# Supplementary Information


| Section | Contents | Page |
|---|---|---|
| I. | Calculating $I$ and $\Delta I$ | 2 |
| II. | Measuring correlation widths | 14 |
| III. | Performing reconstructions with correlated and independent responses | 14 |
| IV. | Rod- and cone-equivalent photon flux | 15 |

## I. Calculating $I$ and $\Delta I$

Our calculation of the mutual information, $I$, and the amount of information lost when cells are treated as independent encoders, $\Delta I$, follows very closely the "direct method" developed by Strong et al.[1] (see also refs. 2-4). The direct method is now relatively standard for computing mutual information[1-4], and can be used with only a minor extension to compute the lost information, $\Delta I$. Thus, our analysis follows published methods in all respects, except one: we provide error bars for the ratio $\Delta I/I$.

### The direct method

While the direct method has been published[1-4], for convenience we outline the main steps here. Neural responses are "constructed" by dividing spike trains into time bins and counting the number of spikes in each bin. The possible spike counts can be thought of as letters, and contiguous sets of letters as words. A response at time $t$ is taken to be the word that starts at time $t$, and the stimulus that drives it is the visual input that occurs before $t$. In our experiments the movie is periodic, so all words that occur a fixed time after movie onset see the same stimulus. We can thus estimate the probability distribution of word pairs given a stimulus, denoted $P_T(r_1, r_2|s)$, by constructing a histogram of words from cells 1 and 2 that occur at time $sT$ after movie onset, where $s$, which labels stimulus, is an integer. Specifically, $P_T(r_1, r_2|s)$ is the number of times words 1 and 2, from cells 1 and 2, occurred at time $sT$ relative to movie onset, divided by the number of movie presentations. The subscript $T$ denotes word length, so the number of letters in a word is $T$ divided by bin size.

Given $P_T(r_1, r_2|s)$, one can calculate both the mutual information and the lost information for words of length $T$, denoted $I_T$ and $\Delta I_T$, respectively. Because of temporal correlations in the spike train, these calculations do not yield true values for either quantity; only in the infinite word limit are the calculations correct. The idea behind the direct method is to use linear regression to extrapolate to infinite word length. Specifically, one fits the curves $I_T/T$ and $\Delta I_T/T$ versus $1/T$ to a line; the intercept at $1/T = 0$ then corresponds to the true values of $I$ and $\Delta I$, in bits/unit time. In what follows we show explicitly how we calculate $I_T$ and $\Delta I_T$ from our data.

The mutual information at word length $T$, $I_T$, is given by the standard formula[5]

$$I_T = -\sum_{r_1,r_2} P_T(r_1,r_2)\log_2 P_T(r_1,r_2) + \sum_s P(s)\sum_{r_1,r_2} P_T(r_1,r_2|s)\log_2 P_T(r_1,r_2|s) \qquad (1)$$

where $P_T(r_1,r_2) \equiv \sum_s P_T(r_1,r_2|s)P(s)$ and $P(s)$ is the probability of observing stimulus $s$. The latter quantity is taken to be uniform; i.e., $P(s)$ equals one over the number of stimuli, independent of $s$. The lost information at word length $T$, $\Delta I_T$, is written (see Eq. (2) of the main text)

$$\Delta I_T = \sum_s P(s)\sum_{r_1,r_2} P_T(r_1,r_2|s)\log_2\left[\frac{P_T(r_1,r_2|s)}{P_{T,IND}(r_1,r_2|s)}\right] - \sum_{r_1,r_2} P_T(r_1,r_2)\log_2\left[\frac{P_T(r_1,r_2)}{P_{T,IND}(r_1,r_2)}\right] \qquad (2)$$

where $P_{T,IND}(r_1,r_2|s) \equiv P_T(r_1|s)P_T(r_2|s)$ and $P_{T,IND}(r_1,r_2) \equiv \sum_s P_{T,IND}(r_1,r_2|s)P(s)$.

Before describing our method for computing $I_T$ and $\Delta I_T$, it is useful to rewrite these two quantities as

$$\begin{aligned} I_T &= H_T(R_1,R_2) - H_T(R_1,R_2|S) \\ \Delta I_T &= H_{T,IND}(R_1,R_2|S) - H_T(R_1,R_2|S) - D(P_T(r_1,r_2)||P_{T,IND}(r_1,r_2)) \end{aligned}$$

where $H_T(R_1,R_2)$ is the total entropy, $H_T(R_1,R_2|S)$ is the conditional entropy with respect to the correlated distribution, $H_{T,IND}(R_1,R_2|S)$ is the conditional entropy with respect to the uncorrelated distribution, and $D(P_T(r_1,r_2)||P_{T,IND}(r_1,r_2))$ is the Kullback-Leibler distance between the correlated and uncorrelated total distributions; these four quantities are given by

$$H_T(R_1, R_2) = -\sum_{r_1,r_2} P_T(r_1, r_2) \log_2 P_T(r_1, r_2) \quad (3a)$$

$$H_T(R_1, R_2|S) = -\sum_{s} P(s) \sum_{r_1,r_2} P_T(r_1, r_2|s) \log_2 P_T(r_1, r_2|s) \quad (3b)$$

$$H_{T,IND}(R_1, R_2|S) = -\sum_{s} P(s) \sum_{r_1,r_2} P_{T,IND}(r_1, r_2|s) \log_2 P_{T,IND}(r_1, r_2|s) \quad (3c)$$

$$D(P_T(r_1, r_2)||P_{T,IND}(r_1, r_2)) = \sum_{r_1,r_2} P_T(r_1, r_2) \log_2 \left[ \frac{P_T(r_1, r_2)}{P_{T,IND}(r_1, r_2)} \right] . \quad (3d)$$

