Coding Strategies in Monkey V1 and Inferior Temporal Cortices

ETHAN D. GERSHON, MATTHEW C. WIENER, PETER E. LATHAM, AND BARRY J. RICHMOND
Laboratory of Neuropsychology, National Institute of Mental Health, and Laboratory of Developmental Neurobiology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland 20892

Gershon, Ethan D., Matthew C. Wiener, Peter E. Latham, and Barry J. Richmond. Coding strategies in monkey V1 and inferior temporal cortices. *J. Neurophysiol.* 79: 1135–1144, 1998. We would like to know whether the statistics of neuronal responses vary across cortical areas. We examined stimulus-elicited spike count response distributions in V1 and inferior temporal (IT) cortices of awake monkeys. In both areas, the distribution of spike counts for each stimulus was well described by a Gaussian distribution, with the log of the variance in the spike count linearly related to the log of the mean spike count. Two significant differences in response characteristics were found: both the range of spike counts and the slope of the log(variance) versus log(mean) regression were larger in V1 than in IT. However, neurons in the two areas transmitted approximately the same amount of information about the stimuli and had about the same channel capacity (the maximum possible transmitted information given noise in the responses). These results suggest that neurons in V1 use more variable signals over a larger dynamic range than IT neurons, which use less variable signals over a smaller dynamic range. The two coding strategies are approximately as effective in transmitting information.

INTRODUCTION

Neurons in different regions of the visual system encode different aspects of visual stimuli. For example, neurons in V1 cortex respond strongly to an oriented bar, whereas those in inferior temporal (IT) cortex often require a more complex stimulus. We would like to know whether the differences in what is encoded are reflected in differences in the neuronal response as this may shed light on strategies for cortical processing. Specifically, we ask two questions. First, in what way does the statistical structure of responses differ across areas? Second, how do the differences, if any, affect information transmission?

The first question can be answered by looking at the statistics of neuronal responses. Depending on the nature of those responses, the relevant statistics may be as simple as mean firing rate or as complicated as high order correlations in spike arrival times. The second question requires a precise definition of information, which is provided by information theory. Information theory tells us that the ability of a neuron to distinguish among members of a set of stimuli depends on two things. The first is the number of distinguishable responses: the more distinguishable responses, the larger the information. The second is the variability in the response to each stimulus. Variability (which is reflected in the probabilistic transformation from stimulus to response) clearly degrades information transmission. If each stimulus produces a very broad range of responses, the information in each response may be very small no matter how many distinguishable responses there are.

One difficulty in applying information theory to neuronal systems is determining what to call a response, i.e., what coding scheme to use. Two extremes are the number of spikes in a fairly wide time window and the spike arrival times measured with high resolution. In general, one looks for the simplest coding scheme that conveys the most information—conflicting constraints the relative importance of which must be decided on a case-by-case basis. Here we have the additional problem that we want to compare brain regions that may use different coding schemes. Fortunately, for V1 and IT, the areas we consider here, it has been shown that spike count in a window ~300 ms wide carries most of the stimulus-related information—about 80% (Heller et al. 1995). The remaining 20% of the information is carried in spike timing with an accuracy of ~30 ms in V1 and 60 ms in IT (Heller et al. 1995). Because neurons in V1 fire at about twice the rate of those in IT (see RESULTS), the spike timing accuracy relative to the mean interspike interval is about the same in the two areas. Thus in this paper, we use the spike count as our neural code. This assumption greatly simplifies our calculations, although in principle everything we do here could be applied to coding schemes that include temporal variations.

Because we are using a spike count code, the number of distinguishable responses is the range of spike counts a neuron is capable of producing in response to all stimuli; we refer to this as the dynamic range. Any single stimulus also elicits a range of spike counts. This variability in the response to a single stimulus is captured in the conditional probability of observing n spikes given stimulus s, $P(n|s)$. Therefore, we can compare responses in the two regions by examining only the dynamic range and the conditional probabilities. In practice, this means we need a stimulus set only large enough to provide accurate estimates of these two quantities. This is a weaker constraint on the stimulus set than is required by other information theoretic analyses, which depend on the frequencies with which stimuli are presented (Cover and Thomas 1991).

Determining the dynamic range of a neuron is relatively straightforward; the difficult part of the analysis is determining the conditional probability distribution, $P(n|s)$. Here we follow previous work, in which it has been shown that the logarithm of the variance of the stimulus-elicited spike count is related linearly to the logarithm of the mean spike count in both cat and monkey V1 cortex (Dean 1981; Tolhurst et al. 1981, 1983; van Kan et al. 1985; Vogels et al. 1989). We confirm the mean-variance relation in monkey V1, and we observe a similar relation in monkey IT cortex. We then go on to show that $P(n|s)$ is well approximated by a modified Gaussian distribution (the main modification

was truncation at 0; see METHODS for details) with mean, μ, that depends on the stimulus, *s*, and variance that depends only on the mean.

What the mean-variance relation gives us is the conditional probability, $P(n|\mu)$, of observing *n* spikes given a mean spike count μ. While $P(n|\mu)$ is not quite the same as the probability of observing *n* spikes given the stimulus, it is in some ways more valuable. For example, given the mean-variance relation, and thus $P(n|\mu)$, in V1 and IT, it is relatively easy to compare the two areas: simply examine which area has a larger variance for a given mean spike count, which has a larger dynamic range, and investigate how the differences affect information transmission. When we carry out this exercise, we find that both the spike count variance and the dynamic range are significantly larger in V1 than in IT. For information transmission, these two trends work in opposite directions: a large dynamic range increases information transmission (more distinguishable responses), whereas a large variance reduces information transmission (more variability). For V1 and IT, the two effects approximately cancel: the maximum amount of information that could be transmitted is about the same in the two areas—assuming a spike count code is used, and the observed dynamic range and mean-variance relations apply. Thus although neurons in V1 and IT implement different coding strategies, as reflected in the significantly larger variability and range of responses in V1 than in IT, neurons in the two areas are capable of transmitting about the same amount of information using a spike count code.

An abstract of these results has appeared (Gershon et al. 1996).

METHODS

Data set

We performed new analyses using previously published data. The data came from two studies of supragranular V1 complex cells, each study using two rhesus monkeys performing a simple fixation task (Kjaer et al. 1997; Richmond et al. 1990), and from one study of neurons in area TE of IT cortex in two other monkeys performing a simple sequential nonmatch-to-sample task (Eskandar et al. 1992). The stimuli were centered on V1 neuronal receptive fields, which were located in the lower contralateral visual field 1–3° from the fovea. The IT visual receptive fields were large and bilateral and included the fovea. Standard extracellular recording methods were used throughout.

In the three experimental studies considered here, the visual stimuli were two-dimensional black-and-white patterns based on the Walsh functions (Fig. 1). For V1 *set 1* (Richmond et al. 1990), 128 stimuli were used: a set of 64 8 × 8 pixel patterns and their contrast-reversed counterparts. For V1 *set 2* (Kjaer et al. 1997), 16 stimuli were used: a set of 8 16 × 16 pixel patterns and their contrast-reversed counterparts. In both sets, the patterns covered the excitatory receptive field. At 3° eccentricity, the stimuli were ~2.5° on a side. For the IT experiments, 32 stimuli were used: 16 4 × 4 pixel patterns and their contrast-reversed counterparts. These patterns were 4° square and centered on the fixation point.

