Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 10749–10753, November 1993
Neurobiology

Simple models for reading neuronal population codes

(population vector/maximum likelihood/tuning curves/perceptual learning/orientation)

H. S. SEUNG AND H. SOMPOLINSKY

AT&T Bell Laboratories, Murray Hill, NJ 07974 and Racah Institute of Physics and Center for Neural Computation, Hebrew University, Jerusalem 91904, Israel

Communicated by Frank H. Stillinger, August 11, 1993

ABSTRACT **In many neural systems, sensory information is distributed throughout a population of neurons. We study simple neural network models for extracting this information. The inputs to the networks are the stochastic responses of a population of sensory neurons tuned to directional stimuli. The performance of each network model in psychophysical tasks is compared with that of the optimal maximum likelihood procedure. As a model of direction estimation in two dimensions, we consider a linear network that computes a population vector. Its performance depends on the width of the population tuning curves and is maximal for width, which increases with the level of background activity. Although for narrowly tuned neurons the performance of the population vector is significantly inferior to that of maximum likelihood estimation, the difference between the two is small when the tuning is broad. For direction discrimination, we consider two models: a perceptron with fully adaptive weights and a network made by adding an adaptive second layer to the population vector network; We calculate the error rates of these networks after exhaustive training to a particular direction. By testing on the full range of possible directions, the extent of transfer of training to novel stimuli can be calculated. It is found that for threshold linear networks the transfer of perceptual learning is nonmonotonic. Although performance deteriorates away from the training stimulus, it peaks again at an intermediate angle. This nonmonotonicity provides an important psychophysical test of these models.**

Empirical studies of neuronal response are yielding increasing knowledge of its statistical properties (see, e.g., refs. 1 and 2). The goal of relating these properties to psychophysical thresholds involves several basic questions. First, given the response of a population of neurons to a stimulus, what are the optimal procedures for performing tasks such as estimation of stimulus parameters or discrimination between two stimuli (3, 4)? Second, what are plausible neuronal mechanisms that "read" the neuronal responses and perform these tasks? Finally, how does the performance depend on the tuning curve properties of the population? For a large population of neurons whose fluctuations are statistically independent, maximum likelihood (ML) procedures are optimal. The dependence of ML estimation error and discrimination thresholds on the population size N is well known to be $1/\sqrt{N}$(5, 6). Using the Fisher information (5) as a tool, we study the dependence of the ML performance on the tuning curve properties in the context of neurons coding for the direction of a stimulus.

Although the ML procedures provide important theoretical bounds on actual performance, in general they do not seem to have plausible neural implementations. We study simple models, linear and threshold linear networks, for estimation and discrimination of directional stimuli and compare their performance with the ML procedures. For direction estimation we focus on a network that computes a population vector by summing the preferred directions of the neurons weighted by their response magnitudes. Some experimental evidence for this scheme has been found in the generation of saccadic eye movements in primates (7). It has also been suggested as a code for the direction of arm movements (8) and as a model of visual orientation estimation (9–11).

Here we study the performance of the population vector relative to the optimal ML estimation. An important outcome of our analysis of direction discrimination is that threshold linear models require adaptation to perform well. We calculate theoretical generalization curves for the amount of transfer of learning from a trained stimulus to novel stimuli. Testing these predictions by psychophysical measurements could shed light on the neuronal mechanisms involved in perception and perceptual learning (12–14).

ML PROCEDURES

Population of Direction Selective Neurons. We consider a population of neurons coding for direction in two dimensions, parametrized by θ from 0 to 2π. For example, these could be simple cells in visual cortex coding the direction of motion of a bar stimulus. We characterize the response of the ith neuron by a single nonnegative integer r_i, the total number of spikes generated by the neuron in a fixed time interval following the onset of the stimulus. Our starting point is the assumption that the response of a neuron to a sensory stimulus is stochastic—namely, that repeated presentations of the same stimulus θ induce responses that vary in a random fashion. The response of a population of N neurons is described by a conditional probability distribution $\mathcal{P}(\mathbf{r}|\theta)$, where the vector notation $\mathbf{r}$ is used for the responses $r_1, \ldots, r_N$.

