Information maximization in a network of linear neurons

Holger Arnold*

1. Introduction. It is known since the work of Hubel and Wiesel [3], that many cells in the early visual areas of mammals have characteristic response properties. For example, there are cells sensitive to light-dark contrast and cells sensitive to edges of a certain orientation. In a series of articles [4–6], Linkser showed experimentally that similar response properties can be developed in a simple feed-forward network of linear neurons by using a Hebbian learning rule [2]. The structures that emerged on the different layers of Linsker's network were spatial-opponent (also called center-surround) cells, orientation-selective cells, and orientation columns. Interestingly, these structures emerged without any structured input to the network. This is an important aspect because cells with such characteristics have been found in monkeys even before birth, i.e., before they could have experienced any visual input.[1]

Linsker showed in [7–9] that the self-organization process leading to these receptive field structures is consistent with the principle of maximum information preservation, or "infomax" principle for short. In the context of neural networks, the infomax principle states that the transfer function from one neural layer to the next should preserve as much information as possible. If we call the state of the first layer the stimulus and the state of the second layer the response, then the infomax principle states that the response should contain as much information about the stimulus as possible.

This text is intended as an overview of a part of Linsker's work. The analysis makes use of some concepts from information theory. The book by Cover and Thomas [1] is an extensive (and expensive) reference on that subject; Shannon's original paper [10] is also a very good reading (and is freely available).

2. Optimal receptive fields for linear neurons. We consider a two-layered linear feed-forward network. Let $X = (X_1, \ldots, X_N)$ be a random vector denoting the state of the input or stimulus layer and let $Y = (Y_1, \ldots, Y_M)$ be a random vector denoting the state of the output or response layer. The mutual information between X and Y is defined as

$$I(X;Y) = H(Y) - H(Y|X). \tag{1}$$

In Shannon's terms, $I(X;Y)$ is the information rate of the channel between X and Y. The higher the value of $I(X;Y)$, the more information the response contains about the

Date: May 30 2005.

*See `http://harnold.org/` for contact information.

[1] Just as a side remark: In the time of Hubel's and Wiesel's work, the response properties of neurons were believed to be changing only on a time scale much slower than the time scale of the activity dynamics. Recently, however, it has been found that they can also change on a rather fast time scale. Considering the receptive fields as time-invariant for the process at hand should therefore be seen as a simplification.

stimulus. $H(Y)$ is the information entropy of Y; it is defined as

$$H(Y) = -\int p_Y(y) \log p_Y(y)\, \mathrm{d}y, \tag{2}$$

where p_Y is the probability density function of Y and the integral is taken over all possible values of Y. We assume that the random variables $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_M$ take real values. $H(Y|X)$ is the conditional entropy of Y given X; it is defined as

$$\begin{aligned} H(Y|X) &= \int p_X(x) H(Y|x)\, \mathrm{d}x \\ &= -\int p_X(x) \int p_{Y|x}(y|x) \log p_{Y|x}(y|x)\, \mathrm{d}y\, \mathrm{d}x. \end{aligned} \tag{3}$$

2.1. A model with additive noise. The output of every unit in our network is a weighted linear summation of its inputs. From a biological point of view, this is of course not a completely realistic assumption, but it simplifies our analysis (and it is Linsker's assumption anyway). We assume that the stimuli are distributed according to a multivariate normal distribution with mean zero:

$$p_X(x) = (2\pi)^{-N/2} (\det C_X)^{-1/2} \exp\left(-\frac{1}{2} x^t C_X^{-1} x\right). \tag{4}$$

