Communicated by Geoffrey Hinton

Towards a Theory of Early Visual Processing

Joseph J. Atick
School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540, USA

A. Norman Redlich
Department of Physics and Center for Neural Science, New York University, New York, NY 10003, USA

We propose a theory of the early processing in the mammalian visual pathway. The theory is formulated in the language of information theory and hypothesizes that the goal of this processing is to recode in order to reduce a "generalized redundancy" subject to a constraint that specifies the amount of average information preserved. In the limit of no noise, this theory becomes equivalent to Barlow's redundancy reduction hypothesis, but it leads to very different computational strategies when noise is present. A tractable approach for finding the optimal encoding is to solve the problem in successive stages where at each stage the optimization is performed within a restricted class of transfer functions. We explicitly find the solution for the class of encodings to which the parvocellular retinal processing belongs, namely linear and nondivergent transformations. The solution shows agreement with the experimentally observed transfer functions at all levels of signal to noise.

In the mammalian visual pathway, data from the photoreceptors are processed sequentially through successive layers of neurons in the retina and in the visual cortex. The early stages of this processing (the retina and the first few layers of the visual cortex) exhibit a significant degree of universality; they are very similar in many species and do not change as a mature animal learns new visual perceptual skills. This suggests that the early stages of the visual pathway are solving a very general problem in data processing, which is independent of the details of each species' perceptual needs. In the first part of this paper, we formulate a theory of early visual processing that identifies this general problem.

The theory is formulated in the language of information theory (Shannon and Weaver 1949) and was inspired by Barlow's redundancy reduction hypothesis for perception (Barlow 1961, 1989). Barlow's hypothesis is, however, applicable only to noiseless channels that are unrealistic. The

theory that we develop here is formulated for noisy channels. It agrees with Barlow's hypothesis in the limit of no noise but it leads to different computational strategies when noise is present. Our theory hypothesizes that the goal of visual processing is to recode the sensory data in order to reduce a redundancy measure, defined below, subject to a constraint that fixes the amount of average information maintained. The present work is an outgrowth of an earlier publication in which we addressed some of these issues (Atick and Redlich 1989). However, in that work the role of noise was not rigorously formulated, and although all solutions exhibited there did well in reducing redundancy, they were not proven to be optimal. For a related attempt to understand neural processing from information theory see Linsker (1986,1989) (see also Uttley 1979).

The problem of finding the optimal redundancy reducing code among all possible codes is most likely impossible to solve. A more tractable strategy is to reduce redundancy in successive stages, where at each stage one finds the optimal code within a restricted class. This appears to be the mechanism used in the visual pathway. For example, in the "parvocellular" portion of the pathway, which is believed to be concerned with detailed form recognition, the first recoding (the output of the retinal ganglion cells) can be characterized as linear and nondivergent (code dimension is unchanged). At the next stage, the recoding of the simple cells is still substantially linear but is divergent (for a review see Orban 1984). In this paper, we solve the problem of redundancy reduction for the class of linear and nondivergent codes and we find that the optimal solution is remarkably similar to the experimentally observed ganglion cell recoding. We leave the solution for the next stage of linear divergent codes, where one expects a simple cell like solution, for a future publication.

1 Formulation of the Theory

For concreteness, we shall start by formulating our theory within the specific context of retinal processing. The theory in its more general context will become clear later when we state our redundancy reduction hypothesis. It is helpful to think of the retinal processing in terms of a pair of communication channels, as pictured in Figure 1. In this flow chart, the center box represents the retinal transfer function A, with the signal x representing the visual input including noise ν, and y the output of the ganglion cells. Here, we do not concern ourselves with the detailed implementation of this transfer function by the retina, which involves a fairly complicated interaction between the photoreceptors and the layers of cells leading to the ganglion cells.

