathsf{K}$ such that

$$\mathsf{K}_{x,x'} = (\mathsf{U}\mathsf{g}\mathsf{K}_o)_{x,x'} = \frac{1}{\sqrt{N}} \sum_k e^{ikx} (\frac{1}{\sqrt{N}} g_k e^{-ikx'}) \tag{2.77}$$

$$= \frac{1}{N} \sum_k g_k e^{ik(x-x')} \tag{2.78}$$

The right hand side of equation (2.77) indicates that this filter is a weighted sum of all the Fourier wave filters $\frac{1}{\sqrt{N}} g_k e^{-ikx'}$, with a frequency $k$ specific weight $\frac{1}{\sqrt{N}} e^{ikx}$. Consequently, the resulting filter $\mathsf{K}_{x,x'}$ is sensitivity to all Fourier frequencies with a sensitivity $g_k$. Also, as shown in equation (2.78), the summation weights are such that the component Fourier filters sum constructively at location $x$ and destructively at locations sufficient away from $x$, so that the filter is now spatially localized around location $x$, which is then the center of the corresponding RF. Different filters are indexed by different $x$. Thus different output neurons have the same shape of the RFs, and differ only by their RF center locations. All these neurons have good spatial selectivity but poor frequency selectivity. Each neuron's output multiplexes the inputs from all frequencies.

The two examples of $\mathsf{U}$ above are the two extremes of all possible $\mathsf{U}$'s, one multiplexes no Fourier wave filters and gives RFs of no spatial selectivity, and the other multiplexes all Fourier wave filters and gives RFs of good spatial selectivity. The $\mathsf{U}$ transform that can account for the multiscale, orientation tuned, RFs in Fig. (2.14) is inbetween these two extremes. It multiplexes the Fourier wave filters within a finite frquency range

$$\mathbf{k} \in \pm(\mathbf{k_1^s}, \mathbf{k_2^s}) \tag{2.79}$$

so that the RF

$$\mathsf{K}^s(x - x') \sim \sum_{k \in \pm(\mathbf{k}_1^s, \mathbf{k}_2^s)} g(k) e^{ik(x'-x)+\phi(x)} \tag{2.80}$$

is responsive only to a restricted range of orientations and the magnitudes of $\mathbf{k}$. Here, a bold-faced $\mathbf{k}$ is used to indicate that $\mathbf{k} = (k_x, k_y)$ is in fact a vector of two components $k_x$ and $k_y$. Hence, the frequency range $\mathbf{k} \in \pm(\mathbf{k_1^s}, \mathbf{k_2^s})$ is not isotropic, i.e., is oriented. Meanwhile, the superscript $s$ in $\mathsf{K}^s$ and $(\mathbf{k_1^s}, \mathbf{k_2^s}$ is an index for the particular range of the magnitude of the frequency and for the orientations of the frequencies. Different cortical cells have different RF center locations $x'$ and frequency/orientation ranges $(\mathbf{k}_1^s, \mathbf{k}_2^s)$ to give a complete sampling (Li and Atick 1994a). The code can be viewed as an intermediate between the Fourier wave code, in which each RF is infinitely large and responds to only one frequency and orientation, and the retinal code, in which each RF is small and responsive to all frequencies $k$ and all orientations (i.e., orientation untuned).

Without diving too mch into the mathematical technicalities available in the reference,[51] the results are presented here in some more details. The U matrix for the multiscale code takes the form of a block diagonal matrix:

$$\mathsf{U} = \begin{bmatrix} \mathsf{U}^{(0)} & & & \\ & \mathsf{U}^{(1)} & & \\ & & \mathsf{U}^{(2)} & \\ & & & \ddots \end{bmatrix}$$

U is such that each sub-matrix $\mathsf{U}^{(s)}$ is itself unitary, and is concerned with the finite frequency range $\mathbf{k} \in \pm(\mathbf{k_1^s}, \mathbf{k_2^s})$ in a form like the U in equation (2.76) as a Fourier inverse transform for the whole frequency range. Hence, at the inter-block level, U is like an identity matrix that does no multiplexing between different frequency ranges. At the intra-block level, $\mathsf{U}^{(s)}$ multiplexes all frequency filters $\sim e^{ikx}$ within the frequency range $\mathbf{k} \in \pm(\mathbf{k_1^s}, \mathbf{k_2^s})$. Specifically,[51]

$$\mathsf{U}_{nk}^{(s)} = \begin{cases} \frac{1}{\sqrt{N^{(s)}}} e^{i(-\phi^{(s)}\ n + k x_n^s + \theta)} & \text{if } k > 0 \\ \\ \frac{1}{\sqrt{N^s}} e^{-i(-\phi^s\ n + |k| x_n^{(s)} + \theta)} & \text{if } k < 0 \end{cases} \tag{2.81}$$

where $\theta$ is an arbitrary phase which can be thought of as zero for simplicity at the moment, $N^{(s)}$ is the number of neurons or frequencies in the block $s$, and

$$\phi^{(s)} = \frac{p}{q}\pi, \tag{2.82}$$

for two relatively prime integers $p$ and $q$. So for the $n^{th}$ neuron selective to the this frequency range, its response is:

$$O_n^{(s)} = \frac{1}{\sqrt{N N^{(s)}}} \sum_x [\sum_{\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})} g_k \cos(k(x_n - x) + \phi^{(s)} n + \theta)] S_x \tag{2.83}$$

$$\equiv \sum_x \mathsf{K}^{(s,n)}(x - x') S_x \tag{2.84}$$

with a receptive field centered at the lattice location

$$x(s)_n = (N/N^{(s)})n \tag{2.85}$$

and tuned to frequency (and orientation) range $\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})$, and has a receptive field phase of

$$\phi^{(s)}(n) \equiv \phi^{(s)} n + \theta \tag{2.86}$$

that changes from cell to cell within this frequency tuning band. In particular, different cell $n$ within this band can have different RF phases $\phi^{(s)}(n)$ or shapes, and there are all together $q$ different RF types. When $q = 2$ and $p = 1$, we have quadrature phase relationship between RFs of the two neighboring cells $n$ and $n+1$. This particular requirement on $\phi^{(s)}(n)$, and thus $p$ and $q$, is the result of requiring U to be unitary. The particular choice of $p = 1$ and $q = 2$ also correspond to a choice on the frequency bandwidth of $\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})$, making the bandwidth in octaves as

$$\log_2[(p+q)/p] \approx 1.5\text{octave} \tag{2.87}$$

close to that of frequency tuning width of the V1 cells.

In the same way that coupling color coding with spatial coding gives the red-center-green-surround retinal ganglion cells, coupling coding in space with coding in stereo, color, and time gives the varieties of V1 cells, such as double opponent color-tuned cells (Li and Atick 1994a), direction selective cells (Li 1996, van Hateren and Ruderman 1998), and disparity selective cells (Li and Atick 1994b). It leads also to correlations between selectivities to different feature dimensions within a cell, e.g., cells tuned to color are tuned to lower spatial frequencies. Many of these correlations, analyzed in detail in (Li and Atick 1994ab, Li 1995, 1996), are interesting and illustrative (not elaborated here because of space) and provide many testable predictions. For instance, Li and Atick (1994b) predicted that cells tuned to horizontal (than vertical) orientation are more likely binocular when they are tuned to medium-high spatial frequencies, as subsequently confirmed in single cell and optimal imaging data (Zhaoping et al 2006). Similarly, the predicted poor sensitivity to color and motion combination (Li 1996) has also been observed (Horwitz and Albright 2005).

# Chapter 3

# V1 and information coding

So far, the efficient coding principle seems to account for not only RF properties for retinal cells, but also for the vast diversity of RF properties in V1: tuning to orientation, color, ocularity, disparity, motion direction, scale, and the correlations between these tunings in individual cells. This suggests that the principle of data compression by efficient coding, with minimal information loss, may progress from retina to V1. However, this section discusses two large problems with this argument: (1) there is no quantitative demonstration that V1 significantly improves coding efficiency over retina; and no apparent bit rate bottleneck after the optic nerve; and (2) efficient coding has difficulty in explaining some major aspects of V1 processing.