We model the responses $\{r_i\}$ of the population as independent Poisson random variables. The mean of the spike count of the ith neuron is denoted by $\langle r_i \rangle = f_i(\theta)$, where $\langle \cdots \rangle$ denotes an average with respect to $\mathcal{P}(\mathbf{r}|\theta)$. The variance of a Poisson process equals its mean—i.e., $\langle(\delta r_i)^2\rangle = f_i(\theta)$—where $\delta r_i = r_i - \langle r_i \rangle$. A similar linear relationship between the mean and variance of neural responses in cortex has been observed (1, 2), although with coefficient of proportionality different from 1.

The dependence of the mean response on the stimulus direction is called the direction tuning curve of the cell. In our idealized population the tuning curves all have the same shape, $f_i(\theta) = f(\theta - \theta_i)$, where $f(\theta)$ is assumed to be an even periodic function of θ with a maximum at $\theta = 0$. The angle θ_i denotes the preferred direction of the ith neuron. The preferred directions θ_i are evenly spaced on the circle, $\theta_i = 2\pi i/N$, and $Na/2\pi \gg 1$, where a is the width of f. For concreteness, we will analyze model tuning curves of the form

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: ML, maximum likelihood.

$$f(\theta) = \begin{cases} f_{\min} + (f_{\max} - f_{\min})\cos^m(\pi\theta/2a), & |\theta| < a \\ f_{\min}, & \text{otherwise.} \end{cases} \qquad [1]$$

If the stimulus θ is close to $\theta_i \pm a$, we say that the neuron is near *threshold*. Thus the exponent m controls the rise of the tuning curve at threshold, and the parameter a controls its width. The cases of $m = 1$ and $m = 2$ are shown in Fig. 1*A*.

ML Estimation. An estimation task is one that requires an estimate of a continuously varying stimulus parameter. Here we assume that the animal's estimate $\hat{\theta}$ is based on the responses $\mathbf{r}$ of the neuronal population—i.e., can be written as a function $\hat{\theta}(\mathbf{r})$. The mean-square error of this estimate can be due to both systematic bias $\langle\hat{\theta}(\mathbf{r})\rangle - \theta$ and the fluctuations of $\mathbf{r}$ from trial to trial.

The ML estimate is the value of θ that maximizes the likelihood $\mathcal{P}(\mathbf{r}|\theta)$. For a large population, its variance is given by $\langle(\hat{\theta} - \theta)^2\rangle = 1/J[\mathbf{r}](\theta)$, where $J[\mathbf{r}](\theta)$ is the Fisher information, defined as

$$J[\mathbf{r}](\theta) = \left\langle -\frac{\partial^2}{\partial\theta^2} \log \mathcal{P}(\mathbf{r}|\theta) \right\rangle. \qquad [2]$$

The Fisher information is a functional of $\mathcal{P}(\mathbf{r}|\theta)$ and can be interpreted as the amount of information in $\mathbf{r}$ about the stimulus θ. Because of the independence of the r_i, $J[\mathbf{r}](\theta) = \Sigma_{i=1}^{N} J[r_i](\theta)$, so that J is of the order of N, implying that the typical fluctuation of the ML estimate scales as $\hat{\theta} - \theta \propto N^{-1/2}$. This is in contrast to the bias of the ML estimate, which is of the order of $1/N$. Hence the variance is the dominant contribution to the error in the large N limit. The information contained in the response of neuron i is $J[r_i](\theta) = f_i'(\theta)^2/f_i(\theta)$, which can be interpreted as the square of the signal to noise ratio of the neuron. The total information is given by the sum of $J[r_i](\theta)$ for all neurons, which for large N is

$$J[\mathbf{r}] = N \int_0^{2\pi} \frac{d\phi}{2\pi} \frac{f'(\phi)^2}{f(\phi)}. \qquad [3]$$

Note that because of the isotropic distribution of the preferred directions θ_i, $J[\mathbf{r}]$ is the same for any stimulus θ in the continuum limit.