Although the input x is the actual input to the visual system, we have introduced an earlier input communication channel in the flow diagram with s representing an *ideal* signal. This earlier communication channel

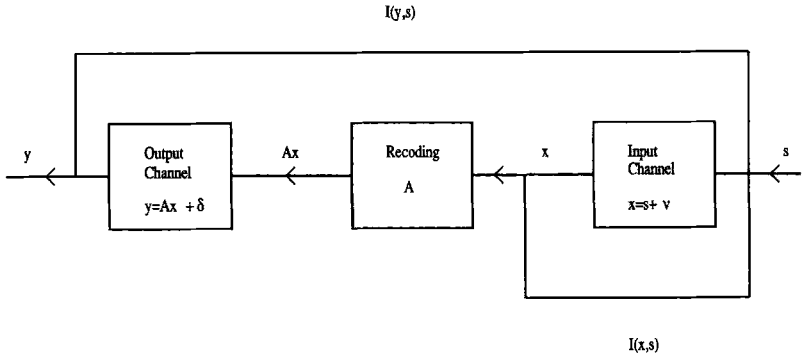

Figure 1: Schematic diagram showing the three stages of processing that the ideal signal s undergoes before it is converted to the output y. ν (δ) is the noise in the input (output) channel.

is taken to be the source of all forms of noise ν in the signal x, including quantum fluctuations, intrinsic noise introduced by the biological hardware, and semantic noise already present in the images. It must be kept in mind, however, that neither *noise* nor *ideal* signal is a universal concept, but depend on what is useful visual information to a particular organism in a particular environment. Here we assume that, minimally, the ideal signal does not include noise ν in the form of completely random fluctuations. It is reasonable to expect this minimal definition of noise to apply to the early visual processing of all organisms.

In this paper, we take the input x to be the discrete photoreceptor sampling of the two-dimensional luminosity distribution on the retina. For simplicity, we use a cartesian coordinate system in which the sampled signal at point $\mathbf{n} = (n_1, n_2)$ is $x[\mathbf{n}]$. However, all of the results below can be rederived, including the sampling, starting with spatially continuous signals, by taking into account the optical modulation transfer function of the eye (see the analysis in Atick and Redlich 1990).

The output y is the recoded signal Ax plus the noise δ in the output channel. This channel may be thought of as the optic nerve. Since the recoding of x into y is linear and nondivergent, its most general form can be written as $y[\mathbf{m}] = \sum_{\mathbf{n}} A[\mathbf{m}, \mathbf{n}]\, x[\mathbf{n}] + \delta[\mathbf{m}]$, where the transfer function $A[\mathbf{m}, \mathbf{n}]$ is a square matrix.

In order to formulate our hypothesis, we need to define some quantities from information theory (Shannon and Weaver 1949) that measure

how well the visual system is communicating information about the visual scenes. We first define the *mutual information* $I(x, s)$ between the ideal signal s and the actual signal x:

$$I(x,s) = \sum_{s,x} P(x,s) \log \left[\frac{P(x,s)}{P(s)\,P(x)} \right] \tag{1.1}$$

with a similar formula for $I(y, s)$. In equation 1.1, $P(s)$ [or $P(x)$] is the probability of the occurrence of a particular visual signal $s[\mathbf{n}]$ (or $x[\mathbf{n}]$), and $P(x, s)$ is the probability of the joint occurrence of $s[\mathbf{n}]$ together with $x[\mathbf{n}]$. $I(x, s)$ measures the actual amount of useful information available at the level of the photoreceptors x, given that the desired signal is s. Likewise, $I(y, s)$ measures the useful information available in the signal y. Also, for continuous signals, the mutual information is a well-defined, coordinate invariant quantity.

To calculate the mutual information $I(x, s)$ explicitly, it is necessary to know something about the probabilities $P(s)$, $P(x)$, and $P(x, s)$. These probability functions, together with the relationship $y = A\,x + \delta$, are also sufficient to calculate the mutual information $I(y, s)$. Although $P(s)$, $P(x)$, and $P(x, s)$ cannot be known completely, we do assume knowledge of the second-order correlators $\langle s[\mathbf{n}]s[\mathbf{m}]\rangle$, $\langle x[\mathbf{n}]x[\mathbf{m}]\rangle$, and $\langle x[\mathbf{n}]s[\mathbf{m}]\rangle$, where $\langle\ \rangle$ denotes the average over the ensemble of all visual scenes. We assume that these correlators are of the form