If one approximates all signals as Gaussian, the V1 cortical code is no more efficient than the retinal code, in terms of information bits transmitted and the cost of neural power, since they both belong to the set of degenerate optimal solutions of $\partial E/\partial \mathsf{K} = 0$. Is the cortical code more efficient due to the higher order input statistics beyond the Gaussian approximation of $P(\mathbf{S})$ (that breaks the degeneracy of the optimal codes)? If so, bar stereo, why isn't it adopted by the retina? In fact, it has been shown that the dominant form of visual input redundancy (in terms of entropy bits) arises from second order rather than higher order input statistics, e.g., correlation between three pixels beyond that predicted from second order statistics (Schreiber 1956, Li and Atick 1994a, Petrov and Zhaoping 2003). This motivated a hypothesis that the V1's multiscale coding serves the additional goal of translation and scale invariance (Li and Atick 1994a) to facilitate object recognition presumably occurring only beyond retina. However, this does not explain the even more puzzling fact of a 100 fold expansion from retina to V1 in the number of neurons (Barlow 1981) to give a hugely overcomplete representation of inputs. For instance, to represent input orientation completely at a particular spatial location and scale, only three neurons tuned to three different orientations would be sufficient (Freeman and Adelson 1991). However, many more V1 cells tuned to many different orientations are actually used. It is thus highly unlikely that the neighboring V1 neurons have decorrelated outputs, even considering the nonlinearity in the actual receptor-to-V1 transform. This contradicts the goal of efficient coding of reducing redundancy and revealing the independent entities in high S/N. Nor does such an expansion improve signal recovery at low S/N ratios since no retina-to-V1 transform could generate new information beyond that available at retina. It has been argued that such an expansion can make the code even sparser (Olshausen and Field 1997, Simoncelli and Olshausen 2001), making each neuron silent for most inputs except for very specific input features. Indeed, $M = 10^6$ bits/second of information, transmitted by $M$ retina ganglions at 1 bits/second by each neuron, could be transmitted by $100M$ V1 neurons at $0.01$ bits/second each (Nadal and Parga 1993), if, e.g., each V1 neuron is much less active with a higher neural firing threshold. Such a sparser V1 representation however gains no coding efficiency. There is yet no reliable quantitative measure of the change in efficiency or data rate by the V1 representation. It would be helpful to have quantitative analysis regarding how this representation sufficiently exposes the underlying cognitive (putatively independent) components to justify the cost of vastly more neurons. Minimizing energy consumption in neural signaling has also been proposed to account for sparser coding (Levy and Baxter 1996, Lennie 2003), possibly

favoring overcompleteness.

As argued in section (2.3), the sparse coding formulation (Olshausen and Field 1997) is an alternative formulation of the same efficient coding principle. Hence, those V1 facts puzzling for efficient coding are equally so for the sparse coding formulation, whose simulations typically generate representations much less overcomplete than that in V1 (Simoncelli and Olshausen 2001). Often (e.g., Bell and Sejnowski 1997), kurtosis (defined as $\langle x^4\rangle/\langle x^2\rangle^2 - 3$ for any probability distribution $P(x)$ of a random variable $x$) of response probabilities $P(\mathbf{O})$ is used to demonstrate that visual input is highly non-Gaussian, and that the responses from a filter resembling a V1 RF have higher kurtosis (and are thus sparser) than those from a center-surround filter resembling a retinal RF. However, one needs to caution that a large difference in kurtosis is only a small difference in entropy bits. For instance, two probability distributions $P_1(x) \propto e^{-x^2/2}$ and $P_2(x) \propto e^{-|x/0.1939|^{0.6}}$ of equal variance $\langle x^2\rangle$ have differential entropies 2 and 1.63 bits, respectively, but kurtosis values of 0 and 12.6, respectively. While Fig. (3.1) demonstrates that higher order statistics (redundancy) causes much or most of the relevant visual perception of object forms, this perception is after the massively lossy visual selection (beyond efficient coding) through the attentional bottleneck.

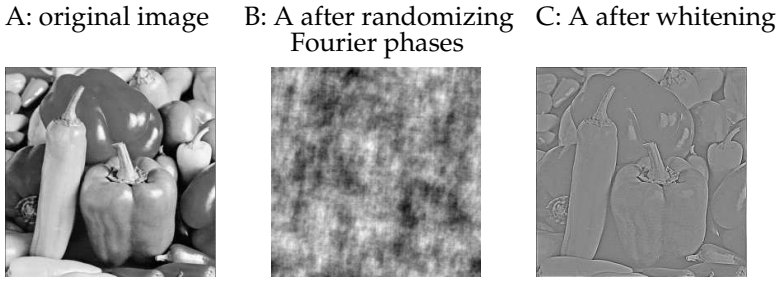



Figure 3.1: An original image in A becomes meaningless when the phases of its Fourier transform are replaced by random numbers, shown in B (After Field 1989). Hence, A and B have the same first and second order statistics characterized by their common Fourier powers $\mathcal{S}_k^2 \sim 1/k^2$, but B has no higher order statistics. Whitening A eliminates the second order correlation in C, but preserves the meaningful form information in the higher order statistics.

For discussion, we divert in this paragraph from the processing goal of data reduction. First, from the perspective of form perception, the redundancy in the higher order statistics (Fig. (3.1)) should be kept, while that in the lower order statistics (which is useless for form perception) should be removed. Second, the sparse coding formulation (Olshausen and Field 1997) also motivated a generative model of visual inputs $\mathbf{S}$ by causes $\mathsf{K}^{-1}$ with amplitudes $\mathbf{O}$ (see section (2.3)). It was argued that overcomplete representations allow more and even non-independent causes, so that some causes can explain away others given any inputs. For instance, a bar oriented at $0^o$ could be best generated by a cause (basis function) of $0^o$ but not of $5^o$, thus the response amplitude $O_i$ for $0^o$ should explain away another $O_{i'}$ for $5^o$, i.e., $O_i \gg O_{i'}$ (Olshausen and Field 1997). This would however require a severe nonlinearity in responses that, e.g., orientation tuning curves would be much narrower than those of V1 RFs. While generative models for vision are expected to be very helpful to understand top-down effects in higher level vision and their top-down feedbacks to V1, they are beyond our scope here and our current knowledge about V1.

Additional difficulties for the coding theories arise from observations made since the 1970's that stimuli in the context outside a neuron's RF significantly modulate its response in a complex manner (Allman et al 1985). For instance, a neuron's response to an optimally oriented bar within its RF can be suppressed by up to $80\%$ when there are surrounding bars of similar orientations outside the RF (Knierim and Van Essen 1992, Sillito et al 1995, Nothdurft et al 1999). This is called iso-orientation suppression. The contextual suppression is weaker when the surrounding bars are randomly oriented, and weakest when they are oriented orthogonally to the bar within the RF. Meanwhile, the response to a weak contrast bar within the RF can be enhanced by up to 3-4 fold when contextual bars are aligned with this bar, as if they are segments of a smooth contour —

i.e., colinear facilitation (Kapadia et al 1995). The horizontal intra-cortical connections (Gilbert and Wiesel 1983, Rockland and Lund 1983), linking nearby cells with overlapping or non-overlapping classical receptive fields (CRFs), are plausible neural substrates mediating the contextual influences. These observations seem like nuisances to the classical view of local feature detectors, or CRFs, and were not taken very seriously immediately, partly due to a lack of theoretical frameworks to understand them. Contextual suppressions maybe viewed as additional mechanisms for redundancy reduction (Rao and Ballard 1999, Schwartz and Simoncelli 2001), leaving contextual facilitation and the neural proliferation still unaccounted for.

To an animal, one bit of information about visual object identity typically has a very different relevance from another bit of information on light luminance. Information Theory can quantify the *amount* of information, and thereby help the design of optimal codes for information *transmission*, a likely goal for the retina. However, it does not assess the *meaning* of information to design optimal representations for information *discrimination or selection (or discarding)*. Information selection and distortion is a critical concern of the cortex that requires losing rather than preserving Shannon Information. Rather than being a nuisance for a classical coding view, intra-cortical interactions can be a wonderful means of implementing other goals. V1, the largest visual area in the brain, equipped with additional neural mechanisms unavailable to retina, ought to be doing important cognitive tasks beyond information transmission. One of the most important and challenging visual task is segmentation, much of it involves selection. To understand V1, we thus turn to the second data reduction strategy for early vision (see section (1)), namely to build a representation that facilitate bottom up visual selection.

# Chapter 4

# The V1 hypothesis — creating a bottom up saliency map for pre-attentive selection and segmentation

At its heart, vision is a problem of object recognition and localization for (eventually) motor responses. However, before this end comes the critical task of input selection of a limited aspects of input for detailed processing by the attentional bottleneck. As discussed in section 1, it is computationally efficient to carry out much of this selection quickly and by bottom up mechanisms by directing attention to restricted visual space. Towards this goal, it has been recently proposed that (Li 1999ab, 2002, Zhaoping 2005) V1 creates a bottom up saliency map of visual space, such that a location with a higher scalar value in this map is more likely to be selected for further visual processing, i.e., to be salient and attract attention. The saliency values are represented by the firing rates $\mathbf{O} = (O_1, O_2, ..., O_M)$ of the V1 neurons, such that the RF location of the most active V1 cell is most likely to be selected, regardless of the input feature tunings of the V1 neurons. Let $(x_1, x_2, ..., x_M)$ denote the RF locations of the V1 cells, the most salient location is then $\hat{x} = x_{\hat{i}}$ where $\hat{i} = \text{argmax}_i O_i$. This means $\hat{x} = \text{argmax}_x(\text{max}_{x_i=x} O_i)$, where $x_i = x$ means that the RF of the $i^{th}$ cell covers location $x$, and the saliency map, SMAP(x), is

$$\text{SMAP}(x) \propto \text{max}_{x_i=x} O_i, \tag{4.1}$$

Hence, the saliency value at each location $x$ is determined by the maximum response to that location. So for instance, a red-vertical bar excites a cell tuned to red color, another cell to vertical orientation, and other cells to various features. Its saliency may be signaled by the response of the red tuned cell alone if this is the maximum response from all cells at that location. Algorithmically, selection of $\hat{x} = x_{\hat{i}}$ does not require this maximum operation at each location, but only a single maximum operation $\hat{i} = \text{argmax}_i O_i$ over all neurons $i$ regardless of their RF locations or preferred input features. This is algorithmically perhaps the simplest possible operation to read a saliency map, and can thus be performed very quickly — essential for bottom up selection. An alternative rule $\text{SMAP}(x) \propto \sum_{x_i=x} O_i$ for saliency would be more complex to execute. It would require an additional, non-trivial, processing to group responses $O_i$, from neurons with overlapping but most likely non-identical RF spans, according to whether they are evoked by the same or different input items around the same location, in order to sum them up. V1's saliency output is perhaps read by (at least) the superior colliculus (Tehovnik et al 2003) which receive inputs from V1 and directs gaze (and thus attention). The maximum operation is thus likely performed within the read out area.