$C_X = (c_{ij}^X)_{ij}$ with $c_{ij}^X = \int p_X(x) x_i x_j\, \mathrm{d}x$ and $x = (x_1, \ldots, x_N)$ is the positive definite covariance matrix of the stimulus distribution; x^t denotes the transpose of x. In biological networks, noise plays a considerable role and should therefore not be neglected. As a first attempt, we assume that the components of the response vector are disturbed by independent additive noise of constant variance. In this case, the response can be written as

$$Y = WX + R \tag{5}$$

for some weight matrix $W = (w_{ij})_{ij}$. The element w_{ij} is the strength of the connection from unit j in the stimulus layer to unit i in the response layer. $R = (R_1, \ldots, R_M)$ denotes the noise vector. We assume that $R_1, \ldots, R_M$ are i.i.d. random variables distributed according to a normal distribution with mean zero and common variance ρ:

$$p_{R_i}(r_i) = (2\pi\rho)^{-1/2} \exp\left(-\frac{r_i^2}{2\rho}\right), \quad i = 1, \ldots, M. \tag{6}$$

Because the elements of R are independent, R is distributed according to a multivariate normal distribution with mean zero and diagonal covariance matrix $C_R = (c_{ij}^R)_{ij}$, where $c_{ij}^R = \delta_{ij}\rho$ (δ being the Kronecker delta). Y is a linear combination of independent normally distributed random vectors and is therefore normally distributed as well; its

probability density function is

$$p_Y(y) = (2\pi)^{M/2}(\det C_Y)^{-1/2} \exp\left(-\frac{1}{2} y^t C_Y^{-1} y\right) \tag{7}$$

with positive definite covariance matrix $C_Y = W C_X W^t + C_R$ and mean zero.

To maximize the mutual information $I(X;Y)$, we need to compute the entropies $H(Y)$ and $H(Y|X)$. The entropy of the normally distributed random vector Y is

$$H(Y) = \log\left((2\pi e)^{M/2}(\det C_Y)^{1/2}\right). \tag{8}$$

See [10] for a derivation of this result. Note that when a stimulus value x is fixed, the covariance matrix of the conditional response vector $(Y|x)$ is equal to that of the noise vector R. Therefore, the conditional entropy $H(Y|X)$ can be expressed as

$$\begin{aligned} H(Y|X) &= \int p_X(x) H(Y|x)\, \mathrm{d}x \\ &= \int p_X(x) \log\left((2\pi e)^{M/2}(\det C_R)^{1/2}\right)\, \mathrm{d}x \\ &= \log\left((2\pi e)^{M/2}(\det C_R)^{1/2}\right) \\ &= \log\left((2\pi e)^{M/2}\rho^{M/2}\right), \end{aligned} \tag{9}$$

where the last line follows because C_R is a diagonal matrix with M-fold eigenvalue ρ. By inserting equations (8) and (9) into equation (1), we obtain

$$\begin{aligned} I(X;Y) &= \log\left(\rho^{-M/2}(\det C_Y)^{1/2}\right) \\ &= \frac{1}{2}\log\left(\det\left(\rho^{-1} C_Y\right)\right). \end{aligned} \tag{10}$$

Note that $I(X;Y)$ becomes infinite as ρ goes to 0. This effect is a result of our use of continuous probability distributions, which can have infinite entropy. Maximizing $I(X;Y)$ given a certain noise level ρ is equivalent to maximizing the determinant of the covariance matrix of the response distribution, which is equal to the product of its eigenvalues. The elements of $C_Y = (c_{ij}^Y)_{ij}$ are given by

$$c_{ij}^Y = \sum_{m=1}^{N} w_{jm} \sum_{l=1}^{N} w_{il} c_{lm}^X + \delta_{ij}\rho. \tag{11}$$

Without bounding the elements of the weight matrix W, $I(X;Y)$ could become arbitrarily large, and the effect of noise could be made arbitrarily small. We can avoid this by

normalizing the connections to each response unit using the constraints

$$\sum_{j=1}^{N} w_{ij}^2 = 1 \text{ for all } i = 1, \ldots, M. \tag{12}$$

In the next section, we shall see how we can get rid off this normalization condition by modifying our noise model.