$$\begin{aligned} \langle s[\mathbf{n}]s[\mathbf{m}]\rangle &= R_0[\mathbf{n},\mathbf{m}] \\ \langle x[\mathbf{n}]x[\mathbf{m}]\rangle &= R_0[\mathbf{n},\mathbf{m}] + N^2\delta_{\mathbf{n},\mathbf{m}} \equiv R[\mathbf{n},\mathbf{m}] \\ \langle x[\mathbf{n}]s[\mathbf{m}]\rangle &= \langle s[\mathbf{n}]s[\mathbf{m}]\rangle \end{aligned} \tag{1.2}$$

where $R_0[\mathbf{n}, \mathbf{m}]$ is some yet unspecified correlation matrix, and we have defined $\langle \nu[\mathbf{n}]\nu[\mathbf{m}]\rangle \equiv N^2\delta_{\mathbf{n},\mathbf{m}}$. Using $x = s + \nu$, equations 1.2 imply that there are no correlations between the noise ν and s. Given these correlators, we assume that the probability distributions are those with maximal entropy:

$$\begin{aligned} P(u) &= \left[(2\pi)^d \det(R_{uu})\right]^{-1/2} \\ &\quad \exp\left[-\frac{1}{2}\sum_{\mathbf{n},\mathbf{m}} (u[\mathbf{n}] - \bar{u}) R^{-1}_{uu}[\mathbf{n},\mathbf{m}](u[\mathbf{m}] - \bar{u})\right] \end{aligned} \tag{1.3}$$

for $u = s, x, y$ and $R_{uu}[\mathbf{n}, \mathbf{m}] \equiv \langle u[\mathbf{n}]\,u[\mathbf{m}]\rangle$ (d is the dimension of $\mathbf{n}$). We have included here the mean $\bar{u} \equiv \langle u\rangle$, although in all of our results it drops out. Equation 1.3 can also be used to determine the joint proba-

bilities $P(x,s)$ and $P(y,s)$, since these are equal to $P(z_{xs})$ and $P(z_{ys})$ for the larger sets of stochastic variables $z_{xs} = (x,s)$ and $z_{ys} = (y,s)$ whose correlators R_{zz} are calculated from R_{xx}, R_{xs}, R_{ss}, R_{yy}, and R_{ys}.

It is not difficult to show, using the explicit expressions for the various probability distributions, that

$$I(x,s) = \frac{1}{2}\log\left[\frac{\det\left(R_0 + N^2\right)}{\det N^2}\right] \tag{1.4}$$

$$I(y,s) = \frac{1}{2}\log\left\{\frac{\det\left[A(R_0 + N^2)A^T + N_\delta^2\right]}{\det\left(AN^2A^T + N_\delta^2\right)}\right\} \tag{1.5}$$

(In 1.5, we used $\langle \delta[\mathbf{n}]\, \delta[\mathbf{m}]\rangle \equiv N_\delta^2 \delta_{\mathbf{n},\mathbf{m}}$.) The mutual informations depend on both the amount of *noise* and on the amount of *correlations* in the signals. Noise has the effect of reducing $I(x,s)$ [or $I(y,s)$] because it causes uncertainty in what is known about s at x (or y). In fact, infinite noise reduces $I(x,s)$ and $I(y,s)$ to zero. This becomes clear in equations 1.4 and 1.5 as N^2 goes to infinity, since then the ratio of determinants goes to one causing I to vanish.

Increasing spatial correlations in equations 1.4 and 1.5 also has the effect of reducing I because correlations reduce the information in the signals. Correlations indicate that some scenes are far more common than others, and an ensemble with this property has lower average information than one in which all messages are equally probable. The effect of increasing correlations is most easily seen, for example, in $I(x,s)$ in the limit of N^2 very small, in which case $I(x,s) \sim \log[\det(R_0)]$. If the average signal strengths $\langle s^2[\mathbf{n}]\rangle$ are held constant, then $\log[\det(R_0)]$ is maximum when R_0 is diagonal (no correlations) and vanishes when the signal is completely correlated, that is, $R_0[\mathbf{n},\mathbf{m}] = \text{constant}$, $\forall$ $\mathbf{n}$, $\mathbf{m}$. In fact, by Wegner's theorem (Bodewig 1956, p. 71), for positive definite matrices (correlation matrices are always positive definite) $\det(R_0) \leq \Pi_i (R_0)_{ii}$, with equality only when R_0 is completely diagonal.