The overcomplete representation of inputs in V1, puzzling in the efficient coding framework, greatly facilitates fast bottom up selection by V1 outputs (Zhaoping 2006). For instance, having

many different cells tuned to many different orientations (or features in general) near the same location, the V1 representation $\mathbf{O}$ helps to ensure that there is always a cell $O_i$ at each location to *explicitly* signal the saliency value of this location if the saliency is due to an input orientation (feature) close to any of these orientations (or features), rather than having it signalled *implicitly* by activities of a group of neurons (and thus disabling the simple maximum operation $\hat{i} = \text{argmax}_i O_i$ to locate it)[1]. It is apparent that V1's overcomplete representation should also be useful for other computational goals which could also be served by V1. Indeed, V1 also sends its outputs to higher visual areas for operations, e.g., recognition and learning, beyond selection. Within the scope of this paper, I do not elaborate further our poor understanding of what constitutes the best V1 representation for computing saliency as well as serving other goals.

Meanwhile, contextual influences, a nuisance under the classical view of feature detectors, enable the response of a V1 neuron to be context or global input dependent. This is necessary for saliency computations, since, e.g., a vertical bar is salient in a context of horizontal but not vertical bars. The dominant contextual influence in V1 is iso-feature suppression, i.e., nearby neurons tuned to similar features such as orientation and color are linked by (di-synaptic) inhibitory connections (Knierim and Van Essen 1992, Wachtler et al, 2003 Jones et al 2001), and, in particular, iso-orientation suppression. Consider an image containing a vertical bar surrounded by many horizontal bars, and the responses of cells preferring the locations and orientations of the bars. The response to the vertical bar (in a vertical preferring cell) escapes the iso-orientation suppression, while those to the horizontal bars do not since each horizontal bar has iso-orientation neighbors. Hence, the highest V1 response is from the cell responding to the vertical bar, whose location is thus most salient by the V1 hypothesis, and pops out perceptually. By this mechanism, even though the RFs and the intra-cortical connections mediating contextual influences are *local* (i.e., small sized or of a finite range), V1 performs a *global* computation to enable cell responses to reflect context beyond the range of the intra-cortical connections (Li 1998a, 1999a, 2000). Retinal neurons, in contrast, respond in a largely context independent manner, and would not be adequate except perhaps for signalling context independent saliency such as at a bright image spot.

Ignoring eccentricity dependence for simplicity (or consider only a sufficiently small range of eccentricities), we assume that the properties of V1 RFs and intra-cortical interactions are translation invariant, such that, neural response properties to stimulus within its RF are regardless of the RF location, and interaction between two neurons depends on (in addition to their preferred features) the relative rather than absolute RF locations. Then, the V1 responses should be translation invariant when the input is translation invariant, e.g., an image of a regular texture of horizontal bars, or of more general input symmetry such as in an image of a slanted surface of homogeneous texture. However, when the input is not translation invariant, V1 should produce corresponding variabilities in its responses. The contextual influences, in particular iso-feature suppression, are particularly suited to amplify such variances, which are often at salient locations, e.g., at the unique vertical bar among the horizontal bars, or the border between a texture of horizontal bars and another of vertical bars (Li 2000). Therefore, V1 detects and highlights the locations where input symmetry breaks, and saliency could be computationally defined by the degree of such input variance or spatial/temporal symmetry breaking (Li 1998ac, 1999a, 2000). The salient locations of input symmetry breaking typically correspond to boundaries of object surfaces. Since the selection of these locations proposed for V1 is executed before object recognition or classification, it has also been termed as pre-attentive segmentation without classification (Li 1998c, 1999a).

Conditional on the context of background homogeneity, input variance at a texture border or a pop out location is a rare or low probability event. Hence, the saliency definition by the degree

[1] As discussed in Li (1996), V1 could have many different copies $\mathbf{O}^1, \mathbf{O}^2, ...\mathbf{O}^p, ...$ (where superscript $p$ identifies the particular copy) of complete representation of $\mathbf{S}$, such that each copy $\mathbf{O}^p = \mathsf{U}^p\mathsf{g}\mathsf{K}_o\mathbf{S}$ has as many cells (or dimensions) as the input $\mathbf{S}$, and is associated with a particular choice of unitary matrix $\mathsf{U}^p$. Each choice $\mathsf{U}^p$ specifies a particular set of preferred orientations, colors, motion directions, etc. of the resulting RFs whose responses constitute $\mathbf{O}^p$, such that the whole representation $(\mathbf{O}^1, \mathbf{O}^2, ...\mathbf{O}^p, ...)$ covers a whole spectrum of feature selectivities to span these feature dimensions (although the gain matrix $\mathsf{g}$ assigns different sensitivities, some very small, to different feature values and their combinations). In reality, the V1 representation is more like a tight frame of high redundant ratio (Daubechies 1992, Lee 1996, Salinas and Abbott 2000) than a collection of complete representations (from the degenerate class), which would require (Li and Atick 1994a), in addition to the oriented RFs, checker shaped RFs not typically observed physiologically.

of input symmetry breaking is related to the definition of saliency by surprise or novelty (Itti and Baldi 2006, Lewis and Zhaoping 2005). Other definitions of saliency include: a salient location is where an "interest point" detector (for a particular geometric image feature like a corner) signals a hit, or where local (pixel or feature) entropy (i.e., information content) is high (Kadir and Brady 2001). While it can be shown that saliency by novelty and saliency by high local entropy are related, computational definitions of bottom up or general purpose saliency have not yet reached a converging answer.

Given the above limitations, we take the behavioral definition of saliency, and the known V1 mechanisms from physiological and anatomical data, to test the V1 saliency hypothesis by comparing V1 predicted saliencies with the behaviorally measured ones. Saliency has been extensively studied psychophysically using visual search tasks or segmentation tasks ( Treisman and Gelade 1980, Wolfe 1998). The saliency of the target in a visual search task, or the border between regions in a segmentation task, is a measure of the target or border location to attract attention, i.e., be selected, in order to be processed. Thus it can be measured in terms of the reaction time to perform the task. For instance, searching for a vertical bar among horizontal ones, or a red dot among green ones, is fast, with reaction times that are almost independent of the number of distractors (Treisman and Gelade 1980, Julesz 1981). These are called feature search tasks since the target is defined by a unique basic feature, e.g., vertical or red, which is absent in the distractors. In contrast, conjunction search is difficult, for a target defined by a unique conjunction of features, e.g., a red-vertical bar among red-horizonal bars and green-vertical bars (Treisman and Gelade 1980).

In the rest of the section, we will test the V1 hypothesis, through a physiologically based V1 model, to see if saliencies predicted by V1 responses agree with existing behavioral data. This section will then end with analysis to show that the V1 saliency theory, motivated by understanding early vision in terms of information bottlenecks, better agrees with new experimental data than the traditional frameworks of saliency (Treisman and Gelade 1980, Julesz 1981, Wolfe, Case, Franzel 1989, Koch and Ullman 1985, Itti and Koch 2000), which were developed mostly from behavioral data.

## 4.1 Testing the V1 saliency map in a V1 model

We should ideally examine if higher V1 responses predict higher saliencies, namely, behaviorally faster visual selections. Many behavioral data on saliency in terms of the reaction times in visual search and segmentation tasks are available in the literature (Wolfe, 1998). However, physiological data based on stimuli like those in the behavioral experiments are few and far between. Furthermore, to determine the saliency of, say, the location of a visual target, we need to compare its evoked V1 responses to responses to other locations in the scene, since, as hypothesized, the selection process should pick the classical RF of the most active neuron responding to the scene. This would require the simultaneous recordings of many V1 units responding to many locations, a very daunting task with current technology.