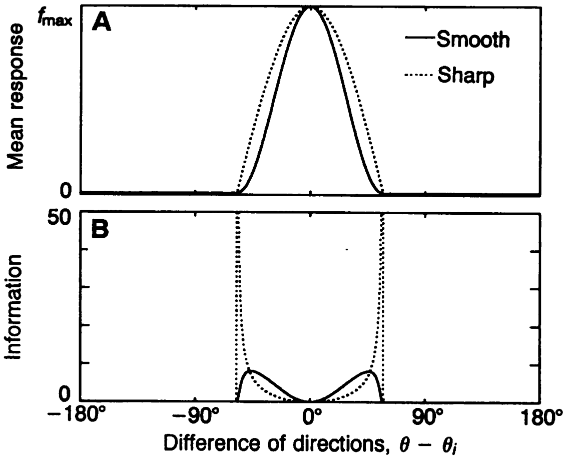

FIG. 1. (*A*) Two tuning curves of the form given in Eq. **1**. Both smooth ($m = 2$, solid line) and sharp ($m = 1$, dotted line) thresholds are shown. The ratio of background to peak response is $\rho = f_{\min}/f_{\max} = 0.01$, and the width is $a = 1$. (*B*) Information $J[r_i](\theta)$ in neuron i as a function of $\theta_i - \theta$ for tuning curves with $a = 1$ and $\rho = 0.01$. There is no information in the neurons with $\theta_i \approx \theta$, at their maximal firing rates. For the sharp threshold population, the peak of J is at $|\theta - \theta_i| = a$ and extends well beyond the top of the figure.

By the Cramér-Rao inequality (15), no unbiased estimator can have smaller variance than $1/J$. Hence the ML estimate is asymptotically optimal, since its variance saturates this bound in the large N limit.

ML Discrimination. A discrimination task involves a finite set of alternatives rather than a smoothly varying parameter. In each trial of a *single interval* discrimination, either stimulus θ or $\theta + \delta\theta$ is presented at random and the task is to determine which of the two stimuli was presented. Given the response $\mathbf{r}$, the ML discrimination is according to which likelihood, $\mathcal{P}(\mathbf{r}|\theta)$ or $\mathcal{P}(\mathbf{r}|\theta + \delta\theta)$, is greater. In a *two-interval* discrimination (also known as two-alternative forced choice), each trial contains a presentation of both stimuli in random order, and the task is to determine in which order they were presented. Here ML weighs the relative likelihoods of the two orders. For a large population of uncorrelated neurons, it can be shown that the probabilities of error for ML discrimination are $H(d'/2)$ for *single-interval* and $H(d'/\sqrt{2})$ for *two-interval*, where $H(x) = (2\pi)^{-1/2} \int_x^\infty dx\, e^{-x^2/2}$ is the area under the normal distribution between x and infinity. The quantity d' is the *discriminability* d' of the two stimuli and is given by

$$d' = |\delta\theta| \sqrt{J[\mathbf{r}](\theta)}, \qquad [4]$$

provided that $|\delta\theta|$ is scaled so that d' is of order unity in the large N limit. This is the relevant scaling for psychophysical experiments, in which stimulus differences are adjusted so that discrimination error is neither too small nor too large. These results reflect two facts. First, for large N, the ML discrimination between two nearby stimuli on the basis of the response $\mathbf{r}$ is equivalent to one based on the ML estimate $\hat{\theta}(\mathbf{r})$. Second, the fluctuations in $\hat{\theta}$ are asymptotically normal. If the two alternatives have equal prior probabilities of presentation, ML discrimination is optimal.

Threshold Effect. A striking feature of Eq. **3** is its sensitivity to the shape of $f(\theta)$ near threshold. Most importantly, if the ratio of background to peak response $\rho \equiv f_{\min}/f_{\max}$ is small, and the slope $f'(\theta)$ remains finite as the mean response $f(\theta)$ drops to $f_{\min}$, neurons just above threshold can carry anomalously high information, and their contribution may dominate the total information $J[\mathbf{r}]$. For example, for the tuning curve given by Eq. **1** with $m = 1$, the neurons at the threshold contribute the maximal information (per neuron), $\max_i\{J[r_i]\} = (\pi/2a)^2/\rho$, which can be large if ρ is small. This *threshold effect* is evident in the sharp peaks at the thresholds seen in Fig. 1*B*, where we plot the information in neuron i as a function of preferred direction θ_i. In contrast, there is no threshold effect for the tuning curve with *smooth threshold* corresponding to Eq. **1** with $m = 2$ because $f'(\theta)$ vanishes at threshold. Instead, the most informative neurons are separated from the threshold by $a\sqrt{\rho}$ for small ρ, as demonstrated in Fig. 1*B*.