Having introduced a measure $I(y,s)$ of the actual average information available at y, we now define the channel *capacity* $C_{\text{out}}(y)$ which measures the *maximal* amount of information that could flow through the output channel. Here, we define the capacity $C_{\text{out}}(y)$ as the maximum of $I(y,w)$ varying freely over the probabilities $P(w)$ of the inputs to the *output* channel, holding the average signal strengths $\langle y^2[\mathbf{n}]\rangle$ fixed:

$$\begin{aligned} C_{\text{out}}(y) &= \max_{P(w)} \; I(y,w)\Big|_{\langle y^2\rangle = \text{const.}} \\ &= \max_{P(w)} \; \log\det\left(\frac{R_{ww} + N_\delta^2}{N_\delta^2}\right)\Big|_{(R_{ww})_{ii} = \text{const.}} \end{aligned} \tag{1.6}$$

where $y = w + \delta$ and $(R_{ww})_{ii}$ are the diagonal elements of the autocorrelator of w (w is a dummy variable, which in Fig. 1 corresponds to Ax). A constraint of this sort is necessary to obtain a finite capacity for continuous signals and is equivalent to holding constant the average "power" expenditure or the variance in the number of electrochemical spikes sent along each fiber of the optic nerve.[1] Using Wegner's theorem, (1.6) the maximum occurs for the probability distribution $P(w)$ for which R_{ww}, or equivalently R_{yy}, is diagonal:

$$C_{\text{out}}(y) = \frac{1}{2} \log \prod_i \left[\frac{R_{yy}}{N_\delta^2} \right]_{ii} \tag{1.7}$$

which for $y = Ax + \delta$ is explicitly

$$C_{\text{out}}(y) = \frac{1}{2} \log \prod_i \left[\frac{A(R_0 + N^2)A^T + N_\delta^2}{N_\delta^2} \right]_{ii} \tag{1.8}$$

At this point, we are ready to state our generalized redundancy reduction hypothesis. We propose that the purpose of the recoding A of the visual signal in the early visual system is to *minimize the "redundancy"*

$$\mathcal{R} = 1 - I(y, s)/C_{\text{out}}(y) \tag{1.9}$$

subject to the constraint that $I(y, s)$ be equal to the minimum average information I^* that must be retained to meet an organism's needs. $I(y, s)$ is therefore constrained to be a fixed quantity and redundancy is reduced by choosing an A that minimizes $C_{\text{out}}(y)$. To avoid confusion, we should emphasize that C_{out} is fixed only at fixed "power," but can be lowered by choosing A to lower the output "power."

Although, in practice we do not know precisely what the minimal I^* is, we assume here that it is the information available to the retina, $I(x, s)$, lowered slightly by the presence of the additional noise δ in the output channel. We therefore choose the constraint

$$I(y, s) = I^* = I(x + \delta, s) \tag{1.10}$$

but our results below do not depend qualitatively on this precise form for the constraint. Since I^* does not depend on A, it can be determined from physiological data, and then used to predict independent experiments (see Atick and Redlich 1990).

The reader should be cautioned that equation 1.9 is *not* the conventional definition of redundancy for the total channel from s to y. The standard redundancy would be $\mathcal{R} = 1 - I(y, s)/C_{\text{tot}}(y)$ where C_{tot} is the maximum of $I(y, s)$ varying freely over the input probabilities $P(s)$,

[1] Since a ganglion cell has a nonvanishing mean output, "power" here is actually the cell's dynamic range.

keeping $\langle y^2 \rangle$ fixed. In contrast to C_{tot}, C_{out} is directly related to the "power" in the optic fiber, so reducing equation 1.9 in the manner just described always leads to lower "power" expenditure. Also, since $C_{\text{out}} > C_{\text{tot}}$, lowering C_{out} necessarily lowers the "power" expenditure at all stages up to y, which is why we feel equation 1.9 could be biologically more significant.

Our hypothesis is similar to Barlow's redundancy reduction hypothesis (Barlow 1961), with the two becoming identical when the system is free of noise ν. In this limit, redundancy is reduced by diagonalizing the correlation matrix R_0 by choosing the transfer matrix A such that $R_{yy} = AR_0A^T$ is diagonal. With R_{yy} diagonal, the relationship $\det(R_{yy}) \leq \Pi_i(R_{yy})_{ii}$ becomes an equality giving $C(y) = I(y, s)$ so the redundancy (1.9) is eliminated. [In reality, the redundancy (1.9) is a lower bound reflecting the fact that we chose probability distributions which take into account only second-order correlators. More complete knowledge of $P(s)$ would lower $I(x, s)$ and $I(y, s)$ and therefore increase $\mathcal{R}$.]