We thus resort to the simpler (though incomplete) alternative of simultaneously recording from all neurons in a simulated V1 model (Li, 1999a, Fig. (4.1)). (Such a simplification is, in spirit, not unlike recording under anesthesia *in vivo* or using *in vitro* slices, with many physiological mechanisms and parameters being altered or deleted.) Our model includes only the most relevant parts of V1, namely simplified models of pyramidal cells, interneurons, and intra-cortical connections, in layer 2-3 of V1 (which mediate contextual influences). As a first demonstration of principle, a further simplification was made by omitting input variations in color, time (except to model stimulus onset), stereo, and scale without loss of generality. The neurons are modelled by membrane potentials, e.g., $x_{i\theta}$ and $y_{i,\theta}$ denote the membrane potentials of the pyramidal and interneurons, whose RFs are centered at $i$ (here $i$ denotes location within this V1 model rather than an index for a cell elsewhere in this paper) and oriented at angle $\theta$. Their outputs are modelled by firing rates $g_x(x_{i\theta})$ and $g_y(y_{i\theta})$ which are sigmoid-like functions of the membrane potentials. The equations of

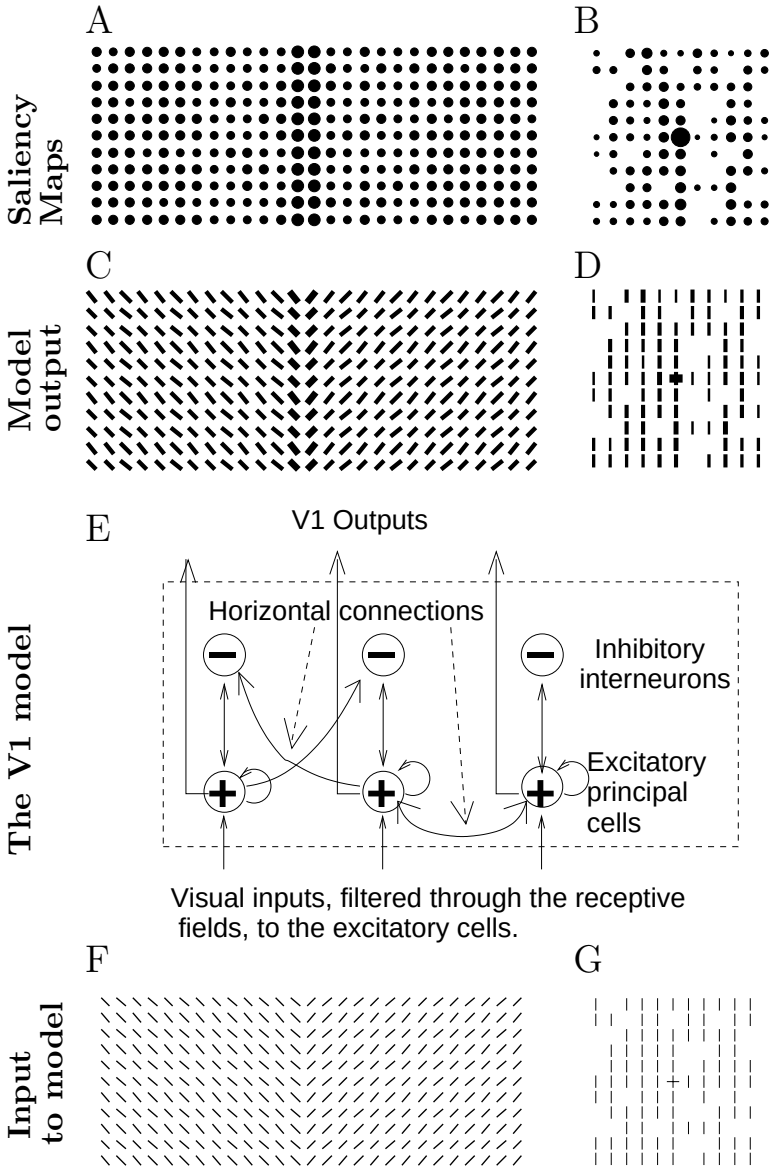



Figure 4.1: The V1 model and its function. The model (E) focuses on the part of V1 responsible for contextual influences: layer 2-3 pyramidal cells, interneurons, and intra-cortical connections. Pyramidal cells and interneurons interact with each other locally and reciprocally. A pyramidal cell can excite other pyramidal cells monosynaptically, or inhibit them disynaptically, by projecting to the relevant inhibitory interneurons. General and local normalization of activities are also included in the model. Shown are also two input images (F, G) to the model, and their output response maps (C,D). The input strengths are determined by the bar's contrast. Each input bar in each example image has the same contrast in these examples. A principal (pyramidal) cell can only receive direct visual input from an input bar in its CRF. The output responses depend on both the input contrasts and the contextual stimuli of each bar due to contextual influences. Each input/output image plotted is only a small part of a large extended input/output image. In many figures in the rest of this paper, the thicknesses of the stimulus or response bars are plotted as proportional to their input/output strengthes for visualization. At top (A, B) are saliency maps where the size of the circle at each location represents the firing rate of the most active cell responding to that visual location. A location is highly salient if its saliency map value has a high $z$ score compared to the values in the background.

motion are

$$\begin{aligned}\dot{x}_{i\theta} &= -\alpha_x x_{i\theta} - g_y(y_{i,\theta}) - \sum_{\Delta\theta\neq 0} \psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) \\ &\quad + J_o g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o \qquad (4.2)\\ \dot{y}_{i\theta} &= -\alpha_y y_{i\theta} + g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c \qquad (4.3)\end{aligned}$$

where $\alpha_x x_{i\theta}$ and $\alpha_y y_{i\theta}$ model the decay to resting potentials, $I_{i\theta}$ model external visual inputs, $I_c$ and $I_o$ model background inputs, including noise and feature unspecific surround suppressions, and the rest of the terms on the right hand side model interactions between neurons for feature specific contextual influences with finite range neural connections like $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ for example. The pyramidal outputs $g_x(x_{i\theta})$ (or their temporal averages) represent the V1 responses. Equations (4.2) and (4.3) specify how the activities are initialized by external inputs and then modified by the contextual influences via the neural connections.

This model (Li 1999a) has a translation invariant structure, such that all neurons of the same type have the same properties, and the neural connections $J_{i\theta,j\theta'}$ (or $W_{i\theta,j\theta'}$) have the same structure from all the pre-synaptic neuron $j\theta'$ except for a translation and rotation to suit $j\theta'$ (Bressloff et al 2002). The dynamics of this model are such that (1) model response does not spontaneously break input translation symmetry when $I_{i\theta}$ is independent of $i$ (otherwise, the model would hallucinate salient locations when there is none); (2) when the inputs are not translation invariant, the model manifests these variant locations by response highlights whose magnitudes reflect, with sufficient sensitivity, the degrees of input variances; and (3) the model reproduces the most relevant physiological observations of neural responses with and without the contextual influences, particularly the phenomena of iso-orientation suppression and colinear facilitation, etc outlined in section (3). Condition (3) ensures that model sufficiently resembles reality to offer reasonable basis for hypothesis testing. Conditions (1) and (2) are computational requirements for saliency computation. The fact that a single set of model parameters can be found to satisfy all three conditions supports the hypothesis that V1 creates a saliency map.

Nonlinear dynamic analysis ensures that this recurrent network of interacting neurons is well behaved in terms of stability and robustness (Li 1999a, 2001, Li and Dayan 1999). It can be shown that equations (4.2) and (4.3) describe a minimal model, which has to include the inhibitory interneurons but not necessarily neural spikes, for the required computation, particularly to satisfy conditions (1) and (2) above simultaneously. The model design and analysis are mathematically challenging and I omit the details (Li 1999a, 2001, Li and Dayan 1999). However, they are not as formidable as simultaneous *in vivo* recordings from hundreds of V1 neurons using visual search stimuli. Following the design and calibration, all model parameters are fixed (as published in Li 1998b, 1999a) for all input stimuli. The saliency of a visual location $i$ is assessed by a z score, $z_i = (S_i - \bar{S})/\sigma$, where $S_i = \max_\theta(g_x(x_{i\theta}))$ (here $S$ links to word "saliency" rather than "signal") is the highest model response to that location, while $\bar{S}$ and $\sigma$ are the mean and standard deviations of the population responses from the active neurons. Obviously, the z score is only used for hypothesis testing and is not calculated by V1.

The model responses to stimuli of visual search agree with human behavior. In orientation feature search in Fig. (4.1 BDG), the target possessing a uniquely oriented bar pops out from distractors of uniformly oriented bars since a neuron responding to the uniquely oriented target bar escapes the strong iso-orientation suppression experienced by neurons responding to the distractor bars. In orientation conjunction search in Fig. (4.2B), the target does not pop out since neurons responding to each of its component bars experience iso-orientation suppression from the contextual input just as neurons responding to a typical distractor bar. A vertical target bar among distractor crosses in Fig. (4.2A) does not pop out since its evoked response is suppressed by the contextual vertical bars in the crosses. This is the basis of typical observations that a target lacking a feature present in distractors does not pop out (Treisman and Gelade 1980). The model shows that visual searches become more difficult when the distractors are less similar as in Fig. (4.2CD), or are less