The total information is given by the sum of the $J[r_i]$, which is the area under the curves of Fig. 1*B*. For the *smooth threshold* ($m = 2$) case, the total $J[\mathbf{r}]$ is not sensitive to the value of ρ for $\rho \ll 1$. In this case, $J[\mathbf{r}] \approx Nf_{\max}\pi/2a$, which is finite for all values of $a > 0$. This is demonstrated in Fig. 2, where we present J for the $m = 2$ tuning curve as a function of a for $\rho = 0.01$ and 0.1. In contrast, for the *sharp threshold* tuning curve ($m = 1$) the total information is $J[\mathbf{r}] \approx (Nf_{\max}/8a)|\log \rho|$. Thus, the contribution of the neurons near the threshold of activity leads to a logarithmic divergence of J as $f_{\min} \rightarrow 0$. Finally, for all values of m, $J \propto a^{-1}$, implying that narrower tuning improves performance.

MODELS OF "READOUT"

Direction Estimation by Population Vector. Although ML estimation is asymptotically optimal, it is not clear whether

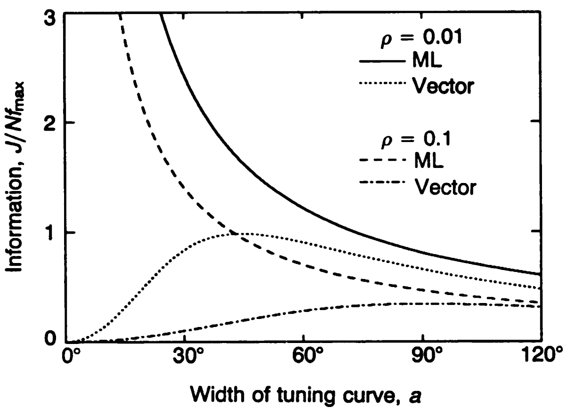

FIG. 2. Total information $J[\mathbf{r}]$ and information in the population vector $J[\hat{z}]$ as functions of a, for the smooth threshold ($m = 2$) tuning curve of Eq. 1 with $\rho = 0.01$ and 0.1. For large a, $J[\hat{z}]$ behaves like $J[\mathbf{r}]$, but as $a \to 0$, $J[\mathbf{r}]$ diverges, while $J[\hat{z}]$ falls to zero. The peak in $J[\hat{z}]$ is broad and shifts to larger a as ρ increases (see Eq. 7).

it possesses a plausible biological implementation. A biologically plausible alternative to ML is the population vector (7–11), which can be interpreted as a linear unbiased estimator of the 2d vector representation of the stimulus, ($\cos\theta$, $\sin\theta$). In the following it is more convenient to use the equivalent complex number representation $z = e^{i\theta}$. The population vector is given by the sum of N complex numbers, each pointing in the preferred direction of a neuron and weighted by its response,

$$\hat{z} = \frac{1}{N|\tilde{f}_1|} \sum_{k=1}^{N} r_k e^{i\theta k}. \qquad [5]$$

In the prefactor, $\tilde{f}_1$ is the first Fourier component of f, where the nth Fourier component is defined by $\tilde{f}_n = (2\pi)^{-1} \int_0^{2\pi} d\theta\, e^{in\theta} f(\theta)$. For large N, $\langle\hat{z}\rangle = (\tilde{f}_1/|\tilde{f}_1|)e^{i\theta}$. If $f(\theta)$ is nonnegative and has a single maximum at the origin, as in the cases considered here, $\tilde{f}_1$ is real and positive, $\langle\hat{z}\rangle = z$, and the population vector (Eq. 5) is an unbiased estimator of z. The population vector can be interpreted as the output of a single layer network with two linear nodes whose weight vectors are proportional to $\cos\theta_i$ and $\sin\theta_i$, respectively.

Due to the Poisson fluctuations of r_k in Eq. 5, both the magnitude $\hat{R}$ and direction $\hat{\theta}$ of $\hat{z} \equiv \hat{R}e^{i\hat{\theta}}$ fluctuate. The performance of the population vector in the estimation of *direction* is measured by the variance of the directional fluctuations. This can be calculated by considering the fluctuations of $\hat{z}$ perpendicular to z. Alternatively, one can calculate the information in $\hat{z}$ about θ, using Eq. 2 and the fact that for large N, $\hat{z}$ has a two-dimensional Gaussian distribution. The result is

$$\langle(\hat{\theta} - \theta)^2\rangle^{-1} = J[\hat{z}] = N\frac{2\tilde{f}_1^2}{\tilde{f}_0 - \tilde{f}_2}. \qquad [6]$$

An important outcome of Eq. 6 is that the performance of the population vector is insensitive to the shape of the tail of the tuning curve. It depends mainly on the width a and the background noise $f_{\min}$. As a result, $J[\hat{z}]$ is almost identical for the sharp and smooth threshold tuning curves. This can be understood from Eq. 6, which can be interpreted as the square of the ratio of a population-averaged signal to a population-averaged noise. Because of this averaging, the population vector is not as sensitive to the presence of highly informative neurons near threshold, as is the full information, Eq. 3, which sums the squared signal to noise ratios of the individual neurons.