Where reducing $\mathcal{R}$ in equation 1.9 differs considerably from Barlow's hypothesis is in the manner of redundancy reduction when noise is significant. Under those circumstances, $\mathcal{R}$ in equation 1.9 is sizable, not because of correlations in the signal, but because much of the channel capacity is wasted carrying noise. Reducing equation 1.9 when the noise is large has the effect of increasing the signal-to-noise ratio. To do this the system actually *increases* correlations (more precisely increases the amplitude of the correlated signal relative to the noise amplitude), since correlations are what distinguish signal from noise. For large enough noise, more is gained by lowering the noise in this way than is lost by increasing correlations. For an intermediate regime, where signal and noise are comparable, our principle leads to a compromise solution, which locally accentuates correlations, but on a larger scale reduces them. All these facts can be seen by examining the properties of the explicit solution given below.

Before we proceed, it should also be noted that Linsker (1986) has hypothesized that the purpose of the encoding A should be to maximize the mutual information $I(y, s)$, subject to some constraints. This differs from the principle in this paper which focuses on lowering the output channel capacity while maintaining the *minimum* information needed by the organism. While both principles may be useful to gain insight into the purposes of neural processing in various portions of the brain, in the early visual processing, we believe that the primary evolutionary pressure has been to reduce output channel capacity. For example, due to much lower resolution in peripheral vision, the amount of information arriving at the retina is far greater than the information kept. It is difficult to believe that this design is a consequence of inherent local biological hardware constraints, since higher resolution hardware is clearly feasible, as seen in the fovea.

2 Explicit Solution

To actually minimize $\mathcal{R}$ we use a lagrange multiplier λ to implement the constraint (equation 1.10) and minimize

$$E\{A\} = C(y) - \lambda[I(y,s) - I(x+\delta,s)] \tag{2.1}$$

with respect to the transfer function A, where $C(y)$, $I(x,s)$, and $I(y,s)$ are given in equations 1.8, 1.4, and 1.5, respectively. One important property of $R[\mathbf{n},\mathbf{m}]$ (in equation 1.2) that we shall assume is translation invariance, $R[\mathbf{n},\mathbf{m}] = R[\mathbf{n}-\mathbf{m}]$, which is a consequence of the homogeneity of the ensemble of all visual scenes. We can take advantage of this symmetry to simplify our formulas by assuming $A[\mathbf{n},\mathbf{m}] = A[\mathbf{n}-\mathbf{m}]$. With this assumption, the diagonal elements $(R_{yy})_{ii}$ in equation 1.7 are all equal and hence minimizing $C(y)$ is equivalent to minimizing the simpler expression $\mathrm{Tr}(A\,R\,A^T)$.

Using the identity $\log(\det B) = \mathrm{Tr}(\log B)$ for any positive definite matrix B, and replacing $C(y)$ by $\mathrm{Tr}(A\,R\,A^T)$, equation 2.1 becomes

$$\begin{aligned} E\{A\} \;=\; & \frac{1}{N_\delta^2}\int_{-\pi}^{\pi} d\mathbf{w}\, A(\mathbf{w})\, R(\mathbf{w})\, A(-\mathbf{w}) \\ & -\frac{\lambda}{2}\left\{\int_{-\pi}^{\pi} d\mathbf{w}\log\left[\frac{A(\mathbf{w})R(\mathbf{w})A(-\mathbf{w})+N_\delta^2}{A(\mathbf{w})N^2A(-\mathbf{w})+N_\delta^2}\right]\right. \\ & \left. -\int_{-\pi}^{\pi} d\mathbf{w}\log\left[\frac{R(\mathbf{w})+N_\delta^2}{N^2+N_\delta^2}\right]\right\} \end{aligned} \tag{2.2}$$

where all variables are defined in momentum space through the standard discrete two-dimensional fourier transform, for example,

$$A(\mathbf{w}) \equiv A(w_1,w_2) = \sum_{\mathbf{m}} e^{-i\mathbf{m}\cdot\mathbf{w}} A[\mathbf{m}]$$