The stimulus was on for 320 ms in V1 and 352 ms in IT. To account for latencies and to avoid contamination from off-responses, spikes were counted during the interval from 30 to 300 ms after stimulus onset for the V1 neurons and 50 to 350 ms after stimulus onset for the IT neurons. For every neuron, each stimulus was presented approximately the same number of times (±2) in randomized order. Different neurons received different numbers of presentations. The number of stimulus presentations was between 3 and 34 in V1 *set 1*, between 18 and 231 in V1 *set 2*, and between 21 and 51 in IT. The timing of events, including spikes, was recorded with 1-ms resolution.

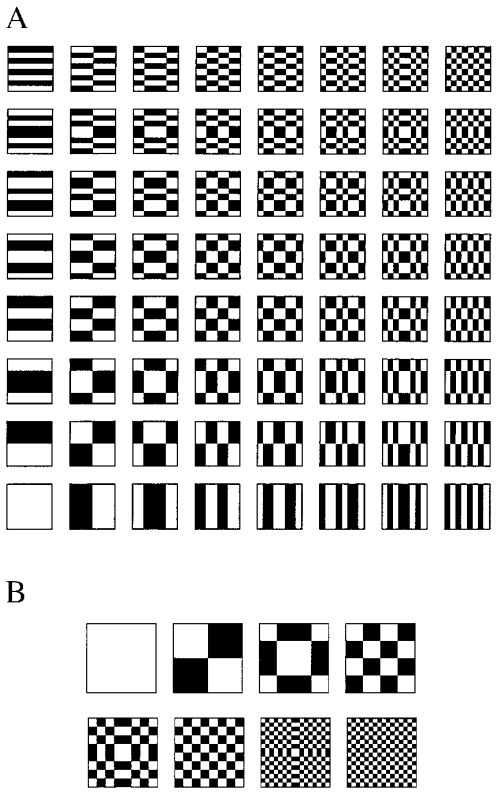

FIG. 1. Walsh patterns. For V1 *set 1*, the 64 stimuli (*A*) and the corresponding contrast-reversed set were presented on the receptive fields while the monkey fixated. Stimuli were 2.5° on a side (covering the excitatory receptive field and some of the surround). For V1 *set 2*, the 8 stimuli (*B*) and the corresponding contrast-reversed set were presented on the receptive field while the monkey fixated. For inferior temporal (IT), the 4 × 4 set (16 stimuli) in the *lower left corner* of *A* and the corresponding contrast-reversed set were used as the monkey performed a nonmatch-to-sample task. The stimuli were 4° on a side and were centered at the point of fixation.

Relationship between mean spike count and its variance

For each cell, each stimulus produces a sample mean spike count, μ_i, and a sample variance in spike count, σ_i^2, where the subscript *i* labels stimulus. We use linear regression to fit the curve $\log \sigma^2 = b + m \log \mu$ to the set of points (μ_i, σ_i^2). This results in a slope, *m*, and intercept, *b*, for each cell.

Estimates of $\log(\mu)$ and $\log(\sigma^2)$ obtained by taking the logarithm of the sample mean and variance are biased and result in underestimation of the variance of response distributions and overestimation of transmitted information. We corrected for the bias using a Taylor series expansion; only a few terms are needed for good results. See Kendall and Stuart (1961), p. 4–6.

Fitting analytic distributions to the data

We seek a model for the conditional probability, $P(n|s)$, of observing *n* spikes in response to stimulus *s*. We examined two widely used probability distributions—the Poisson distribution and a modified Gaussian distribution. The Gaussian distribution was modified by truncation to eliminate the negative portion followed

by normalization. Such distributions have been considered for neural data before (Foldiak 1993). The probability of seeing n spikes was taken to be the integral of this density function between $n - \frac{1}{2}$ and $n + \frac{1}{2}$ (0 and $\frac{1}{2}$ for $n = 0$). A χ^2 test was used to compare each of the analytic distributions to the histogram of experimentally observed spike counts. To have enough data for this analysis, only the responses to stimuli that had been presented ≥ 12 times to a given cell were considered.

As an alternative to integrating the Gaussian between $n - \frac{1}{2}$ and $n + \frac{1}{2}$, we took the probability of observing n spikes, $P(n|s)$, to be proportional to the Gaussian density evaluated at n. The constant of proportionality was chosen to ensure that the total probability summed to one. This alternative method resulted in negligible differences in all quantities we calculated.

Information measures

The information carried in a neuron's response about which member of a set of stimuli is present is defined as (Cover and Thomas 1991)

$$I(S; R) = \sum_{s,r} P(s)P(r|s) \log_2 \frac{P(r|s)}{P(r)} \tag{1}$$

where S is the set of stimuli s, R is the set of responses r, $P(r|s)$ is the conditional probability of response r given stimulus s, $P(s)$ is the probability that stimulus s occurred, and $P(r) = \sum_s P(r|s)P(s)$ is the probability of response r. *Equation 1* is general, but here we confine ourselves to the case where the response r is taken to be the number of spikes elicited by the stimulus. Thus in what follows, we replace $P(r|s)$ with $P(n|s)$ and $P(r)$ with $P(n)$ where n is the number of spikes.

The transmitted information, $I(S; R)$, given in *Eq. 1* is a function of the stimulus probability distribution, $P(s)$. The channel capacity is the maximum value of $I(S; R)$ with respect to the probability distribution $P(s)$. Here we take S to contain all visual stimuli. Clearly channel capacity depends on what we take for the response, i.e., what we choose for the neural code. However, once we choose a code and a stimulus set, the channel capacity is well defined, and it represents a lower bound on the maximum amount of information that could be transmitted. In this analysis, we use a spike count code. Such a code has been shown to carry ~80% of the stimulus-related information (Heller et al. 1995), so we suspect that the lower bound we compute will not be far from the true maximum transmitted information.

Because the channel capacity is independent of the frequencies with which stimuli are presented in any single experiment, it is a robust measure that can be used to compare information transmission rates across brain regions. However, it is more difficult to compute than transmitted information for purely experimental reasons: we can measure the conditional probability distribution, $P(n|s)$, for a relatively small number of stimuli, s, but to accurately estimate the channel capacity, we need to know $P(n|s)$ for all stimuli. We can get around this problem by first constructing $P(n|\mu)$ from $P(n|s)$, where $P(n|\mu)$ is the probability of observing spike count n given the mean spike count μ, and second, developing an analytic model for $P(n|\mu)$.

The expression for transmitted information must be rewritten in terms of $P(n|\mu)$. We start by writing the transmitted information, $I(S; R)$, in the form

$$I(S; R) = \sum_{\mu} \sum_{s \in [\mu, \mu+\Delta\mu], n} P(s)P(n|s) \log_2 \frac{P(n|s)}{P(n)} \tag{2}$$

where the notation $s \in [\mu, \mu + \Delta\mu]$ means restrict s to only those stimuli that produce a response the mean spike count of which lies between μ and $\mu + \Delta\mu$ and the sum over μ runs in increments of $\Delta\mu$. *Equation 2* is exact; all we have done is order stimuli by the mean spike count they produce. The next step is to replace $P(n|s \in [\mu, \mu + \Delta\mu])$ with $P(n|\mu)$. This also would be exact in the limit $\Delta\mu \to 0$ if the distribution of spike counts depended only on the mean. We show in the results that it is a good approximation to assume that the distribution of spike counts does depend only on the mean; in particular, it provides an estimate of the transmitted information that is consistent with estimates reached by other accepted methods. Thus we will adopt that approximation here.