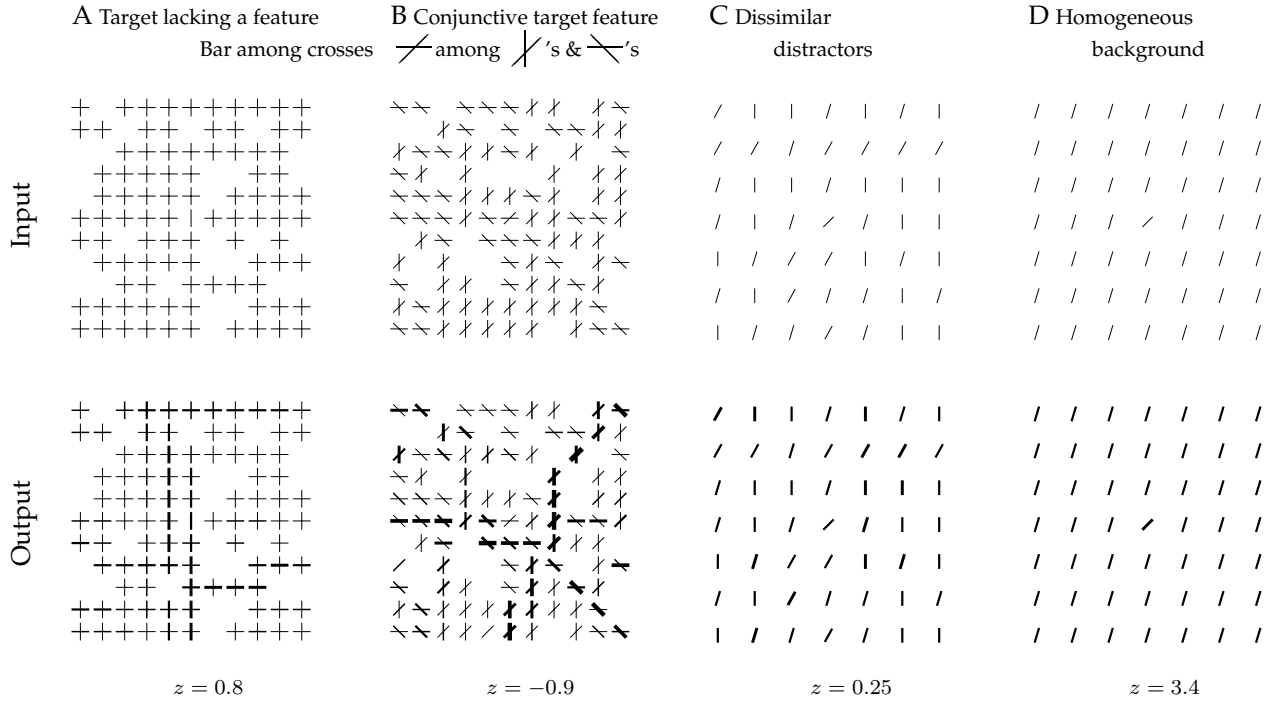



Figure 4.2: Stimulus (top), model responses (bottom), and the $z$ scores for the target (in the center of each pattern), for four examples. A: A difficult search for a target bar lacking a horizontal bar present in distractors. It forms a trivial pair of search asymmetry with Fig. (4.1DG). B: Difficult search for a unique conjunction of orientation features. Searching for a target of a $45^o$ bar among distractors of different orientations $0^o$, $15^o$, or $30^o$ from vertical in C is more difficult than among identical distractor of $15^o$ from vertical in D.

regularly placed in space (Li 2002), as is also known psychophysically (Duncan and Humphreys 1989). This is because (see a related argument (Rubenstein and Sagi 1990)) such stimulus changes increase the variance of surround influences experienced by neurons responding to individual distractors, thereby increasing $\sigma$ and decreasing the target $z$ score.

A more stringent test comes from applying the V1 model to the subtle examples of visual search asymmetry, when the ease of visual search tasks changes slightly upon swapping the target and the distractor. The direction of these slight changes would be difficult to predict much beyond a chance level by an incorrect model or hypothesis without any parameter tuning. Nevertheless, the model predicted (Li 1999b) these directions correctly in all of the five best known examples of asymmetry (Treisman and Gormican 1988) shown in Fig. (4.3).

## 4.2 Psychophysical test of the V1 theory of bottom up saliency

Motivated by understanding early vision in terms of information bottlenecks, the V1 saliency hypothesis has some algorithmically simple but conceptually unusual or unexpected properties which should be experimentally verified. In particular, the saliency of a location is signalled by the most active neuron responding to it regardless of its feature tuning. For instance, the cross among bars in Fig. (4.1G) is salient due to the more responsive neuron to the horizontal bar, and the weaker response of another neuron to the vertical bar is ignored. This means the "less salient features" at any location are invisible to bottom up saliency or selection, even though they are visible to attention attracted to the location by the response to another feature at the same location.

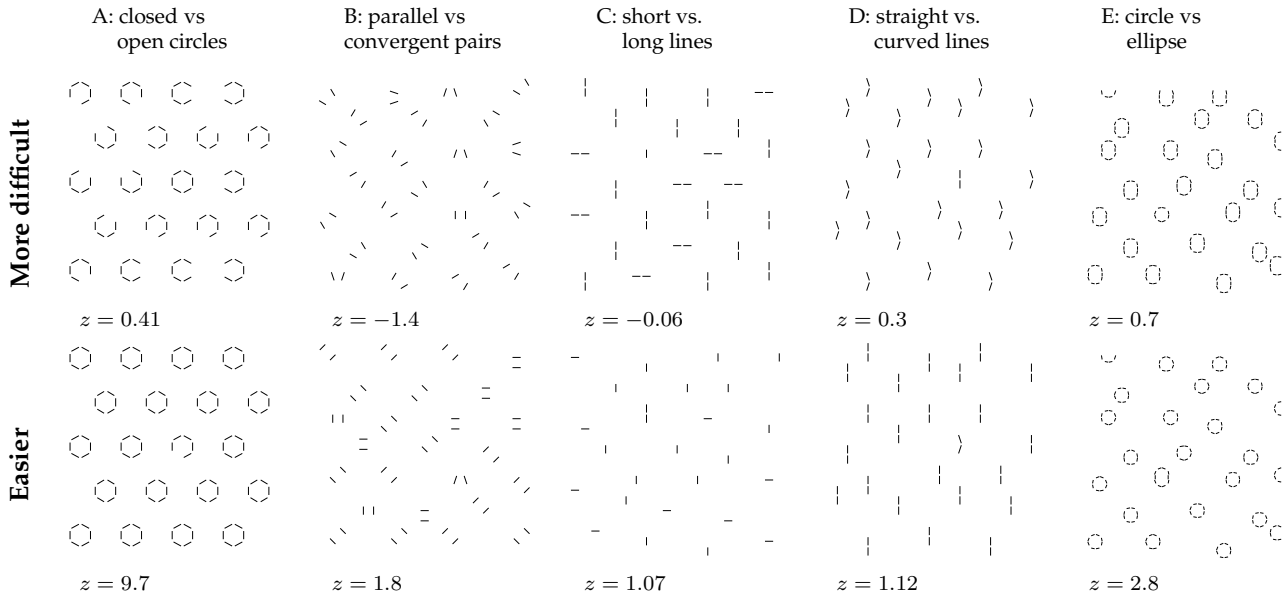



Figure 4.3: Five subtle pairs of the visual search asymmetry studied by Treisman and Gormican (1988) and directionally accounted for by the model. Stimulus patterns are shown with the targets' z scores from the model marked under them.

While this algorithmically simple selection can be easily executed even by a feature blind reader of the saliency map, it seems a waste not to consider the contributions of the "less salient features" to obtain a "better" saliency measure of a location $x$ as the summation $\sum_{x_i=x} O_i$, rather than the maximum $\max_{x_i=x} O_i$, of all responses to this location (see Lewis and Zhaoping (2005) for comparing the two measures based on input statistics). If there is a task in which task relevant features are less salient and "invisible" to bottom up selection by the V1 hypothesis (the maximum rule), the task will be predicted as difficult if saliency plays a significant role, such as in reaction time conditions.

Fig. (4.4) shows texture patterns **a, b, c** that illustrate and test the prediction. Pattern **a** has two iso-orientation textures, activating two populations of neurons, one for left tilt and another for right tilt orientation. Pattern **b** is a uniform texture of a checkerboard of horizontal and vertical bars, evoking responses from another two groups of neurons for horizontal and vertical orientations respectively. With iso-orientation suppression, neurons responding to the texture border bars in pattern **a** are more active than those responding to the background bars; since each border bar has fewer iso-orientation neighbors to exert contextual iso-orientation suppression on the evoked response. For ease of explanation, let us say, the responses from the most active neurons to a border bar and a background bar are $10$ and $5$ spikes/second respectively. This response pattern makes the border location more salient, making texture segmentation easy. Each bar in pattern **b** has as many iso-orientation neighbors as a texture border bar in pattern **a**, hence evokes also a response of $10$ spikes/second. The composite pattern **c**, made by superposing patterns **a** and **b**, activates all neurons responding to patterns **a** and **b**, each neuron responding roughly as it does to **a** or **b** alone (omitting for simplicity any interactions between neurons tuned to different orientations, without changing the conclusion). Now each texture element location evokes the same maximum response of 10 spikes/second, and, by the V1 hypothesis, is as salient (or non-salient) as another location. Hence the V1 theory predicts no saliency highlight at the border, thus texture segmentation is predicted to be much more difficult in **c** than **a**, as is apparent by viewing Fig. (4.4). The task relevant tilted bars are "invisible" to V1 saliency to guide segmentation, while the task irrelevant horizontal and vertical bars interfere with the task.