It is straightforward to see from equation 2.2 that the optimal transfer function $A(\mathbf{w})$ satisfies the following quadratic equation:

$$[F(\mathbf{w})R(\mathbf{w})+1][F(\mathbf{w})N^2+1] = \frac{\lambda}{2}\frac{R_0(\mathbf{w})}{R(\mathbf{w})} \tag{2.3}$$

where we have defined $F(\mathbf{w}) = A(\mathbf{w})\cdot A(-\mathbf{w})/N_\delta^2$. The fact that A appears only through F, is a manifestation of the original invariances of I and C under orthogonal transformations on the transfer function $A[\mathbf{m}]$, that is, under $A \to U\,A$ with $U^T U = 1$. Equation 2.3 has only one positive solution for F, which is given explicitly by

$$N^2F = \frac{1}{2}\frac{R_0}{R}\left(1+\sqrt{1+\frac{2\lambda N^2}{R_0}}\;\right) - 1 \tag{2.4}$$

where λ is determined by solving $I(y,s) = I(x+\delta, s)$. After eliminating F the latter equation becomes

$$\int d\mathbf{w} \log\left(\sqrt{\frac{R_0}{2\lambda N^2}} + \sqrt{1+\frac{R_0}{2\lambda N^2}}\right) = \frac{1}{2}\int d\mathbf{w} \log\left(\frac{R}{N^2}\cdot\frac{N^2+N_\delta^2}{R+N_\delta^2}\right) \quad (2.5)$$

In general, equation 2.5 must be solved for λ numerically.

The fact that the transfer function A appears only through F leads to a multitude of degenerate solutions for A, related to each other by orthogonal transformations. What chooses among them has to be some principle of minimum effort in implementing such a transfer function. For example, some of the solutions are nonlocal (by local we mean a neighborhood of a point $\mathbf{n}$ on the input grid is mapped to the neighborhood of the corresponding point $\mathbf{n}$ on the output grid), so they require more elaborate hardware to implement; hence we examine local solutions. Among these is a unique solution satisfying $A(\mathbf{w}) = A(-\mathbf{w})$, which implies that it is rotationally invariant in coordinate space. We compare it to the observed retinal transfer function (ganglion kernel), known to be rotationally symmetric.

Since rotation symmetry is known to be broken at the simple cell level, it is significant that this formalism is also capable of producing solutions that are not rotationally invariant even when the correlation function is. It may be that the new features of the class of transfer functions at that level (for example, divergence factor) will lift the degeneracy in favor of the nonsymmetric solutions. (In fact, in one dimension we find solutions that break parity and look like one-dimensional simple cells kernels.)

The rotationally invariant solution is obtained by taking the square root of F in equation 2.4 (we take the positive square root, corresponding to on-center cells). In what follows, we examine some of its most important properties. To be specific, we parameterize the correlation function by a decaying exponential

$$R[\mathbf{n},\mathbf{m}] = N^2\delta_{\mathbf{n}-\mathbf{m},0} + S^2 e^{-\|\mathbf{n}-\mathbf{m}\|/D} \quad (2.6)$$

with D the correlation length measured in acuity units and S the signal amplitude. We have done numerical integration of equations 2.4 and 2.5 and determined $A[\mathbf{m}]$ for several values of the parameters. In Figure 2, we display one typical solution, which was obtained with $S/N = 2.0$, $D = 50$, and $N_\delta = 0.025$. In that figure, empty disks represent positive (excitatory), while solid disks represent negative (inhibitory) components of $A[\mathbf{m}]$. Also, the logarithm of the area of a disk is directly related to the amplitude of the component of $A[\mathbf{m}]$ at that location. As one can see, the solution has a strong and rather broad excitatory center with a weaker and more diffuse surround. A very significant feature of the theoretical profiles is their insensitivity to D (and to N_δ), which is necessary to account for the fact that the observed profiles measured in acuity units are similar in different species and at different eccentricities.