Before replacing $P(n|s \in [\mu, \mu + \Delta\mu])$ with $P(n|\mu)$, we need to express $P(\mu)$ in terms of $P(s)$. This can be done by noting that $P(s)$ induces a probability distribution $P(\mu)$

$$P(\mu)\Delta\mu = \sum_{s \in [\mu, \mu+\Delta\mu]} P(s) \tag{3}$$

Then ignoring the error associated with the approximation $P(n|s \in [\mu, \mu + \Delta\mu]) \approx P(n|\mu)$, we write the probability of observing spike count n, averaged over all mean spike counts, as

$$P(n) = \int \mathrm{d}\mu P(n|\mu)P(\mu) \tag{4}$$

where we replaced the sum over μ that appeared in *Eq. 2* with an integral, valid in the limit of small $\Delta\mu$. Finally, we can rewrite *Eq. 2* for $I(S; R)$ in terms of probability distributions over n and μ

$$I(S; R) = \int \mathrm{d}\mu P(\mu) \sum_n P(n|\mu) \log_2 \frac{P(n|\mu)}{P(n)} \tag{5}$$

with $P(\mu)$ and $P(n)$ given in *Eqs. 3* and *4*, respectively. Again we use an integral over μ rather than a sum.

We show in the results that $P(n|\mu)$ is well approximated by a modified Gaussian distribution the variance of which is a function of mean spike count. Using this modified Gaussian, we can determine channel capacity by finding the distribution of mean spike counts, $P(\mu)$, that maximizes transmitted information, *Eq. 5*. That distribution must be found numerically, and the numerical implementation requires that we discretize the continuous space of mean responses. We denote these discretized probabilities by $\bar{P}(\mu) = \int_{\mu}^{\mu+\Delta\mu} \mathrm{d}\mu P(\mu)$.

The search for the maximizing set of probabilities is subject to three constraints: the probabilities must be nonnegative, the probabilities must sum to one, and the range of means must be finite. The first two constraints arise from intrinsic properties of probability distributions. If the third constraint is violated, the transmitted information can be infinite and the problem of maximizing transmitted information is ill-posed.

The first constraint is implemented by restricting the search space such that $0 \leq \bar{P}(\mu)$ for all μ. The second constraint is implemented by requiring that

$$\sum_{\mu} \bar{P}(\mu) = 1 \tag{6}$$

The third constraint is implemented by requiring that the distribution of spike counts be consistent with the observed data; that is, the distribution of means must not lead to a distribution of spike counts with many counts outside the observed range. Specifically, if $n_{\min}$ and $n_{\max}$ are the minimum and maximum observed spike counts over all stimuli for a particular cell, then we demand that

$$\sum_{n > n_{\max}} C_+(n - n_{\max})P(n) + \sum_{n < n_{\min}} C_-(n_{\min} - n)P(n) = \epsilon \tag{7}$$

where $P(n)$ is defined in *Eq. 4*, both $C_+(n)$ and $C_-(n)$ are nondecreasing functions of n, and ϵ is small. *Equation 7* ensures that $P(n)$ falls off rapidly for spike counts outside the observed range. To implement the optimization procedure, we need to translate this into a constraint on $\bar{P}(\mu)$ because the search for the maximum

value of the transmitted information occurs in $\bar{P}(\mu)$ space. Defining the function

$$C(\mu) \equiv \sum_{n>n_{\max}} C_+(n - n_{\max})P(n|\mu) + \sum_{n<n_{\min}} C_-(n_{\min} - n)P(n|\mu) \quad (8)$$

and combining *Eqs. 4* and *7*, we arrive at

$$\sum_{\mu} C(\mu)\bar{P}(\mu) \leq \epsilon \quad (9)$$

Equation 9 represents our third constraint. In practice, because expanding the range of spike counts increases transmitted information, we do not have to worry about our range being too small, only too large. Therefore, in *Eq. 9*, only the equality constraint is important.

In our numerical calculations we use

$$C_+(n) = C_-(n) = n^2$$

$$\epsilon = 0.1$$

To find the channel capacity, we minimize the function

$$F[\bar{P}(\mu)] = -I(R; S) + h_1(\sum_{\mu} \bar{P}(\mu) - 1)^2 + h_2[\sum_{\mu} C(\mu)\bar{P}(\mu) - \epsilon]^2 \quad (10)$$

where h_1 and h_2 are large constants. ($h_1 = 10^{12}$ and $h_2 = 10^{15}$ in the calculations presented here. Other large values for the constants give similar results.) The second and third terms of this expression are penalty functions that increase the value of $F[\bar{P}(\mu)]$ when the second and third constraints are not met.

Any standard minimization algorithm can be used. We performed the minimization using the Splus (v. 3.4, Mathsoft, Seattle WA) gradient-descent function *nlminb*.

A minimum may be either global or local. However, in our problem, the minimum is global. This is guaranteed because the space we are searching [$\bar{P}(\mu) \geq 0$ combined with two linear constraints] is convex, and transmitted information is a concave function with respect to $\bar{P}(\mu)$ (Cover and Thomas 1991, p. 31). Therefore, we are guaranteed a single global minimum, and the gradient descent method must converge to that minimum.

RESULTS

We performed new analyses using previously published data from 42 V1 complex cells from two separate data sets (13 from V1 *set 1* and 28 from V1 *set 2*) and 19 IT neurons (Eskandar et al. 1992; Kjaer et al. 1997; Richmond et al. 1990).

Log(variance) is linearly related to log(mean)

Various researchers have demonstrated a linear relation between the logarithm of the mean stimulus-elicited spike count and the logarithm of its variance in V1 neurons (Dean 1981; Tolhurst et al. 1981, 1983; van Kan et al. 1985; Vogels et al. 1989). Using linear regression, we find such a relation for both our V1 complex cells and IT neurons (see Fig. 2).

The slopes of all 13 regressions for neurons in V1 *set 1*, in 27 of 28 neurons in V1 *set 2*, and in 17 of 19 of the IT neurons were significant ($P < 0.01$). The minimum, median, and maximum values of r^2 were 0.14, 0.61, and 0.83 in V1 *set 1*, 0.11, 0.86, and 0.97 in V1 *set 2*, and 0.02, 0.59, and 0.82 in IT, respectively. The median slopes from the regressions were 1.43 (range 0.91–2.67, 12/13 slopes >1) and 1.18 (range 0.38–1.74, 18/28 slopes >1) for neurons in V1 *set 1* and V1 *set 2*, respectively, and 0.82 (range 0.41–1.53, 14/19 slopes <1) for IT neurons. The median intercepts from the regressions were 0.26 (range −2.42–1.45, 3/13 constants <0) and 0.60 (range −0.79–2.10, 5/28 <0) in V1 *set 1* and V1 *set 2*, respectively, and 0.31 (range −1.03–1.82, 5/13 <0) in IT.