The population vector and the ML estimator also differ in their dependence on the width a of the tuning curve and on the ratio ρ of background to peak activity, for small values of a and ρ. Whereas $J[\mathbf{r}]$ diverges as $a \to 0$ like $J \sim Nf_{\max}a^{-1}$ relatively independently of ρ (for the smooth threshold case), $J[\hat{z}]$ behaves roughly as $J[\hat{z}] \sim Nf_{\max}a^2/(a^3 + \rho)$. Thus, $J[\hat{z}]$ vanishes as $a \to 0$ for any fixed $\rho > 0$ and has a maximum at a finite value of a (see Fig. 2). The location of the maximum increases with ρ as

$$a_{\max} \sim \rho^{1/3} \qquad [7]$$

and the value of $J[\hat{z}]$ at $a_{\max}$ grows with decreasing ρ as $J[\hat{z}] \propto \rho^{-1/3}$. This behavior of the population vector results from the fact that for small a most neurons are below threshold and produce noise, with variance $f_{\min}$, without contributing to the signal. In contrast, the total information $J[\mathbf{r}]$ increases for small a because ML can make use of the gradient of the response. As the tuning curves become narrower, the increase in signal $|f'|$ more than offsets the decrease in the number of neurons above threshold. In addition, the neurons below threshold are completely ignored by the ML estimator.

Discrimination by a Threshold Linear Network. The simplest neural network that is capable of performing discrimination between two directions is a single-layer perceptron, which computes a linear sum of its inputs, followed by a thresholding. For concreteness we consider a two-interval discrimination task, where the network has to signal 1 if θ preceded $\theta + \delta\theta$ and -1 otherwise. For this task, the output of the perceptron is assumed to be $\mathrm{sgn}(R - R')$ where

$$R = \sum_{i=1}^{N} \omega_i r_i, \qquad R' = \sum_{i=1}^{N} \omega_i r_i' \qquad [8]$$

and $\{r_i\}$ and $\{r_i'\}$ are the neural responses to the first and second stimuli, respectively. This model presupposes the existence of a short-term memory that stores the summed response R of the first stimulus. By evaluating the probability of error of such a decision rule, we find that for large N the error of the perceptron is given by $H(d'/\sqrt{2})$ with discriminability $d' = |\delta\theta|\sqrt{J[R](\theta_0)}$ and $J[R]$ is the information in R,

$$J[R](\theta) = \frac{[\Sigma_{i=1}^{N}\, \omega_i f_i'(\theta)]^2}{\Sigma_{i=1}^{N}\, \omega_i^2 f_i(\theta)}. \qquad [9]$$

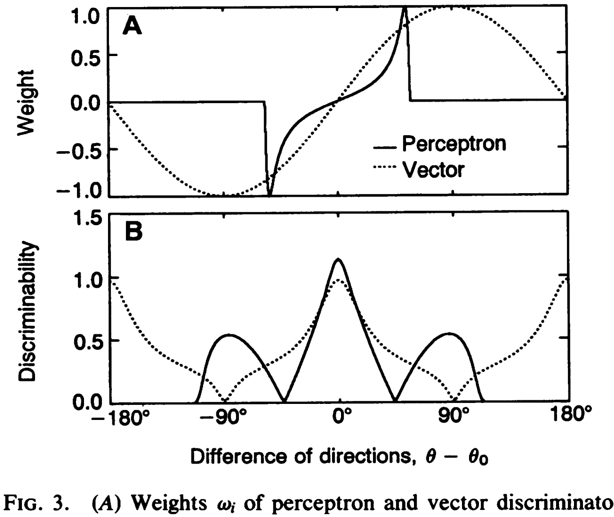

FIG. 3. (A) Weights ω_i of perceptron and vector discriminator adapted to θ_0, for the smooth threshold tuning curves. Each curve is normalized so that the maxima are at ±1. For the perceptron, the largest weights are $f_{\min}^{1/2}$ away from threshold. (B) Discriminability d' in units of $|\delta\theta|\sqrt{Nf_{\max}}$ of stimuli θ and $\theta + \delta\theta$ for the perceptron and vector discriminator after adaptation to θ_0.