Note that if saliency of location $x$ were determined by the summation rule $\sum_{x_i=x} O_i$, responses to various orientations at each texture element in pattern **c** could sum to preserve the border high-

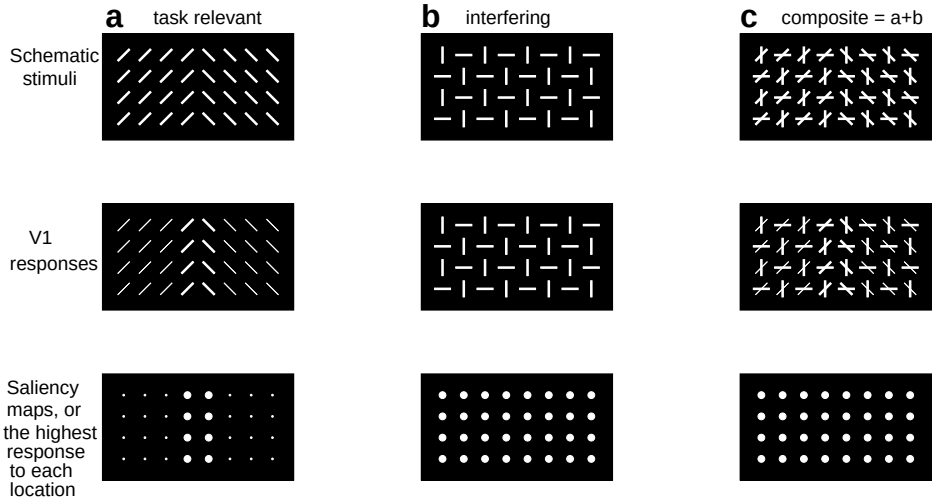
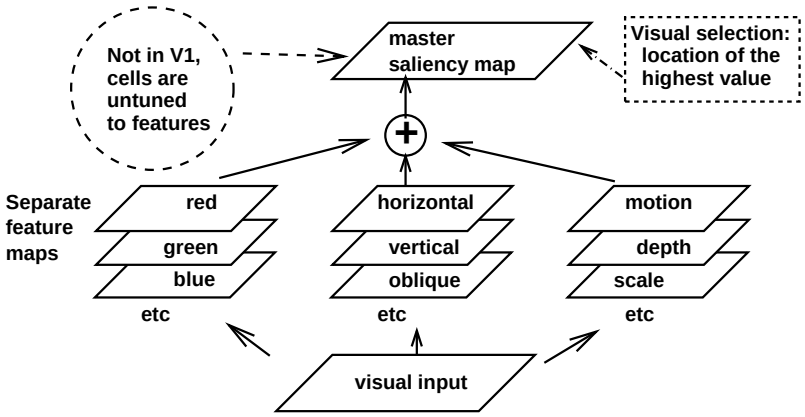



Figure 4.4: Psychophysical test of the V1 saliency hypothesis. **a, b, c**: schematics of texture stimuli (extending continuously in all directions beyond the portions shown), each followed by schematic illustrations of V1's responses and saliency maps, formulated as in Fig. (4.1). Every bar in **b**, or every texture border bar in **a**, has fewer iso-orientation neighbours to induce iso-orientation suppression, thus evoking less suppressed responses. The composite stimulus **c**, made by superposing **a** and **b**, is predicted to be difficult to segment, since the task irrelevant features from **b** interfere with the task relevant features from **a**, giving no saliency highlights to the texture border.

light as 20 spikes/second against a background of 15 spikes/second, thus predicting easy texture segmentation. The V1 theory prediction (by the maximum rule) is confirmed by psychophysically measuring the reaction times of subjects to locate the texture border (Zhaoping and May 2004). Additional data (Zhaoping and May 2004) confirmed other unique predictions from the V1 theory, such as predictions of interference by irrelevant color on orientation based tasks, and predictions of some phenomena of visual grouping due to the anisotropic nature of the contextual influences involving orientation features (arising from combining colinear facilitation with iso-orientation suppression).

The V1 saliency theory bears an interesting relationship with previous, traditional, theories of bottom up saliency (Treisman and Gelade 1980, Julesz 1981, Wolfe, Case, Franzel 1989, most of which also include top-down components). These theories were based mainly on behavioral data, and could be seen as excellent phenomenological models of behavioral saliency. They can be paraphrased as follows (Fig. (4.5A)). Visual inputs are analyzed by separate feature maps, e.g., red feature map, green feature map, vertical, horizontal, left tilt, and right tilt feature maps, etc., in several basic feature dimensions like orientation, color, and motion direction. The activation of each input feature in its feature map decreases roughly with the number of the neighboring input items sharing the same feature. Hence, in an image of a vertical bar among horizontal bars, the vertical bar evokes a higher activation in the vertical feature map than those of each of the many horizontal bars in the horizontal map. The activations in separate feature maps are summed to produce a master saliency map. Accordingly, the vertical bar produces the highest activation at its location in this master map and attracts visual selection. In contrast, a unique red-vertical bar, among red-horizontal and green-vertical bars, does not evoke a higher activation in any one feature

A: Previous theories of bottom up visual saliency map




B: The theory of bottom up saliency map from V1, and its cartoon interpretation

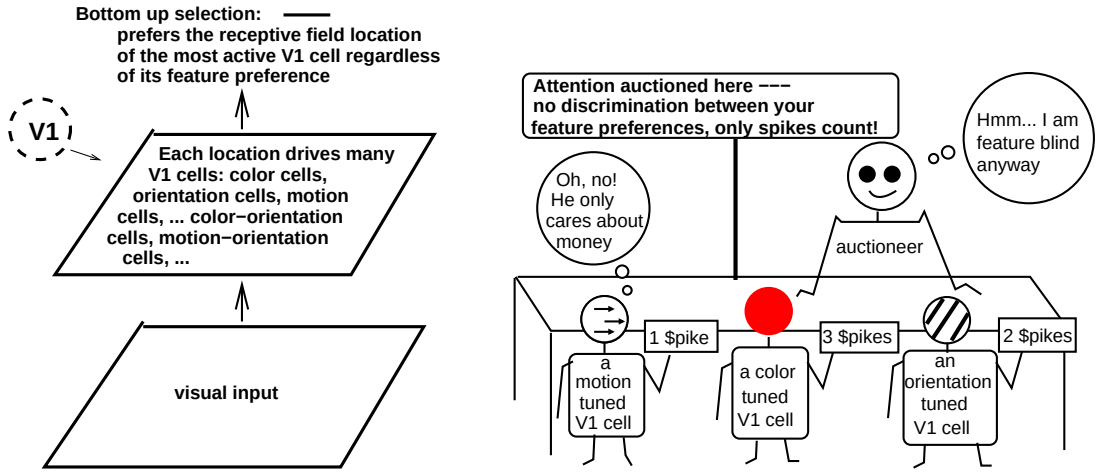



Figure 4.5: Schematic summaries of the previous and the V1 theories of the bottom up visual saliency map. No separate feature maps, nor any summation of them, are needed in the V1 theory, in contrast to previous theories. The V1 cells signal saliency despite their feature tuning, whereas the previous theories explicitly or implicitly assumes a saliency map in a brain area (such as lateral intraparietal area (LIP), Gottlieb et al 1998) where cells are untuned to features.

map, red, green, vertical, or horizontal, and thus not in the master map either. The traditional theories have been subsequently made more explicit (Koch and Ullman 1985) and implemented by computer algorithms (Itti and Koch 2000). When applied to the stimuli in Fig. (4.4), it becomes clear that the traditional theories correspond to the summation rule $\sum_{x_i=x} O_i$ for saliency determination when different response $O_i$ to different orientations at the same location $x$ represent responses from different feature maps. Thus, the traditional theory would predict easy segmentation for the composite pattern of Fig. (4.4c), contrary to data.

The V1 saliency theory differs from the traditional theories mainly because it was motivated by understanding V1. It aims for fast computation, thus requires no separate feature maps or any combinations of them, nor any decoding of the input features to obtain saliency. Indeed, many V1 neurons, e.g., an orientation and motion direction tuned neuron, are tuned to more than one feature dimension (Livingstone and Hubel 1984), making it impossible to have separate groups of V1 cells

for separate feature dimensions. Furthermore, V1 neurons signal saliency by their responses *despite* their feature tunings, hence their firing rates are the universal currency for saliency (to bid for selection) regardless of the feature selectivity of the cells, just like the purchasing power of Euro is independent of the nationality or gender of the currency holders (Fig. (4.5B)). In contrast, the traditional theories were motivated by explaining the behavioral data by a natural framework, without specifying the cortical location of the feature maps or the master saliency map, or a drive for algorithmic simplicity. This in particular leads to the feature map summation rule for saliency determination, and implies that the master saliency map should be in a higher level visual area where cells are untuned to features.

# Chapter 5

# Summary

This paper reviews two lines of works to understand early vision by its role of data reduction in the face of information bottlenecks. The efficient coding principle views the properties of input sampling and input transformations by the early visual RFs as serving the goal of encoding visual inputs efficiently, so that as much input information as possible can be transmitted to higher visual areas through information channel bottlenecks. It not only accounts for these neural properties, but also, by linking these properties with visual sensitivity in behavior, provides an understanding of sensitivity or perceptual changes caused by adaptation to different environment (Atick et al 1993), and of effects of developmental deprivation (Li 1995). Non-trivial and easily testable predictions have also been made (Dong and Atick 1995, Li 1994b, 1996), some of which have subsequently been confirmed experimentally, for example on the correlation between the preferred orientation and ocularity of the V1 cells (Zhaoping et al 2006). The V1 saliency map hypothesis views V1 as creating a bottom up saliency map to facilitate information selection or discarding, so that data rate can be further reduced for detailed processing through the visual attentional bottleneck. This hypothesis not only explains the V1 properties not accounted for by the efficient coding principle, but also links V1's physiology to complex visual search and segmentation behavior previously thought of as not associated with V1. It also makes testable predictions, some of which have also subsequently been confirmed as shown here and previously (e.g., Li 2002, Zhaoping and Snowden 2006). Furthermore, its computational considerations and physiological basis raised fundamental questions about the traditional, behaviorally based, framework of visual selection mechanisms.