Data arising from a process having equal mean and variance (for example, a Poisson process), would give rise to a regression intercept and slope statistically indistinguishable from 0 and 1, respectively. The regressions from all 13 cells in V1 *set 1*, 24 of 28 cells in V1 *set 2*, and 14 of 19 IT cells had either an intercept significantly different from 0 ($P <$

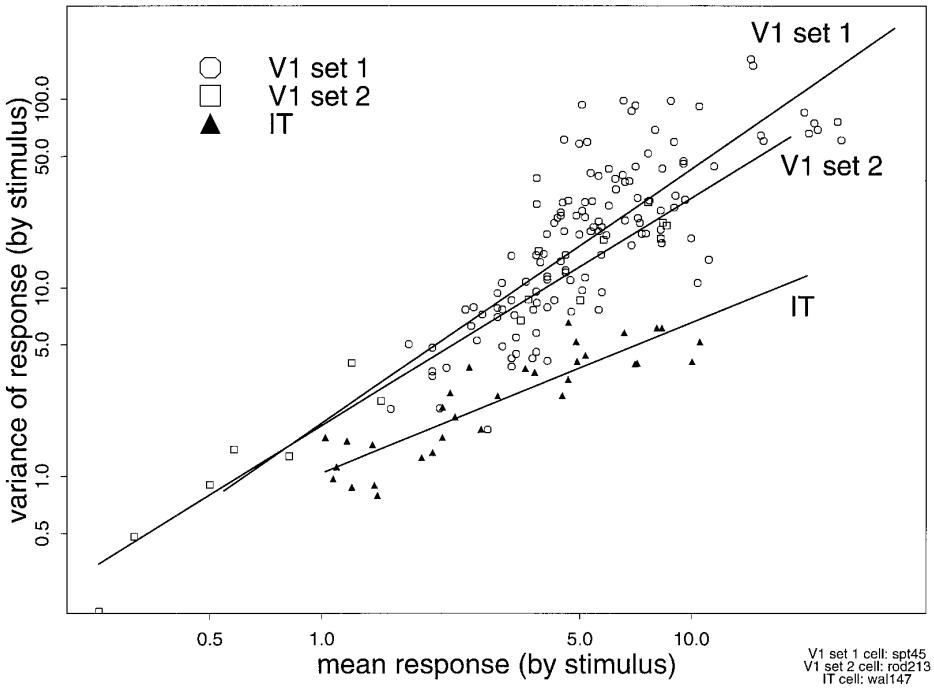

FIG. 2. Log(mean) vs. log(variance) regression. There were 128 stimuli for the V1 *set 1* neuron (○), 16 stimuli for the V1 *set 2* neuron (□), and 32 stimuli for the IT cell (△). Least-squares regression line for each data set is shown. This example shows the cell with the median slope from each data set.

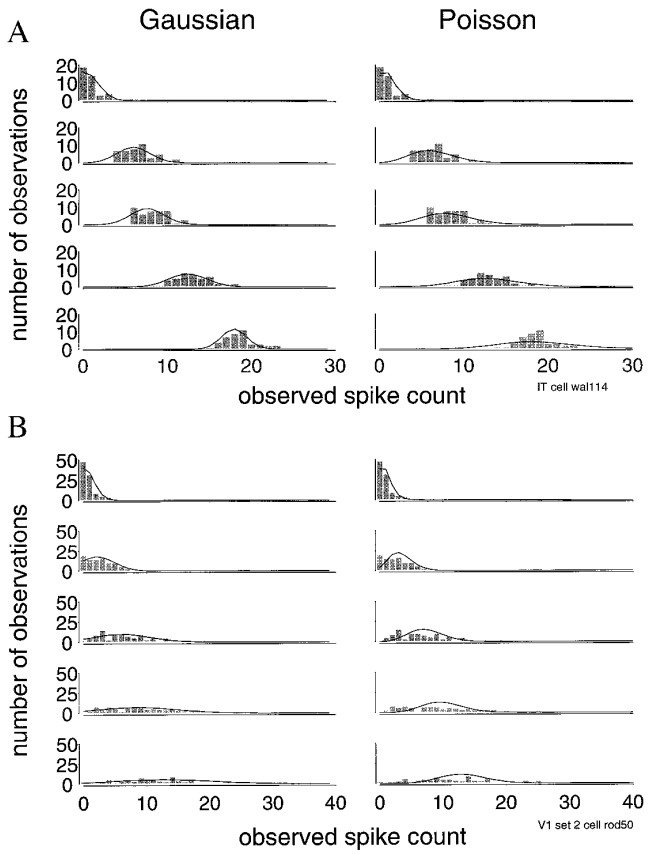

FIG. 3. Sample fits using Poisson and modified Gaussian distributions. *A*: cell from IT. *B*: cell from V1. Each row shows the histogram of responses to 1 of 32 (IT) or 128 (V1) stimuli, along with the best-fit modified Gaussian *(left)* and Poisson *(right)* distributions. Modified Gaussian provides a better fit, especially when the mean firing rate is large. Stimuli presented here were selected to show responses with a range of mean spike counts for each cell. Note that the scales for the 2 sets of graphs are different.

0.05) or a slope significantly different from 1 ($P < 0.05$) or both. These results provide evidence that the Poisson distribution does not provide a good model of the data.

Modified Gaussian fits spike count data better than Poisson

We fit modified Gaussian distributions (as described in METHODS) and Poisson distributions to the empirical distribution of responses elicited by each stimulus. Sample fits are shown in Fig. 3. A χ^2 test was used to evaluate the fits. The requirement that each response distribution analyzed be based on $\geq$12 presentations of the given stimulus excluded 7 of 13 of the neurons from V1 *set 1*. Three of 13 cells had enough presentations per stimulus for all stimuli, and three others had enough presentations for a few stimuli each, for a total of 433 response distributions. All cells from V1 *set 2*, and all cells from the IT set, had enough presentations for all stimuli.

The Poisson distribution requires only a single parameter, the mean spike count, whereas the modified Gaussian requires both the mean and variance of spike count. Parameters were computed in three ways: by choosing the mean (and variance, for the Gaussian) that minimized χ^2, by using the observed mean and (for the Gaussian) the variance predicted by the mean-variance regression, and by using the observed mean and (for the Gaussian) variance. The third method gave such poor results that we dropped it from consideration. The variance of responses to any given stimulus is a sample variance, and therefore is itself a random variable. The regression model uses the variances in response to all stimuli to estimate the variance of response to each stimulus. We believe this explains why the second method is so much more effective than the third method.

Figure 4 shows that the Poisson distribution could be rejected ($P < 0.05$) much more frequently than the modified Gaussian distribution, even when the best-fit parameters were used for the Poisson, and the parameters from the Gaussian were estimated using the data and the mean-variance relation (described in the following text). The difference was even greater when the best-fit parameters were used for the Gaussian as well.

The fact that a χ^2 test based on the observed mean and predicted variance for the Gaussian fails more often than 5% of the time at $P = 0.05$ (6, 25, and 8% in V1 *set 1*, V1 *set 2*, and IT, respectively) suggests that factors other than those identified in this paper may influence the variance of the distributions.