The numerator is the square of the "signal" part of R (for discriminating stimuli near θ) and the denominator is the variance of the "noise" in R (evaluated with the Poisson statistics of $\{r_i\}$).

An important property of the perceptron is that for any choice of weights $\{\omega_i\}$, the information $J[R](\theta)$ vanishes for some angle θ, since the numerator of Eq. **9** is the square of a total derivative of a periodic function of θ. This implies that for any fixed choice of weights, the performance of the perceptron will be close to random in some regime of angles. Thus, in order to perform well for all angles θ the weights have to be adapted to the stimulus θ.

Suppose the system is trained to perform a discrimination task around θ_0, allowing unrestricted changes in the weights to maximize the performance at this angle. To calculate the optimal choice of weights we maximize $J[R](\theta_0)$ with respect to ω_i. The result is $\omega_i(\theta_0) \sim f_i'(\theta_0)/f_i(\theta_0)$. Fig. 3*A* shows these weights as a function of $\theta_i - \theta_0$ for the tuning curve of Eq. **1** with $m = 2$. Note that the neurons with maximal response ($\theta_i = \theta_0$) are given zero weight. This is due to their vanishing derivative f_i', which means that they carry little information relevant to discrimination at θ_0. Substituting the optimal ω_i in Eq. **9** demonstrates that for this choice of weights $J[R](\theta_0)$ equals $J[\mathbf{r}]$, Eq. **3**. Thus the discriminability of the stimuli θ_0 and $\theta_0 + \delta\theta$ to a perceptron that is fully adapted to this particular discrimination is identical to that of the ML discrimination. In other words, for $\theta = \theta_0$ there is no loss of information in the transformation $\mathbf{r} \rightarrow R$ for the fully adapted perceptron.

If the network is trained for a given stimulus θ_0, its ability to generalize to $\theta \neq \theta_0$ without further modification of weights is determined by d' or equivalently $J[R](\theta)$ for $\theta \neq \theta_0$. This measure of "transfer" of learning to other angles is shown in Fig. 3*B* as a function of $\theta - \theta_0$. These results show that if a perceptron is fully adapted to θ_0, its ability to generalize decreases rapidly as a function of $\theta - \theta_0$. An important feature of Fig. 3*B* is the nonmonotonicity in the performance of the perceptron. The discriminability drops to zero quickly but then increases again manifesting auxiliary peaks, away from the center, and finally drops to zero again for all values of $|\theta - \theta_0|$ that are larger than $2a$.

Discrimination with Population Vector. The fully adaptive perceptron performs optimally in the adapted angle but requires the adaptation of all its N weights. An alternative network that requires less adaptation can be constructed by adding a second layer to the population vector network. The second layer consists of a single output neuron that sums linearly the outputs of the two components of the population vector with weights c_α, $\alpha = 1, 2$. Thus, this network is equivalent to a single perceptron with N weights $\omega_i = c_1 \mathrm{w}_i^1 + c_2 \mathrm{w}_i^2$, where w_i^1 and w_i^2 are the $2N$ weights of the first layer—namely, $\mathrm{w}_i^1 = \cos\theta_i$, and $\mathrm{w}_i^2 = \sin\theta_i$. Suppose the network learns to perform the discrimination around a particular angle θ_0. It is reasonable to consider a situation in which the first layer of $2N$ weights $(\mathrm{w}_i^1, \mathrm{w}_i^2) = (\cos\theta_i, \sin\theta_i)$ that give rise to the population vector is fixed, and learning occurs only in the second layer weights, c_1 and c_2. The optimal second layer weights can be calculated by maximizing $J[R](\theta_0)$ with respect to c_α, yielding $c_1 = \sin\theta_0$ and $c_2 = -\cos\theta_0$. The meaning of this optimum can be understood by noting that with these values of c_α the weights of the equivalent single layer perceptron are $\omega_i \sim \sin(\theta_i - \theta_0)$. These weights are plotted as a function of $\theta_i - \theta_0$ in Fig. 3*A*. Note that unlike the fully adapted perceptron the weights from the background neurons in the population vector network are not zero.