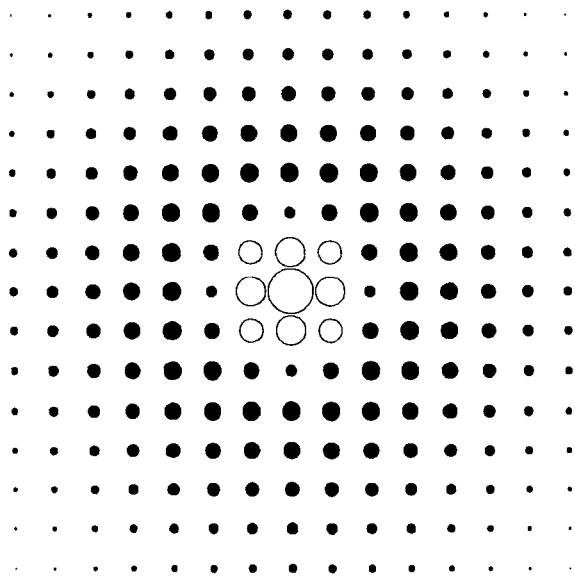

Figure 2: Optimal transfer function, $A[\mathbf{m}]$, for nondivergent linear codes, with $D = 50$, $S/N = 2$, and $N_\delta = 0.025$. Open disks denote positive (excitatory) components of $A[\mathbf{m}]$ while solid disks denote negative (inhibitory) components. The area of a disk is directly related to the logarithm of $A[\mathbf{m}]$ at that location.

To get more insight into this solution, let us qualitatively examine its behavior as we change S/N (for a detailed quantitative comparison with physiological data see Atick and Redlich 1990). For that, we find it more convenient to integrate out one of the dimensions (note this is not the same as solving the problem in one dimension). The resulting profile, corresponding to Figure 2, is shown in Figure 3b. In Figure 3, we have also plotted the result for two other values of S/N, namely for low and high noise regimes (Fig. 3a and c, respectively). These show that an interpolation is happening as S/N changes between the two extremes. Analytically, we can also see this from equation 2.4 for any R_0 by taking the limit $N/S \to 0$, where $A(\mathbf{w})$ becomes equal to

$$\sqrt{(\lambda/2 - 1)/R_0}$$

One recognizes that this is the square root of the solution one gets by carrying out *prediction* on the inputs, a signal processing technique which we advocated for this regime of noise (see also Srinivasan et al. 1982)

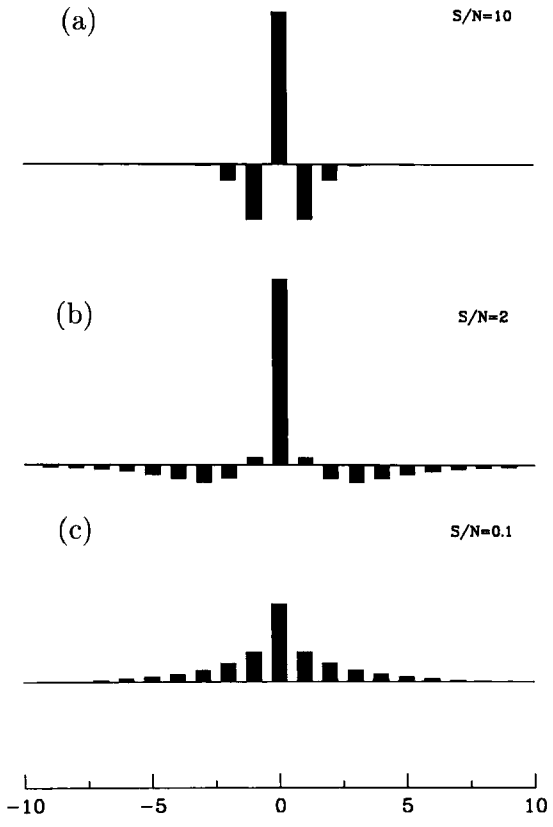

Figure 3: (a–c) Optimal solution at three different values of S/N. These profiles have been produced from the two-dimensional solution by summing over one direction and normalizing the resulting profile such that the height central point is equal to the center height in the two-dimensional solution.

as a redundancy reduction technique in our previous paper (Atick and Redlich 1989). The spatial profiles for the square root solution are very similar to the prediction profiles, albeit a bit more spread out in the surround region. This type of profile reduces redundancy by reducing the amount of correlations present in the signal.