Information estimates using a modified Gaussian distribution

Because a modified Gaussian distribution modeled the data better than a Poisson distribution in all three data sets, we used the modified Gaussian to describe the conditional probabilities $P(n|s)$ needed to compute transmitted information. We chose the mean and variance of the modified Gaussian in three ways: by using the observed mean together with the variance predicted by the mean-variance relation,

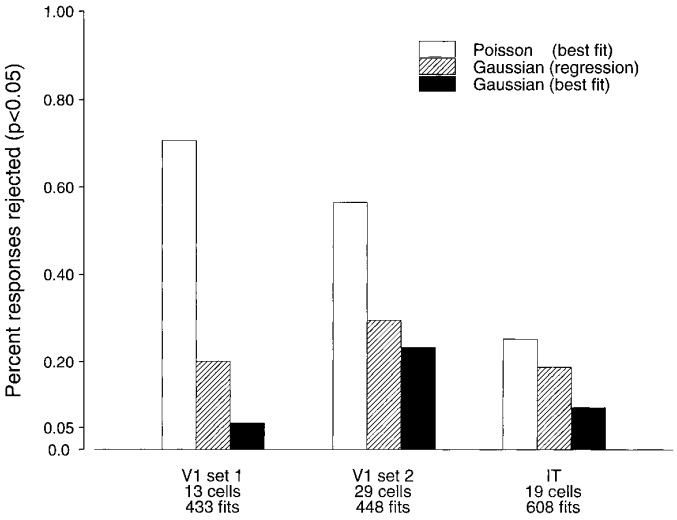

FIG. 4. Chi-squared test of response distributions. Each bar shows the percent of response distributions for which the hypothesis that the data came from the Poisson or modified Gaussian distribution can be rejected ($P = 0.05$). Modified Gaussian distribution using the best-fit parameters is rejected less often than the distribution using the observed mean and variance calculated using the log(variance) vs. log(mean) regression, indicating that other factors probably influence the variance.

by calculating the mean and variance directly from the data, and by using the mean and variance obtained from the fitting procedure. For comparison we also computed the information using an artificial neural network (Golomb et al. 1997; Heller et al. 1995; Kjaer et al. 1994).

The estimates obtained using the first method—the regression method—and the network method are nearly equal (Fig. 5). The second method always calculates higher values for transmitted information than the first method [mean difference = 0.047 ± 0.063 (SD) bits], and the third method calculates even higher values (mean difference = 0.072 ± 0.036 bits). These represented median percent differences of 8 and 20%, respectively.

As a check, we also calculated the transmitted information on the assumption that the responses were distributed according to the Poisson distribution. As expected, given that the Poisson distribution fit the data poorly, the information estimates showed large deviations from the network estimates. The information calculated using the Poisson distribution was higher than the information calculated using the modified Gaussian regression method in 51 of 60 cells (mean difference = 0.23 ± 0.25 bits).

The transmitted information depends on the width of the counting window. We examined windows ranging from 30 to 270 ms in V1 *set 1*, from 30 to 320 ms in V1 *set 2*, and from 50 to 350 ms in IT. The log(mean) versus log(variance) regression was calculated using the spike count distributions in each window. The mean information in the largest time window was 0.33 ± 0.16 bits (n = 13) for V1 *set 1*, 0.40 ± 0.25 bits (n = 28) for V1 *set 2*, and 0.41 ± 0.38 bits (n = 19) for IT. Information rose quickly in the two V1 data sets—most information accumulated in just 50 ms (Fig. 6). Information in IT rose much more slowly, beginning to level off after ~150 ms. The early dip in transmitted information in cells in IT is due to latency effects: some stimuli elicit spikes earlier than others, and in small windows this produces information. Because we are using a spike count code, information is reduced as more of the stimuli elicit spikes. Information rises again as different spike counts become distinguishable. This is evidence that latency carries stimulus-related information in IT neuronal responses. Latency has been shown to carry stimulus-related information in V1 (Gawne et al. 1996).

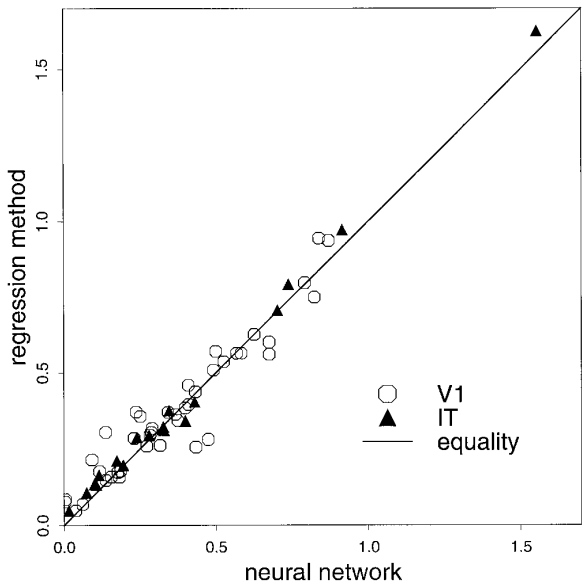

FIG. 5. Two methods for estimating transmitted information. The *x* axis shows the mean value calculated using the neural network (Kjaer et al. 1994); the *y* axis shows the value calculated using the method described in the text. Values calculated using the 2 methods are nearly identical. All cells with enough data to allow analysis (60) are represented.

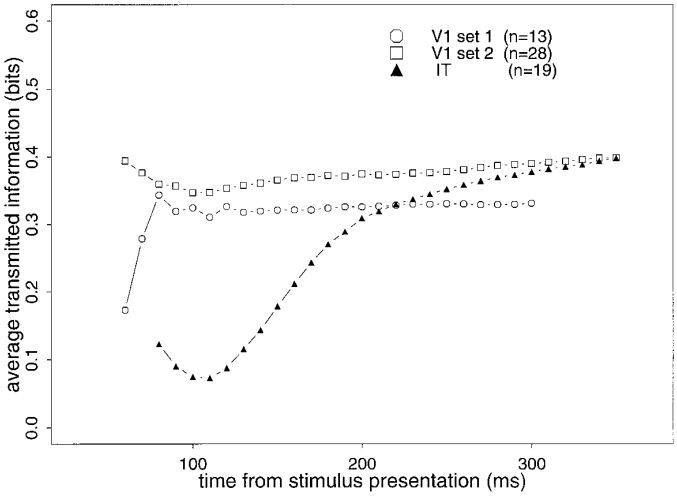

FIG. 6. Transmitted information as a function of the counting window size. The *x* axis shows the time from stimulus presentation. IT starts later than V1 because it has a longer latency. The *y* axis shows the transmitted information accumulated from stimulus presentation (time 0) to the time indicated on the *x* axis. Information accumulates significantly more quickly in neurons from V1 than in neurons from IT.

Channel capacity is approximately the same in V1 and IT

We can compute the channel capacity (assuming a spike count code) by finding the distribution of mean spike counts

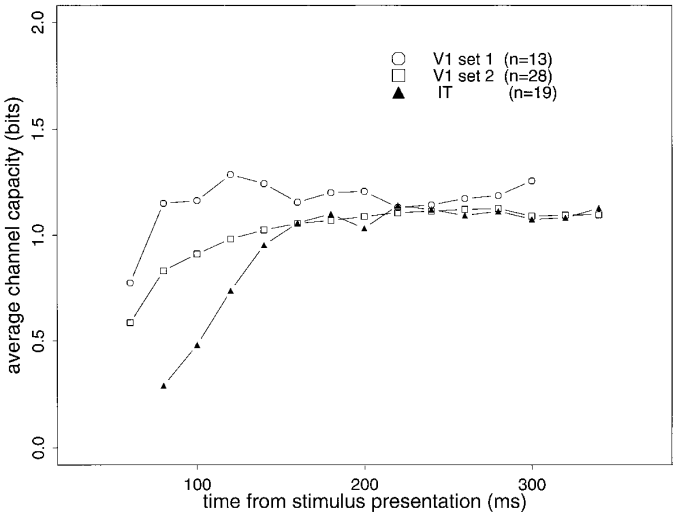

FIG. 7. Channel capacity as a function of the counting window size. The *x* axis shows the time from stimulus presentation. IT starts later than V1 because it has a longer latency. The *y* axis shows the channel capacity accumulated from stimulus presentation *(time 0)* to the time indicated on the *x* axis. Channel capacity rises more quickly in V1 than in IT, although the difference is not as pronounced as for transmitted information.