With the above optimal c_α the performance of the network in the discrimination task at θ_0 is given by $H(d'/\sqrt{2})$ with discriminability $d' = |\delta\theta|\sqrt{J[R](\theta_0)}$, where $J[R](\theta_0)$ is given by Eq. **9** with the weights of the equivalent perceptron ω_i above. Evaluating Eq. **9** with these weights we find that $J[R](\theta)$ is equal to $J[\hat{z}]$, Eq. **6**. Thus at θ_0 the optimal discriminator based on the population vector utilizes all the information in $\hat{z}$. This performance is somewhat inferior to that of the ML discriminator, since $J[\hat{z}] < J[\mathbf{r}]$, as shown in Fig. 2.

The transfer curve for this network is determined by $J[R](\theta)$ for $\theta \neq \theta_0$, which is plotted in Fig. 3*B*. As seen in this figure $J[R]$ is a periodic function of $\theta - \theta_0$ with a periodicity of π. The central maximum has a width of the order of the tuning curve width a. There is no transfer for directions at right angles to θ_0, and transfer is maximal for a direction that is opposite to the direction used in training.

DISCUSSION

Using the Fisher information, we have compared the performance of linear and threshold linear networks with ML, which is optimal for a large uncorrelated population. If the variance of neuronal response is linearly proportional to the mean and background activity is low, as is the case for simple cells in visual cortex, the ML performance exhibits a strong sensitivity to the shape of the tail of the tuning curve, because weakly active neurons have minimal noise. Depending on the shape of this tail the information can be highly concentrated in the neurons near threshold.

A biologically plausible alternative to ML estimation is the population vector, which provides an unbiased estimator of direction. Since the population vector gives relatively uniform weight to every neuron, it depends primarily on the width of the tuning curve and on the activity level of the background neurons and is insensitive to the tails of the tuning curve. The information in the population vector is optimized by a width that increases with ρ, Eq. **7**. A value of $\rho \approx 0.01$ is a reasonable estimate for the ratio of the background to peak activity in simple cells in cat primary visual cortex. For this value of ρ our predicted value of optimal tuning width corresponds to a half width at half max of 24° for direction tuning. Applying our theory to orientation selective cells yields an optimal value of 12°. The maxima in the performance are broad, especially for larger widths (Fig. 2). This is not far from the range 14° to 22° observed experimentally for orientation tuning in simple cells (16). Furthermore, a clear tendency for broader tuning in direction selective cells has been observed (17). We conjecture that a change in the background activity might trigger adaptation in the tuning width so that it remains close to the optimal in accordance with Eq. **7**. It may be possible to test this conjecture experimentally. Although our discussion has addressed explicitly only the coding of sensory stimuli, our analysis of the performance of the population vector can be applied also to the coding of directions of movements in motor systems (7, 8), where population vector codes have been implicated.

In psychophysical experiments on primates, the just noticeable difference in orientation is roughly 0.5 degree (9). Assuming a tuning curve peak of 50 spikes (corresponding to peak response during 500 msec) and width as above, of order 100 neurons are required to yield this performance. This is consistent with simulations by Vogels of a population of neurons with identical tuning properties (9). Our results (Fig. 2) show that, whereas for a narrowly tuned population the population vector code is in general significantly inferior to the ML performance, there is little difference in their performance for a broadly tuned population with smooth tuning curves. For the $m = 2$ tuning curve with the width quoted above, the number of neurons needed for the population code is larger by a factor of 1.4 than the number required to achieve the same performance with ML. For general tuning curves the difference between the two performances is smaller the smoother are the tails of the curve. Exact equality of ML and population vector estimation performance holds only for tuning curves of the form $\log f(\theta) = A + B \cos\theta$, assuming a Poisson population.