To get more analytical results, we need to introduce a few more assumptions. We assume that stimulus and response layer have the same size: $N = M$ (this assumption can be generalized to $N = kM$ for $k \in \mathbb{N}$). Let the units of both layers be uniformly arranged along two rings (the analysis can be extended to higher dimensions, leading to n-tori, and to an infinite set of units). From this arrangement, we obtain a natural displacement function. Let d_{ij} denote the displacement from unit i to unit j on the ring. In our case, we simply define $d_{ij} = [j - i]$ with $[k] = k \bmod N$. The displacement can be computed between units on the same layer, as well as between units on different layers (by projecting one ring onto the other). Now we assume that the covariance c_{ij}^X between two elements i and j of a random vector from the stimulus distribution X is a function of the displacement d_{ij} from i to j: $c_{ij}^X = c_{d_{ij}}^X$. Note that due to the ring structure of the units, c^X is a periodic function. Because C_X is a symmetric matrix, c^X must satisfy $c_{d_{ij}}^X = c_{d_{ji}}^X$. C_X is a cyclic matrix; therefore, its eigenvalues $\lambda_1, \ldots, \lambda_N$ are given as the components of the discrete Fourier transform of the sequence $c_0^X, \ldots, c_{N-1}^X$:

$$\lambda_k = \mathcal{F}_k \left(c_j^X\right)_j = \sum_{j=0}^{N-1} c_j^X \omega_N^{kj}, \quad k = 1, \ldots, N, \tag{13}$$

with $\omega_N^{kj} = e^{-\frac{2\pi i}{N} kj}$. Further, we assume that the strength w_{ij} of the connection from unit j in the stimulus layer to unit i in the response layer is also a periodic function of the displacement d_{ij} from i to j: $w_{ij} = w_{d_{ij}}$. This means that every unit in the response layer develops the same receptive field structure. Under this condition C_Y is a cyclic matrix as well. Its eigenvalues $\mu_1, \ldots, \mu_N$ are the components of the discrete Fourier transform of the sequence $c_0^Y, \ldots, c_{N-1}^Y$:

$$\mu_k = \mathcal{F}_k \left(c_j^Y\right)_j = \sum_{j=0}^{N-1} c_j^Y \omega_N^{kj}, \quad k = 1, \ldots, N. \tag{14}$$

With equation (11) and the definitions of the functions c^X and w, we get

$$\mu_k = \mathcal{F}_k \left(\sum_{m=0}^{N-1} w_{m-j} \sum_{l=0}^{N-1} w_l \, c_{m-l}^X \right)_j + \rho. \tag{15}$$

Let $(f * g)_k$ denote the convolution of the functions f and g at point k. Then, we can express the last equation as

$$\mu_k = \mathcal{F}_k \left(\sum_{m=0}^{N-1} w_{m-j} \left(w * c^X \right)_m \right)_j + \rho, \tag{16}$$

and with $w_k^r = w_{-k}$ as

$$\begin{aligned} \mu_k &= \mathcal{F}_k \left(w^r * w * c \right) + \rho \\ &= \mathcal{F}_k \left(w^r \right) \mathcal{F}_k \left(w \right) \mathcal{F}_k \left(c^X \right) + \rho. \end{aligned} \tag{17}$$

Note that $\mathcal{F}_k \left(w^r \right) = \overline{\mathcal{F}_k \left(w \right)}$. With $z_k = |\mathcal{F}_k \left(w \right)|^2$, and equation (13), we have

$$\mu_k = \lambda_k z_k + \rho. \tag{18}$$

By inserting the eigenvalues of C_Y into equation (10), we obtain

$$I(X;Y) = \frac{1}{2} \sum_{k=1}^{N} \log \left(1 + \frac{\lambda_k z_k}{\rho} \right). \tag{19}$$