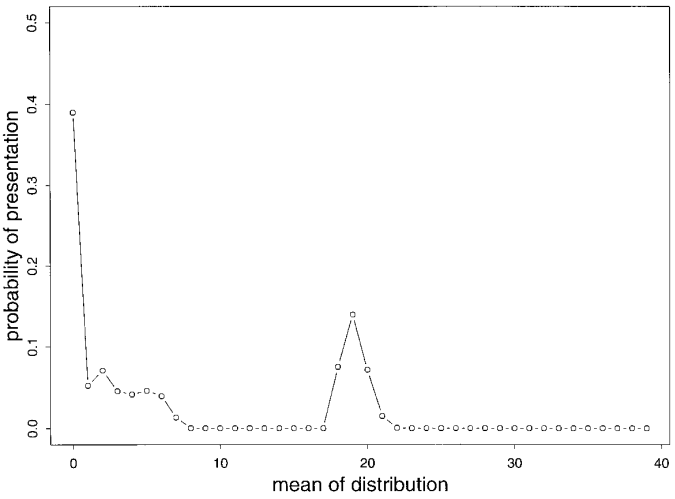

FIG. 8. Distribution of mean responses (each corresponding to a stimulus equivalence class) that maximizes transmitted information. The *x* axis shows the means. The *y* axis shows the probability with which the means should occur to achieve channel capacity. Distribution here was calculated using integer means. As noted in the text, using a finer grid does not materially affect the results.

that yields the highest transmitted information (see METHODS). This requires knowing the minimum and maximum observed spike counts for each cell and the variability in spike count at each mean. The minimum and maximum come directly from the data; for the variability, we assumed that the probability of observing a particular spike count, $P(n|\mu)$, was given by a modified Gaussian distribution with mean μ, and with a variance predicted by a linear relation between log(variance) and log(mean).

In the longest windows available, the minimum spike count for all but six cells was 0; that is, at least one stimulus elicited no spikes. There were two exceptions in each of the V1 data sets (in both, the minimum count was 1 spike) and four exceptions in the IT data set (the minima were 1, 2, 2, and 4). The maximum spikes elicited by any stimulus were fairly evenly spread over a range of 30–75 in V1 *set 1*, 15–45 (with 2 outliers with maxima of 58 and 73) in V1 *set 2*, and 10–30 (with 2 outliers with maxima of 43 and 55) in IT. The median spike count maxima were 54, 31, and 24, respectively.

The average channel capacity was 1.26 ± 0.21 bits in V1 *set 1*, 1.12 ± 0.28 bits in V1 *set 2*, and 1.13 ± 0.47 bits in IT. The median channel capacities were 1.28, 1.02, and 1.23 bits, respectively.

As can be seen in Fig. 7, channel capacity also rose more quickly as a function of time in neurons from the two V1 data sets than in neurons from IT, although the difference is not as pronounced as for transmitted information.

Allowing a larger range of responses increases the amount of information that can be transmitted. Therefore, the channel capacity calculated here depends on the constraints imposed on the dynamic range. To test the robustness of our numerical results, in a few cases we decreased ϵ (which controls how many responses can lie outside the observed dynamic range; see METHODS) by a factor of 10 or used a constant instead of a quadratic weighting function [$C_+(n) = C_-(n) =$ constant; see METHODS]. This did not change the resulting value of the channel capacity by >5% for any of the examples we considered. In addition, numerical implementation of the gradient descent requires that we discretize the probability distribution of the mean spike count. In our simulations, we used a bin size of one spike count so the means took on integer values. Again, to test robustness, in several cases, we decreased the bin size by a factor of 2 and saw little change.

Figure 8 shows a typical example of the probabilities with which various means should occur to achieve channel capacity. In every neuron, mean zero (that is, no spikes) occurs most frequently. A small group of means occurs somewhat less frequently, and the rest of the means occur with extremely low probability. The bumps appear for different non-zero means for different cells; we do not present an average distribution because averaging obscures the fact that each distribution consists of discrete bumps. If the dynamic range of the cell is larger, then additional ''ripples'' may appear, indicating further means that occur with significant probability.

DISCUSSION

In the INTRODUCTION, we posed two questions: in what way does the statistical structure of responses differ across areas and how do the differences, if any, affect information transmission? We found that the responses of neurons in V1 and IT cortex do indeed have different structures. The maximum spike count observed in V1 cortex neurons is generally much higher than that in IT cortex neurons. In addition, responses in V1 are much more variable than those in IT.

Despite the differences in response structure, neurons in V1 and IT cortex carried approximately the same amount of information about the stimulus set used in these experiments. However, transmitted information depends both on the stimuli chosen and on the relative frequencies with which they are presented. With a different stimulus set, we might well have measured a different amount of information. This makes comparison of transmitted information across areas and generalization to other stimuli difficult.

To overcome this limitation, we estimated the channel capacity of the neurons. The channel capacity is the maximum information that a neuron can transmit using a given code, given the constraints of noise in the channel and limited range of responses. The channel capacities were also approximately equal in the two areas. This suggests that, in the course of visual processing, variability is traded off against dynamic range.

The observed differences in spike count and variability are easy both to visualize and to quantify. Transmitted information and channel capacity, on the other hand, are more abstract quantities, and their computation requires a number of assumptions. These assumptions include that spike count is the neural code, that the observed dynamic range is a good estimate of the true dynamic range of the cell, and that the mean-variance relations derived from our data hold for all visual stimuli. We now discuss these and other assumptions and how they influenced our conclusions.

Transmitted information

Computation of the transmitted information between neural responses and a stimulus set requires that we choose a

neural code. Here we chose the number of spikes in a window ≤330 ms wide. In the two areas we examined, V1 and IT, such a spike count code has been shown to carry ~80% of the information contained in the full neuronal response (Heller et al. 1995). Thus the true transmitted information is ~25% higher than the values we report. Because we are comparing areas in which the downward bias caused by using an incomplete code is about the same, this bias should have virtually no effect on our conclusion that the two areas transmit about the same amount of information.

The primary effect of choosing a spike count code was that it allowed us to formulate a simple model for the response distributions. Those distributions turned out to be well approximated by a modified Gaussian (see METHODS for a precise definition of modified). The existence of a model for the response distribution allowed analyses that would have been impossible otherwise. Although in principle it is possible to construct a model for more complicated codes, it is more difficult and larger data sets may be required.