A varying degree of perceptual learning is required by different discrimination mechanisms. If the nervous system in fact managed to implement ML then its performance would be uniformly good for all stimuli. A more complex network than considered here—e.g., a multilayer perceptron—might also attain uniformly good performance (18, 19). In contrast, due to their limited representational power, simple threshold linear circuits can achieve uniformly good performance for all stimuli only after adaptation to each one. If such a network is adapted to one stimulus, it will perform worse with a novel one, until further adaptation is allowed to take place. The extent of transfer does not decrease monotonically with the separation between the novel and adapted stimuli. Instead, after dropping to zero for separations of roughly the tuning width, it increases again. This nonmonotonic transfer is especially pronounced for the population vector, where it reaches a maximum of 100% when the separation is half the period of the underlying tuning curves. There is a trade-off between performance at the adapted stimulus and the degree of transfer. The fully adaptive perceptron performs optimally at the adapted stimulus but its range of transfer is much narrower than that of the vector discriminator, which is suboptimal at the adapted stimulus.

Our analysis has focused on the limits on performance imposed by noise in the input neurons. The analysis could be extended to include noise in the neurons of the readout network itself. Constraints on the learning mechanisms and on representation could also be limiting factors. The issue of representation could be addressed by expanding our scope from linear networks to those with hidden layer nonlinearities. Furthermore, our analysis ignores correlations in the fluctuations in neuronal response (2). As will be discussed elsewhere (unpublished results), the presence of correlations may not change our results drastically.

Simple learning mechanisms, such as Hebb-like rules, can lead to the adaptation required by our networks. The time scale of adaptation depends on the architecture. The vector discriminator is expected to learn fast because it has only two adjustable parameters, in contrast to the fully adapted perceptron, which needs at least order N trials to learn its N weights (20). Our predictions concerning transfer and time scale of perceptual learning can be tested by psychophysical experiments on direction or orientation discrimination. Interestingly, nonmonotonic transfer of learning in direction discrimination tasks has been observed, although the statistical significance of this experimental finding has been questioned (21). Additional, careful experiments could provide an empirical test of the validity of linear and threshold linear models of readout of distributed neural codes.

We thank L. Abbott, M. Ahissar, A. Borst, S. Hochstein, D. Kleinfeld, G. Mato, W. Newsome, N. Rubin, M. Shadlen, and U. Zohary for helpful discussions and the USA–Israel Binational Science Foundation for their support.

1. Vogels, R., Spileers, W. & Orban, G. A. (1989) *Exp. Brain Res.* **77,** 432–436.
2. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, A. (1992) *J. Neurosci.* **12,** 4745–4765.
3. Atick, J. J. (1992) *Network* **3,** 213–251.
4. Bialek, W. (1990) in *1989 Lectures in Complex Systems, SFI Studies in the Sciences of Complexity, Vol. 2*, ed. Jen, E. (Addison-Wesley, Redwood City, CA), pp. 513–595.
5. Paradiso, M. A. (1988) *Biol. Cybern.* **58,** 35–49.
6. Snippe, H. P. & Koenderink, J. J. (1992) *Biol. Cybern.* **66,** 543–551.
7. Lee, C., Rohrer, W. H. & Sparks, D. L. (1988) *Nature (London)* **332,** 357–360.
8. Georgopoulos, A. P., Schwartz, A. B. & Kettner, R. E. (1986) *Science* **233,** 1416–1419.
9. Vogels, R. (1990) *Biol. Cybern.* **64,** 25–31.
10. Gilbert, C. D. & Wiesel, T. N. (1990) *Vision Res.* **30,** 1689–1701.
11. Zohary, E. (1992) *Biol. Cybern.* **66,** 262–272.
12. Bennett, R. G. & Westheimer, G. (1991) *Percept. Psychophys.* **49,** 541–546.
13. Karni, A. & Sagi, D. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 4966–4970.
14. Poggio, T., Fahle, M. & Edelman, S. (1992) *Science* **256,** 1018–1021.
15. Cover, T. M. & Thomas, J. A. (1991) *Elements of Information Theory* (Wiley, New York).
16. Orban, G. A. (1984) *Neuronal Operations in the Visual Cortex* (Springer, Berlin).
17. Hammond, P. & Pomfrett, C. J. D. (1989) *Vision Res.* **29,** 653–662.
18. Orban, G. A., Devos, M. & Vogels, R. (1990) in *Neurocomputing*, eds. Soulié, F. F. & Hérault, J. (Springer, Berlin).
19. Devos, M. & Orban, G. A. (1990) *Neural Comput.* **2,** 152–161.
20. Barkai, N., Seung, H. S. & Sompolinsky, H. (1993) *Phys. Rev. Lett.* **70,** 3167–3170.
21. Ball, K. & Sekuler, R. (1987) *Vision Res.* **27,** 953–965.