To maximize $I(X;Y)$, we use the method of Lagrange multipliers. The constraints in equation (12) can be equivalently expressed as $\sum_{k=1}^{N} z_k = N$. We define the Lagrange function L as

$$L(z_1, \ldots, z_N, \alpha) = I(X;Y) + \alpha \left(\sum_{k=1}^{N} z_k - N \right), \tag{20}$$

where α is the Lagrange multiplier. The partial derivatives of L w.r.t. the z_k are

$$\frac{\partial L}{\partial z_k} = \frac{\lambda_k}{2} (\rho + \lambda_k z_k)^{-1} + \alpha, \quad k = 1, \ldots, N. \tag{21}$$

Solving $\partial L / \partial z_k^0 = 0$ gives the stationary points

$$z_k^0 = -\frac{1}{2\alpha} - \frac{\rho}{\lambda_k}, \quad k = 1, \ldots, N. \tag{22}$$

The second partial derivatives of $I(X;Y)$ w.r.t. z_k and z_l are

$$\frac{\partial^2 I}{\partial z_k \partial z_l} = \begin{cases} -\frac{\lambda_k^2}{2} (\lambda_k z_k + \rho)^{-2} & \text{if } k = l \\ 0 & \text{otherwise} \end{cases}, \tag{23}$$

and the quadratic form $Q_{H_I}(x) = x^t H_I(z_1, \ldots, z_N) x$ of the Hessian matrix $H_I(z_1, \ldots, z_N) =$

$(h_{kl})_{kl}$ of $I(X;Y)$ is

$$\begin{aligned} Q_{H_I}(x) &= \sum_{k=1}^{N}\sum_{l=1}^{N} h_{kl}x_k x_l \\ &= -\sum_{k=1}^{N} \frac{\lambda_k^2}{2}(\lambda_k z_k + \rho)^{-2} x_k^2. \end{aligned} \tag{24}$$

Because C_X and C_Y are positive definite, $Q_{H_I}(x) < 0$ for all x and $H_I(z_1, \ldots, z_N)$ is negative definite at all points. Thus, $I(X;Y)$ is a strictly concave function and its maximum values w.r.t. $z_1, \ldots, z_N$ under the weight constraints are uniquely determined by equation (22). Note that z must be non-negative at all points. Because $I(X;Y)$ is strictly concave and $\partial I(X;Y)/\partial z_k = \frac{\lambda_k}{2}(\rho + \lambda_k z_k)^{-1} > 0$ for all k, the values $\hat{z}_1, \ldots, \hat{z}_N$ that maximize $I(X;Y)$ are given by

$$\hat{z}_k = \max\left\{0, -\frac{1}{2\alpha} - \frac{\rho}{\lambda_k}\right\}, \quad k = 1, \ldots, N, \tag{25}$$

where α is chosen so that $\sum_{k=1}^{N} z_k = 1$. What we actually want, however, is to determine the weights $\hat{w}_0, \ldots, \hat{w}_{N-1}$ that maximize $I(X;Y)$. From the definition of z, we can see that $I(X;Y)$ is invariant under weight transformations that do not change the absolute values of their Fourier transform. For example, $I(X;Y)$ is unaffected by rotations of the weights and by a change of their sign. Consequently, further constraints on the weights are necessary to uniquely determine their optimal values.

2.2. A model with independent noise on each input. The model developed in the previous section has two problems. The first one is the rather arbitrary normalization constraint on the weights. The second one is the under-determinedness of the weights maximizing the mutual information $I(X;Y)$. In this section, we shall see that we can eliminate both problems by modifying our noise model.

So far, we have assumed that the response units are disturbed by independent noise of constant variance ρ. Now, we assume that the signal of each input unit is disturbed by independent noise and that the noise variance can have a different value for each connection. Thus, we have

$$Y_i = \sum_{j=1}^{N} w_{ij}(X_j + R_{ij}), \quad i = 1, \ldots, M, \tag{26}$$

where R_{ij} is a random variable distributed according to a normal distribution with mean zero and variance ρ_{ij}. Then, Y still has a multivariate normal distribution with mean zero and covariance matrix $C_Y = WC_XW^t + C_R$. The elements of $C_R = \left(c_{ij}^R\right)_{ij}$ are

given by

$$c_{ij}^R = \delta_{ij} \sum_{k=1}^{N} w_{ik}^2 \rho_{ik}. \tag{27}$$

As in our first model, C_Y is a cyclic matrix whose eigenvalues $\mu_1, \ldots, \mu_N$ are the components of the Fourier transformation of the sequence $c_0^Y, \ldots, c_{N-1}^Y$.