Channel capacity

An intrinsic drawback of transmitted information is that it depends on the frequencies with which stimuli are presented. This makes the value of the transmitted information somewhat arbitrary—it almost always can be made either larger or smaller simply by changing the probabilities with which stimuli are presented. One could imagine adjusting stimulus probabilities to maximize the transmitted information. Shannon and Weaver (1949) defined the resulting maximum value as the channel capacity. It is a function only of the conditional probability distribution $P(r|s)$.

The use of the term channel capacity to represent the maximum amount of information that can be transmitted using a given code (or ''alphabet'', in Shannon's original terminology) in the presence of noise is well established (Cover and Thomas 1991; Shannon and Weaver 1949). Channel capacity, like transmitted information, depends on how we choose to interpret the cell's response, that is, on our assumption about the neural code. However, once we choose a code, the channel capacity is well defined. Because it is always possible that some other code would allow the cell to transmit more information than the code under examination, the channel capacity based on any given code is a lower bound on the amount of information that the cell can transmit. Because the spike-count code has been shown to carry ~80% of the stimulus-related information (Heller et al. 1995), it provides a reasonable first approximation.

The actual code used by neurons is likely to include some temporal aspects of the response. Although temporal modulation could provide many degrees of freedom, studies of V1 neurons have shown that only a few degrees of freedom are used to carry stimulus-related information (Heller et al. 1995; Richmond et al. 1990; Victor and Purpura 1996). We predict that the increase in channel capacity when temporal modulation is taken into account will be proportional to the increase in transmitted information. If this is true, then the actual channel capacities will be ~25% larger than the values we calculated.

Channel capacity should not be confused with information capacity of the signal (MacKay and McCulloch 1952). The information capacity is the information present in the signal itself, subject to a model of the noise.

Calculating channel capacity

To calculate transmitted information, we need an estimate for the response distribution for each stimulus presented in an experiment. These distributions can be estimated directly from the data. Calculating channel capacity is more difficult, because it requires knowing the response distribution for all stimuli, not only those presented in a particular experiment. This problem can be overcome by sorting stimuli into groups based on the response distribution each evokes. Here we will call each such group of stimuli an equivalence class. The neuron cannot distinguish members of an equivalence class from one another. For example, otherwise identical stimuli of different colors produce the same response distribution in a cell insensitive to color. Therefore, rather than considering each stimulus separately, we work with the equivalence classes.

There are likely to be equivalence classes we do not observe experimentally. However, if we can describe the set of equivalence classes with a model involving a small number of parameters, and if we can show (as we have in RESULTS) that the model adequately describes the distributions at parameter values observed experimentally, then we can assume that the model also describes the distributions for unobserved parameter values. This overcomes the major obstacle to calculating channel capacity.

We found that a modified Gaussian distribution provides a good model of the response distributions in our experiments. We compared transmitted information values obtained using the modified Gaussian to the values obtained by a previously validated method using a neural network (Golomb et al. 1997; Heller et al. 1995; Kjaer et al. 1994); the values obtained by the two methods are indistinguishable. Thus although there may be a distribution that fits these data better than the modified Gaussian, the modified Gaussian is a good model for the calculations we want to perform. A Poisson distribution, although often used to model responses, fit our experimental data poorly. Others have reached the same conclusion (Softky and Koch 1993; Victor and Purpura 1996).

The modified Gaussian distribution is fully specified by two parameters: the mean and variance. This distribution provides a sufficiently simple model of the neuronal responses to allow calculation of the channel capacity. It turns out that we can simplify the problem by estimating the variance of each distribution based on the mean. This simplification is achieved using the linear relation between log(mean) and log(variance) (Dean 1981; Tolhurst et al. 1981, 1983; van Kan et al. 1985; Vogels et al. 1989). If we know the mean of a response, we can calculate its variance. Therefore, any response distribution can be characterized by its mean.

With this model, the equivalence classes are labeled by the mean response. Two stimuli that produce the same mean response also produce identical response distributions. These stimuli will be indistinguishable based on the responses of the cell and need not be considered separately. Note that every stimulus produces some mean spike count, even if it is zero, and therefore is accounted for in this model.

We now invoke our main assumption: that mean responses not observed in our experiments could be observed given appropriate stimuli. Then the channel capacity can be calculated by finding the distribution of mean spike counts that maximizes transmitted information.

It is, of course, necessary to restrict the range of spike counts. There are both biophysical and mathematical reasons for this. Biophysically, we know that all cells have a maximum firing rate. Mathematically, if we allow an infinite number of spikes, the channel capacity would be infinite. We restricted the range of spike counts to be consistent with observed data. Briefly, we required that the spike count fall primarily within the experimentally observed minimum and maximum (see METHODS). The maximum experimentally observed spike count may be an underestimate of the true maximum, as we may not have used stimuli that elicited the highest firing rates. However, the peak firing rates that we saw in V1 and IT are similar to those seen by others (Perrett et al. 1984; Rolls 1984; Rolls et al. 1982; Tolhurst et al. 1981, 1983; Vogels et al. 1989). If new evidence does show that the dynamic range is larger than we observed, the channel capacity can be recalculated. The effects, however, are modest. We calculated the increase in channel capacity assuming that the maximum firing rate for each cell was 25% greater than the measured values. The median increase in channel capacity was 8.5% (range 0.8–20.9%).

To ensure that the estimate of channel capacity is reasonable, it is important to know that the log(mean) versus log(variance) regression is reliable. For all but 3 of 60 neurons the regressions were significant. It is possible that the regression could change for different stimuli, or, for example, in different attentional states. A change in the regression would provide powerful evidence for a state change at a fundamental level of neural function. Such changes have been reported for fly H1 cell (de Ruyter van Steveninck et al. 1997). de Ruyter van Steveninck et al. (1997) found less variance, and more information, in the responses of fly H1 cell to a moving coherent stimulus when the stimulus moved along a ''presumably more naturalistic'' two-dimensional trajectory than when the stimulus moved in one direction at constant speed. However, at least one study that looked for such differences in one monkey visual cortical area (MT) failed to find them (McAdams and Maunsell 1996).

When we calculated the channel capacity, we found that, on average, it is about the same in V1 and IT, and in both areas it is about two to four times the transmitted information. This lends support to the notion that the smaller dynamic range found in IT (as compared with V1), which would tend to decrease the information that can be transmitted, is balanced by less variable neuronal responses, which tend to increase information transmission.

Comparison with other studies

In this study, the information transmitted by the spike count averaged ~1 bit/300 ms (3 bits/s). The channel capacity, although typically two to four times larger in any cell, is also not very large. Other investigators have reported significantly higher transmission rates. A recent preliminary report (Buracas et al. 1996) indicates that the transmitted information rates of MT neurons in the monkey reach 30 bits/s with moving stimuli. de Ruyter van Steveninck et al. (1997) estimated that the responses of the H1 neuron of the fly contain 2.43 bits/30 ms (~80 bits/s). Here we examine factors that may account for the differences in our results.

Both MT cortex and the fly H1 neuron analyze motion. In the experiments noted earlier, monkeys or flies were shown stepped motions of coherent patterns of bars. Analysis was carried out to determine the information transmitted by the neurons about the direction of motion of the (unchanging) coherent pattern. Motion analysis (especially in only 1 or 2 dimensions) is an easier problem than pattern recognition. We expect that it requires less computation than pattern recognition, resulting in increased information transmission rate per neuron.