The goal of theoretical understanding is not only to give insights to the known facts, thus linking seemingly unrelated data, e.g., from physiology and from behavior, but also to make testable predictions and motivate new experiments and research directions. This strategy should be the most fruitful also for answering many more unanswered questions regarding early visual processes, most particularly the mysterious functional role of LGN, which receives retinal outputs, sends outputs to V1, and receives massive feedback fibers from V1 (Casagrande et al 2005). This paper also exposed a lack of full understanding of the overcomplete representation in V1, despite our recognition of its usefulness in the saliency map and its contradiction to efficient coding. The understanding is likely to arise from a better understanding of bottom up saliency computation, and the study of possible roles of V1 (Lennie 2003, Lee 2003, Salinas and Abbott 2000, Olshausen and Field 2005), such as learning and recognition, beyond input selection or even bottom up visual processes. Furthermore, such pursuit can hopefully expose gaps in our current understanding and prepare the way to investigate behavioral and physiological phenomena beyond early vision.

**Acknowledgement** Work supported by the Gatsby Charitable Foundation. I thank Peter Dayan, Richard Turner, and two anonymous reviewers for very helpful comments on the drafts.

# Chapter 6

# References

# Bibliography


[1] Adelson EH and Bergen JR, (1985) Spatiotemporal energy models for the perception of motion *Journal of Optical Society of America. A* 2, 284-299.

[2] Allman J, Miezin F, McGuinness E. Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annu Rev. Neurosci.* 8:407-30 (1985).

[3] Atick JJ and Redlich AN Towards a theory of early visual processing (1990) *Neural Computation* 2:308-320.

[4] Atick JJ, Li Z, Redlich AN (1990) Color coding and its interaction with spatiotemporal processing in the retina. Preprint IASSNS-HEP-90-75 of Institute for Advanced Study, Princeton, USA.

[5] Atick JJ. Could information theory provide an ecological theory of sensory processing. *Network:Computation and Neural Systems* 3: 213-251. (1992)

[6] Atick J.J. Li, Z., and Redlich A. N. Understanding retinal color coding from first principles *Neural Computation* 4(4):559-572 (1992).

[7] Atick J. J., Li, Z., and Redlich A. N. What does post-adaptation color appearance reveal about cortical color representation? (1993) *Vision Res.* 33(1):123-9.

[8] Buchsbaum G, Gottschalk A. (1983) Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proc R Soc Lond B Biol Sci.* 220(1218):89-113.

[9] Barlow HB, 1961, "Possible principles underlying the transformations of sensory messages." In: Sensory Communication W.A. Rosenblith, ed., Cambridge MA, MIT Press, pp. 217-234.

[10] Barlow H.B. (1981) The Ferrier Lecture, 1980: Critical limiting factors in the design of the eye and visual cortex. *Proc. R. Soc. London B* 212, 1-34.

[11] Barlow HB, Fitzhugh R, Kuffler SW (1957) Change of organization in the receptive fields of the cat retina during dark adaptation. *J. Physiol.* 137: 338-354.

[12] Bell AJ Sejnowski TJ (1997) The 'independent components' of natural scenes are edge filters. *Vision Res.* 23: 3327-38.

[13] Bressloff PC, Cowan JD, Golubitsky M, Thomas PJ, Wiener MC. (2002) What geometric visual hallucinations tell us about the visual cortex. *Neural Comput.* 14(3):473-91.

[14] Casagrande VA, Guillery RW, and Sherman SM (2005) Eds. *Cortical function: a view from the thalamus*, volumn 149. of *Progress in Brain Research*, Elsevier 2005.

[15] Dan Y, Atick JJ, Reid RC. (1996) Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory *J Neurosci.* 16(10):3351-62.

[16] Daubechies I. *The lectures on wavelets*, SIAM 1992.

[17] Dong DW, Atick JJ 1995 "Temporal decorrelation: a theory of lagged and non-lagged responses in the lateral geniculate nucleus," Network: Computation in Neural Systems, 6:159-178.

[18] Duncan J., Humphreys G.W. 1989 Visual search and stimulus similarity, *Psychological Rev.* 96(3): 433-58.

[19] Field DJ 1987 Relations between the statistics of natural images and the response properties of cortical cells. Journal of Optical Society of America, A 4(12):2379-94. 1987

[20] Field DJ 1989 What the statistics of natural images tell us about visual coding. *SPIE* vol. 1077 Human vision, visual processing, and digital display, 269-276

[21] Freeman WT and Adelson EH (1991) The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(9):891-906.

[22] Gilbert C.D., Wiesel T.N., Clustered intrinsic connections in cat visual cortex. *J. Neurosci.* 3(5):1116-33 (1983)

[23] Gottlieb JP, Kusunoki M, Goldberg ME. (1998) The representation of visual salience in monkey parietal cortex. *Nature* 391(6666):481-4.

[24] Hamilton DB, Albrecht DG, Geisler WS. (1989) Visual cortical receptive fields in monkey and cat: spatial and temporal phase transfer function. *Vision Research* 29(10):1285-308.

[25] Holub RA, Morton-Gibson M. (1981) Response of Visual Cortical Neurons of the cat to moving sinusoidal gratings: response-contrast functions and spatiotemporal interactions. *Journal of Neurophysiology* 46(6):1244-59

[26] Horton JC and Hocking DR (1996) Anatomical demonstration of ocular dominance columns in striate cortex of the squirrel monkey *J. Neurosci.* 16(17):5510-5522.

[27] Horwitz GD, Albright TD. (2005) Paucity of chromatic linear motion detectors in macaque V1. *J Vis.* 5(6):525-33.

[28] Hubel, D. H. and Wiesel, T. N. 1965. "Binocular interaction in striate cortex of kittens reared with artificial squint." J. Neurophysiol 28: 1041-1059.

[29] Hubel, D. H. and Wiesel, T. N. LeVay, S 1977. "Plasticity of ocular dominance columns in monkey striate cortex." Philosophical Transactions of the Royal Society of London, Series B, 278: 377-409.

[30] Itti L. and Baldi P. (2006) "Bayesian surprise attracts human attention." In *Advances in Neural Information Processing Systems*, Vol. 19 (NIPS2005), pp. 1-8, Cambridge, MA:MIT Press.

[31] Itti L., Koch C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* 40(10-12):1489-506, (2000).

[32] Jones HE, Grieve KL, Wang W, Sillito AM. (2001 Surround suppression in primate V1. *J Neurophysiol.* 86(4):2011-28.

[33] Jonides J. (1981) Voluntary versus automatic control over the mind's eye's movement. In J. B. Long & A. D. Baddeley (Eds.) *Attention and Performance IX* (pp. 187-203). Hillsdale, NJ. Lawrence Erlbaum Associates Inc.

[34] Julesz B. (1981) Textons, the elements of texture perception, and their interactions. *Nature* 290(5802):91-7.

[35] Kadir T. Brady M. (2001) Saliency, scale, and image description. *International J. of Computer Vision* 45(2):83-105.

[36] Kapadia MK, Ito M, Gilbert CD, Westheimer G. Improvement in visual sensitivity by changes in local context: parallel studies in human observers and in V1 of alert monkeys. *Neuron* 15(4):843-56 (1995).

[37] Kaplan E, Marcus S, So YT. Effects of dark adaptation on spatial and temporal properties of receptive fields in cat lateral geniculate nucleus. (1979) *J. Physiol.* 294:561-80.

[38] Kelly D. H. Information capacity of a single retinal channel. *IEEE Trans. Information Theory* 8:221-226, 1962.

[39] Knierim JJ., Van Essen DC, Neuronal responses to static texture patterns in area V1 of the alert macaque monkey. *J. Neurophysiol.* 67(4): 961-80 (1992)

[40] Koch C., Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* 4(4): 219-27 (1985).

[41] Laughlin S B (1981) A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch [C]* 36: 910-2.

[42] Laughlin SB, Howard J, Blakeslee B. (1987) Synaptic limitations to contrast coding in the retina of the blowfly Calliphora. *Proc. R. Soc. Lond. B Biol. Sci.* 231(1265):437-67.

[43] Lennie P. (2003) The cost of cortical computation. *Curr Biol.* 13(6):493-7.

[44] Lewis A, Garcia R, Zhaoping L. (2003) The distribution of visual objects on the retina: connecting eye movements and cone distributions. *Journal of vision* 3(11):893-905.

[45] Lewis AS and Zhaoping L. (2005) Saliency from natural scene statistics. Program No. 821.11. *2005 Abstract Viewer/Itinerary Planner.* Washington, DC: Society for Neuroscience, 2005. Online.

[46] Lewis AS and Zhaoping L. (2006) Are cone sensitivities determined by natural color statistics? *Journal of Vision*. 6(3):285-302. http://www.journalofvision.org/6/3/8/.