(a) Same noise level for all connections. With $\rho_{ij} = \rho$ for all i, j, we have

$$\mu_k = \lambda_k z_k + \rho \sum_{j=0}^{N-1} w_j^2, \tag{28}$$

and

$$I(X;Y) = \frac{1}{2} \sum_{k=1}^{N} \log \left(1 + \frac{\lambda_k z_k}{\rho \sum_{j=0}^{N-1} w_j^2} \right). \tag{29}$$

The sum over the squares of the weights functions as a regularization term, which penalizes large absolute weight values. In this model, the weight constraints from equation (12) are no longer necessary.

(b) Noise level increasing with connection length. To eliminate the underdeterminedness of the optimum weights, we apply a biologically relevant principle: shorter connections should be favored over longer connections. We can achieve this by increasing the noise level with the connection length. Let g be an N-periodic function that is monotonic increasing from 0 to $N/2$ and monotonic decreasing from $N/2$ to N (in other words, g_{j-i} increases with the absolute distance between units i and j). Then, with $\rho_{ij} = \rho_0 g_{j-i}$, we have

$$\mu_k = \lambda_k z_k + \rho_0 \sum_{j=0}^{N-1} g_j w_j^2, \tag{30}$$

and

$$I(X;Y) = \frac{1}{2} \sum_{k=1}^{N} \log \left(1 + \frac{\lambda_k z_k}{\rho_0 \sum_{j=0}^{N-1} g_j w_j^2} \right). \tag{31}$$

2.3. The effect of noise on the redundancy in the optimal responses. Let us study the network from a slightly different point of view. We consider only two response units and analyze how the noise level effects their optimal response. We use our first noise model from section 2.1. For $M = 2$, we get

$$I(X;Y) = \frac{1}{2} \log(\det C_Y) - \log \rho \tag{32}$$

with $\det C_Y = c_{11}^Y c_{22}^Y - (c_{12}^Y)^2$. Let $v_i = c_{ii}^Y - \rho$, $i = 1, 2$, be the variance of Y_i when no noise is present and let $r_{12} = c_{12}^Y / \sqrt{v_1 v_2}$ be the correlation coefficient of Y_1 and Y_2 when no noise is present. Then we can express the determinant of C_Y as

$$\det C_Y = \rho^2 + \rho(v_1 + v_2) + v_1 v_2 (1 - r_{12}^2). \tag{33}$$

References

1. Thomas M. Cover and Joy Thomas. *Elements of Information Theory*. Wiley, 1991.

2. Donald O. Hebb. *The Organization of Behavior*. Wiley, 1949.

3. David H. Hubel and Torsten N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, 195(1):215–243, 1968.

4. Ralph Linsker. From basic network principles to neural architecture: Emergence of orientation columns. *Proceedings of the National Acadamy of Sciences USA*, 83: 8779–8783, 1986.

5. Ralph Linsker. From basic network principles to neural architecture: Emergence of orientation-selective cells. *Proceedings of the National Acadamy of Sciences USA*, 83:8390–8394, 1986.

6. Ralph Linsker. From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Acadamy of Sciences USA*, 83: 7508–7512, 1986.

7. Ralph Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21(3): 105–117, 1988.

8. Ralph Linsker. An application of the principle of maximum information preservation to linear systems. *Advances in Neural Information Processing Systems*, 1:186–194, 1989.

9. Ralph Linsker. Deriving receptive fields using an optimal encoding criterion. *Advances in Neural Information Processing Systems*, 5:953–960, 1992.

10. Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 1948.