In the other regime, where noise is very large compared to the signal, the solution for $A(\mathbf{w}) \sim (R_0/N^2)^{1/4}$ and has the same qualitative features as the *smoothing* solution (Atick and Redlich 1989) which in that limit is $A_{\text{smoothing}} = R_0/N^2$. Smoothing increases the signal to noise of the output and, in our earlier work, we argued that it is a good redundancy reducing technique in that noise regime. Moreover, in that work, we

argued that to maintain redundancy reduction at all signal-to-noise levels a process that interpolates between prediction and smoothing has to take place. We proposed a convolution of the prediction and the smoothing profiles as a possible interpolation (SPI-coding), which was shown to be better than either prediction or smoothing. In the present analysis, the optimal redundancy reducing transfer function is derived, and, although it is not identical to SPI-coding, it does have many of the same qualitative properties, such as the interpolation just mentioned and the overall center-surround organization.

The profiles in Figures 2 and 3 are very similar to the kernels of ganglions measured in experiments on cats and monkeys. We have been able to fit these to the phenomenological difference of gaussian kernel for ganglions (Enroth-Cugell and Robson 1966). The fits are very good with parameters that fall within the range that has been recorded. Another significant way in which the theory agrees with experiment is in the behavior of the kernels as S/N is decreased. In the theoretical profiles, one finds that the size of the center increases, the surround spreads out until it disappears, and finally the overall scale of the profile diminishes as the noise becomes very large. In experiment, these changes have been noted as the luminosity of the incoming light (and hence the signal to noise) is decreased and the retina adapts to the lower intensity (see, for example, Enroth-Cugell and Robson 1966). This active process, in the language of the current theory, is an adjustment of the optimal redundancy reducing processing to the S/N level.

In closing, we should mention that many of the techniques used to derive optimal encoding for the spatial properties of visual signals can be directly applied to temporal properties. In that case, for low noise the theory would lead to a reduction of temporal correlations, which would have the effect of taking the time derivative, while in the high noise case, the theory would lead to integration. Both types of processing play a significant role in visual perception, and it will be interesting to see how well they can be accounted for by the theory. Another issue that should be addressed is the question of how biological organisms evolved over time to have optimal redundancy reducing neural systems. In our previous paper, we discovered an anti-Hebbian unsupervised learning routine which converges to the *prediction* configuration and a Hebbian routine which converges to the *smoothing* profiles. We expect that there exist reasonably local learning algorithms that converge to the optimal solutions described here.

Acknowledgments

Work supported by the National Science Foundation, Grant *PHYS86-20266*.

References

Atick, J. J., and Redlich, A. N. 1989. Predicting the ganglion and simple cell receptive field organizations from information theory. Preprint no. IASSNS-HEP-89/55 and NYU-NN-89/1.

Atick, J. J., and Redlich, A. N. 1990. Quantitative tests of a theory of early visual processing: I. Spatial contrast sensitivity profiles. Preprint no. IASSNS-HEP-90/51.

Barlow, H. B. 1961. Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, W. A. Rosenblith, ed. M.I.T. Press, Cambridge, MA.

Barlow, H. B. 1989. Unsupervised learning. *Neural Comp.* **1**, 295–311.

Bodewig, E. 1956. *Matrix Calculus*. North-Holland, Amsterdam.

Enroth-Cugell, C., and Robson, J. G. 1966. The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol.* **187**, 517–552.

Linsker, R. 1986. Self-organization in a perceptual network. *Computer* (March), 105–117.

Linsker, R. 1989. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems*, D. S. Touretzky, ed., Vol. 1, pp. 186–194. Morgan Kaufmann, San Mateo.

Orban, G. A. 1984. *Neuronal Operations in the Visual Cortex*. Springer-Verlag, Berlin.

Shannon, C. E., and Weaver, W. 1949. *The Mathematical Theory of Communication*. The University of Illinois Press, Urbana.

Srinivisan, M. V., Laughlin, S. B., and Dubs, A. 1982. Predictive coding: A fresh view of inhibition in the retina. *Proc. R. Soc. London Ser. B* **216**, 427–459.

Uttley, A. M. 1979. *Information Transmission in the Nervous System*. Academic Press, London.

Received 9 February 90; accepted 10 June 90.