[47] Levy WB and Baxter RA 1996 Energy efficient neural codes. Neural Computation, 8(3) 531-43.

[48] Lee TS (1996) Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Analysis and Machine Intelligence* 18, 959-971.

[49] Lee TS (2003) Computations in the early visual cortex. *J. Physiology, Paris* 97(2-3):121-39.

[50] Li Zhaoping (1992) Different retinal ganglion cells have different functional goals. *International Journal of Neural Systems*, 3(3):237-248.

[51] Li Zhaoping and Atick J. J. 1994a, "Towards a theory of striate cortex" Neural Computation **6, 127-146**

[52] Li Zhaoping and Atick J. J. 1994b, "Efficient stereo coding in the multiscale representation" Network: computation in neural systems Vol.5 1-18.

[53] Li, Zhaoping Understanding ocular dominance development from binocular input statistics. in *The neurobiology of computation* (Proceeding of Computational Neuroscience Conference, , July 21-23, 1994, Monterey, California, 1994), p. 397-402. Ed. J. Bower, Kluwer Academic Publishers, 1995.

[54] Li Zhaoping 1996 "A theory of the visual motion coding in the primary visual cortex" Neural Computation vol. 8, no.4, p705-30.

[55] Li Z. (1998a) Primary cortical dynamics for visual grouping. *Theoretical aspects of neural computation* Eds. Wong KM, King I, Yeung D-Y. pages 155-164. Springer-verlag, Singapore, January 1998.

[56] Li Z. (1998b) A neural model of contour integration in the primary visual cortex. *Neural Comput.* 10(4):903-40.

[57] Li Z. (1998c) Visual segmentation without classification: A proposed function for primary visual cortex. *Perception* Vol. 27, supplement, p 45. (Proceedings of ECVP, 1998, Oxford, England).

[58] Li Z. Visual segmentation by contextual influences via intra-cortical interactions in primary visual cortex. *Network: Computation and neural systems* 10(2):187-212, (1999a).

[59] Li Z. Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl Acad. Sci USA,* 96(18):10530-5. (1999b)

[60] Li, Zhaoping and Dayan P. (1999) "Computational differences between asymmetrical and symmetrical networks" *Network: Computation in Neural Systems* Vol. 10, 1, 59-77.

[61] Li Z, Pre-attentive segmentation in the primary visual cortex. *Spatial Vision,* 13(1) 25-50. (2000)

[62] Li Z. (2001) Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. *Neural Computation* 13(8):1749-1780.

[63] Li Zhaoping 2002, "A saliency map in primary visual cortex " Trends in Cognitive Sciences Vol 6. No.1. Jan. 2002, page 9-16

[64] Linsker R. 1990 Perceptual neural organization: some approaches based on network models and information theory. Annu Rev Neurosci. 13:257-81.

[65] Livingstone MS, Hubel DH.Anatomy and physiology of a color system in the primate visual cortex. *J. Neurosci.* 4(1):309-56 (1984).

[66] Meister M, Berry MJ (1999) The neural code of the retina *NEURON* 22(3):435-450.

[67] Nadal J.P. and Parga N. 1993. Information processing by a perceptron in an unsupervised learning task. *Network:Computation and Neural Systems* 4(3), 295-312.

[68] Nadal J.P. and Parga N. 1994. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network:Computation and Neural Systems* 5:565-581.

[69] Nakayama K, Mackeben M. (1989) Sustained and transient components of focal visual attention. *Vision Res*. 29(11):1631-47.

[70] Nirenberg, S. Carcieri, S. M. Jacobs A. L. and Latham P. E. (2001) Retinal ganglion cells act largely as independent encoders *Nature* 411:698-701.

[71] Nothdurft HC, Gallant JL, Van Essen DC. Response modulation by texture surround in primate area V1: correlates of "popout" under anesthesia. *Vis. Neurosci.* 16, 15-34 (1999).

[72] Olshausen BA and Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research* 37:3311-3325.

[73] Olshausen BA and Field DJ (2005) How Close Are We to Understanding V1? *Neural Computation* 17:1665-1699

[74] Pashler H. (1998) *Attention* Editor. East Sussex, UK. Psychology Press Ltd.

[75] Petrov Y. Zhaopimg L. Local correlations, information redundancy, and sufficient pixel depth in natural images. *J. Opt Soc. Am. A Opt Image Sci. Vis.* 20(1):56-66. (2003).

[76] Puchalla JL, Schneidman E, Harris RA, Berry MJ. 2005 Redundancy in the population code of the retina. *Neuron* 46(3):493-504.

[77] Rao RPN and Ballard DH Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience* 2: 79-87 (1999).

[78] Rockland KS., Lund JS., Intrinsic laminar lattice connections in primate visual cortex. *J. Comp. Neurol.* 216(3):303-18 (1983).

[79] Rubenstein B.S., Sagi D., Spatial variability as a limiting factor in texture-discrimination tasks: implications for performance asymmetries. *J Opt Soc Am A.*;7(9):1632-43, (1990)

[80] Ruderman DL, Cronin TW, Chiao C-C. Statistics of cone responses to natural images: implications for visual coding. *Journal of Optical Society of America. A* 15(8):2036-45.

[81] Salinas E, Abbott LF. (2000) Do simple cells in primary visual cortex form a tight frame? *Neural Comput.* 12(2):313-35.

[82] Schreiber W. 1956 The measurement of third order probability distributions of television signals *IEEE Trans. Information Theory* 2(3):94-105.

[83] Schwartz O, Simoncelli EP. (2001) Natural signal statistics and sensory gain control. *Nat Neurosci* 4(8): 819-25.

[84] Shannon CE and Weaver W (1949) The mathematical theory of communication (Urbana IL, University of Illinois Press)

[85] Sillito AM, Grieve KL, Jones HE, Cudeiro J, Davis J. Visual cortical mechanisms detecting focal orientation discontinuities. *Nature* 378, 492-496 (1995).

[86] Simons D.J. & Chabris C.F. (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28: 1059-1074

[87] Simoncelli E and Olshausen B. 2001 "Natural image statistics and neural representation" Annual Review of Neuroscience, 24, 1193-216.

[88] Srinivasan MV, Laughlin SB, Dubs A. (1982) Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond B Biol Sci.* 216(1205):427-59.

[89] Stryker M. P. 1986 "The role of neural activity in rearranging connections in the central visual system." In Ruben, R.J. Vater, T.R. Van De, and Rubel, E.W. (Eds.), The biology of Change in Otolaryngology, pp. 211-224, Elsevier Science B.V. Amsterdam.

[90] Sziklai G (1956) Some studies in the speed of visual perception *IEEE Transactions on Information Theory* 2(3):125-8

[91] Tehovnik EJ, Slocum WM, Schiller PH. Saccadic eye movements evoked by microstimulation of striate cortex. *Eur J. Neurosci.* 17(4):870-8 (2003).

[92] Treisman A. M., Gelade G. A feature-integration theory of attention. *Cognit Psychol.* 12(1), 97-136, (1980).

[93] Treisman A, Gormican S. (1988) Feature analysis in early vision: evidence from search asymmetries. *Psychol Rev.* 95(1):15-48.

[94] van Hateren J. (1992) A theory of maximizing sensory information. *Biol Cybern* 68(1):23-9.

[95] van Hateren J. Ruderman D.L. (1998) Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex *Proc. Biol. Sciences.* 265(1412):2315-20

[96] Wachtler T, Sejnowski TJ, Albright TD. (2003) Representation of color stimuli in awake macaque primary visual cortex. *Neuron* 37(4):681-91.

[97] Wolfe J.M., Cave K.R., Franzel S. L. Guided search: an alternative to the feature integration model for visual search. *J. Experimental Psychol.* 15, 419-433, (1989).

[98] Wolfe J. M. (1998) "Visual Search, a review" In H. Pashler (ed.) *Attention* p, 13-74. Hove, East Sussex, UK. Psychology Press Ltd.

[99] Zhaoping L. (2005) The primary visual cortex creates a bottom-up saliency map. In *Neurobiology of Attention* Eds L. Itti, G. Rees and J.K. Tsotsos, Elsevier, 2005, Chapter 93, page 570-575

[100] Zhaoping L. Hubner M., and Anzai A. (2006) Efficient stereo coding in the primary visual cortex and its experimental tests based on optical imaging and single cell data. Presented at Annual Meeting of Computational Neuroscience, Edinburgh, summer, 2006.

[101] Zhaoping L. and May K.A. (2004) Irrelevance of feature maps for bottom up visual saliency in segmentation and search tasks. Program No. 20.1 *2004 Abstract Viewer/Itinerary Planner*, Washington D. C., Society for Neurocience, 2004. Online.

[102] Zhaoping L. and Snowden RJ (2006) A theory of a saliency map in primary visual cortex (V1) tested by psychophysics of color-orientation interference in texture segmentation. in *Visual Cognition* 14(4-8):911-933.

[103] Zhaoping L. (2006) Overcomplete representation for fast attentional selection by bottom up saliency in the primary visual cortex. Presented at *European Conference on Visual Perception*, August, 2006.

[104] Zigmond, M.J., Bloom F. E., Lndis S. C., Roberts J. L., Squire L. R. *Fundamental neuroscience* Academic Press, New York, 1999.