For the neurons in our experiments, almost all of the stimulus-related information that is available in the spike count is available in the first 50 ms of information transmission (after a 30-ms latency period) in V1 cortex and in the first 200 ms of information transmission (after a 50-ms latency period) in IT cortex. If these peak information rates were maintained, the V1 and IT neurons would be able to transmit ~20 and 5 bits/s, respectively. These rates are still smaller than those reported in the motion studies, although 20 bits/s approaches the rates seen by Buracas et al. (1996) in MT.

Given that, in V1, most information that will ever be available is available within 50 ms of the beginning of a response, why flash stimuli on a screen for 300 ms? During normal primate vision, a new image appears on each receptive field one to three times per second due to saccadic eye movement, after which the image is kept nearly still on the retina (compared with saccade velocities). Therefore, to study the processes underlying pattern recognition, flashing stimuli onto the visual field at relatively slow rates seems an appropriate paradigm. It remains to be seen whether more rapid presentation of the images would allow consistent peak information transmission or whether the images would interfere with one another.

In both of the analyses of motion, temporal aspects of neural response were taken into consideration. In our analysis, they were not. Previous analyses (Heller et al. 1995) found that spike count transmitted ~80% of the information available in the full response. Therefore, if we accounted for temporal aspects of the signal, we could expect a 25% rise in transmitted information.

Questions raised

Presumably neurons in these two regions operate according to the same biophysical principles. How is it that the variance is lower in IT neurons than in V1 neurons? Does the larger dynamic range with larger variance offer some advantage that offsets the energy cost of higher firing rates? Finally, why don't all neurons use a large dynamic range with low variability?

The authors thank Drs. Mike W. Oram and Karen D. Pettigrew for helpful discussion and comments on the manuscript.

Present addresses: E. D. Gershon, New York University School of Medicine and Center for Neural Science, New York, NY 10016; P. E. Latham, Dept. of Neurobiology, UCLA, Los Angeles, CA 90095.

Address for reprint requests: B. J. Richmond, Laboratory of Neuropsychology, 49 Convent Dr., Bethesda, MD 20892-4415.

Received 25 September 1997; accepted in final form 1 December 1997.

REFERENCES

BURACAS, G., ZADOR, A., DEWEESE, M., AND ALBRIGHT, T. Measurements of information rates in monkey MT neurons in response to time-varying stimuli. *Soc. Neurosci. Abstr.* 22: 717, 1996.

COVER, T. M. AND THOMAS, J. A. *Elements of Information Theory*. New York: Wiley, 1991.

DE RUYTER VAN STEVENINCK, R. R., LEWEN, G. D., STRONG, S. P., KOBERLE, R., AND BIALEK, W. Reproducibility and variability in neural spike trains. *Science* 275: 1805–1808, 1997.

DEAN, A. F. The variability of discharge of simple cells in the cat striate cortex. *Exp Brain Res* 44: 437–440, 1981.

ESKANDAR, E. N., RICHMOND, B. J., AND OPTICAN, L. M. Role of inferior temporal neurons in visual memory. I. Temporal encoding of information about visual images, recalled images, and behavioral context. *J. Neurophysiol*. 68: 1277–1295, 1992.

FOLDIAK, P. The ''ideal homunculus'': statistical inference from neural population responses. In: *Computation and Neural Systems,* edited by F. H. Eeckman and J. M. Bower. Norwell, MA: Kluwer Academic Publishers, 1993, p. 55–60.

GAWNE, T. J., KJAER, T. W., AND RICHMOND, B. J. Latency: another potential code for feature binding in striate cortex. *J. Neurophysiol.* 76: 1356–1360, 1996.

GERSHON, E. D., LATHAM, P. E., JIN, G.-X., AND RICHMOND, B. J. Stimulus-elicited neuronal responses in striate and inferior temporal cortices are well described by a Gaussian distribution. *Soc. Neurosci. Abstr.* 22: 1612, 1996.

GOLOMB, D., HERTZ, J., PANZERI, S., TREVES, A., AND RICHMOND, B. How well can we estimate the information carried in neuronal responses from limited samples? *Neural Comput.* 9: 649–665, 1997.

HELLER, J., HERTZ, J. A., KJAER, T. W., AND RICHMOND, B. J. Information flow and temporal coding in primate pattern vision. *J. Comput. Neurosci.* 2: 175–193, 1995.

KENDALL, M. G. AND STUART, A. *The Advanced Theory of Statistics: Inference and Relationship.* London: Hafner, 1961, vol. 2.

KJAER, T. W., GAWNE, T. J., HERTZ, J. A., AND RICHMOND, B. J. Insensitivity of V1 complex cell responses to small shifts in the retinal image of complex patterns. *J. Neurophysiol.* 78: 3187–3197, 1997.

KJAER, T. W., HERTZ, J. A., AND RICHMOND, B. J. Decoding cortical neuronal signals: network models, information estimation and spatial tuning. *J. Comput. Neurosci.* 1: 109–139, 1994.

MACKAY, D. M. AND MCCULLOCH, W. S. The limiting information capacity of a neuronal link. *Bull. Math. Biophys.* 14: 127–135, 1952.

MC ADAMS, C. J. AND MAUNSELL, J.H.R. Attention enhances neuronal responses without altering orientation selectivity in macaque area V4. *Soc. Neurosci. Abstr.* 22: 1197, 1996.

PERRETT, D. I., SMITH, P. A. J., POTTER, D. D., MISTLIN, A. J., MILNER, A. D., AND JEEVES, M. A. Neurones responsive to faces in the temporal cortex: studies of the functional organization, sensitivity to identity and relation to perception. *Hum. Neurobiol.* 3: 197–208, 1984.

RICHMOND, B. J., OPTICAN, L. M., AND SPITZER, H. Temporal encoding of two-dimensional patterns by single units in primate visual cortex. I. Stimulus-response relations. *J. Neurophysiol.* 64: 351–369, 1990.

ROLLS, E. T. Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces. *Hum. Neurobiol.* 3: 209–222, 1984.

ROLLS, E. T., PERRETT, D. I., CAAN, A. W., AND WILSON, F.A.W. Neuronal responses related to visual recognition. *Brain* 105: 611–646, 1982.

SHANNON, C. E. AND WEAVER, W. *The Mathematical Theory of Communication.* Urbana, IL: Univ. of Illinois Press, 1949.

SOFTKY, W. R. AND KOCH, C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. *J. Neurosci.* 13: 334–350, 1993.

TOLHURST, D. J., MOVSHON, J. A., AND DEAN, A. F. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vision Res.* 23: 775–785, 1983.

TOLHURST, D. J., MOVSHON, J. A., AND THOMPSON, I. D. The dependence of response amplitude and variance of cat visual cortical neurones on stimulus contrast. *Exp. Brain Res.* 41: 414–419, 1981.

VAN KAN, P. L. E., SCOBEY, R. P., AND GABOR, A. J. Response covariance in cat visual cortex. *Exp. Brain Res.* 60: 559–563, 1985.

VICTOR, J. D. AND PURPURA, K. P. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *J. Neurophysiol.* 76: 1310–1326, 1996.

VOGELS, R., SPILEERS, W., AND ORBAN, G. A. The response variability of striate cortical neurons in the behaving monkey. *Exp. Brain Res.* 77: 432–436, 1989.