To compute $I_T$ and $\Delta I_T$, we use the *naive* estimator of entropy; that is, we replace $P_T(r_1, r_2|s)$ by $n_T(r_1, r_2|s)/N_s$ in Eqs. (1) and (2), where $n_T(r_1, r_2|s)$ is the observed number of times words 1 and 2 appear in response to stimulus $s$ and $N_s \equiv \sum_{r_1,r_2} n_T(r_1, r_2|s)$ is the number of times the movie is shown, and we replace $P_T(r_1, r_2)$ by $\sum_s n_T(r_1, r_2|s)/N$ where $N \equiv \sum_s N_s$ is the total number of words. For the independent distribution, we do essentially the same thing: we replace $P_{T,IND}(r_1, r_2|s)$ by $n_{T,\text{shifted}}(r_1, r_2|s)/N_s$ in Eq. (3c), where $n_{T,\text{shifted}}(r_1, r_2|s)$ is the observed number of times words 1 and 2 appeared in the *shifted* data. By "shifted", we simply mean shifting the response of neuron 2 by one movie repeat; see *Methods* in the main text. As with $P_T(r_1, r_2)$, we replace $P_{T,IND}(r_1, r_2)$ by $\sum_s n_{T,\text{shifted}}(r_1, r_2|s)/N$.

We use $n_{T,\text{shifted}}(r_1, r_2|s)$ rather than $n_T(r_1|s)n_T(r_2|s)$ because the naive estimate of entropy is biased upward, with a bias that depends on the underlying probability distribution[6]. Since $P_T(r_1, r_2|s)$ and $P_{T,IND}(r_1, r_2|s)$ should be reasonably similar, their conditional entropies, $H_T(R_1, R_2|S)$ and $H_{T,IND}(R_1, R_2|S)$, should have similar bias, and their difference should be nearly bias-free. Thus, the estimate of $\Delta I_T$ should be reasonably free of bias, while the estimate of $I_T$ should be biased slightly upward. Below we verified all these "should"s using surrogate data.

**Estimates of the error in $I_T$ and $\Delta I_T$**

To compute the error in $I_T$ and $\Delta I_T$, we assume that the dominant contribution to the error comes from the conditional entropies, $H_T(R_1, R_2|S)$ and $H_{T,IND}(R_1, R_2|S)$. We make this assumption because the number of samples used to compute the conditional entropy for each stimulus is on the order of 300 (the number of times the movie is shown). In contrast,

the number of samples used to compute the total entropy and the Kullback-Leibler distance is a factor of $T_{\rm movie}/T$ larger, where $T_{\rm movie}$ is the length of the movie. Since we show movies for 7 seconds and use word lengths, $T$, ranging from 1 to 10 ms, the number of samples used to compute the total entropy and the Kullback-Leibler distance ranges from about 210,000 to 2.1 million. Thus, we estimate the variance in $I_T$ and $\Delta I_T$ solely from the variances of $H_T(R_1, R_2|S)$ and $H_{T,IND}(R_1, R_2|S)$.

The variance in the naive estimator of entropy can be estimated by assuming the underlying probability distribution is known, writing down an expression for the variance in terms of that distribution, and then Taylor expanding around the true distribution. Since the resulting expression for the variance is general, we will simplify its derivation by denoting the probability of sampling element $j$ simply as $p_j$, denoting the number of times element $j$ is sampled as $n_j$, and assuming that elements are sampled $N$ times. With this notation, the naive estimate for the entropy is

$$H_{\rm naive} = \sum_j h(n_j/N)$$

where

$$h(p) \equiv -p \log_2 p \,.$$

The variance in $H_{\rm naive}$, denoted $\delta H^2_{\rm naive}$, is given by

$$\delta H^2_{\rm naive} = \sum_{ij} \langle h(n_i/N) h(n_j/N) \rangle - \langle h(n_i/N) \rangle \langle h(n_j/N) \rangle \tag{4}$$

where the average is with respect to the multinomial distribution,

$$\langle f(\mathbf{n}) \rangle = \sum_{\{\mathbf{n}\}} f(\mathbf{n})\, N! \prod_j \frac{p_j^{n_j}}{n_j!} \tag{5}$$

(see ref. 6). Expanding $h(n_j/N)$ in a power series around $p_j$ and keeping only terms up to second order, Eq. (4) becomes

$$\delta H^2_{\rm naive} = \sum_{ij} h'(p_i) h'(p_j) \langle (n_i/N - p_i)(n_j/N - p_j) \rangle$$

where a prime denotes a derivative. Using Eq. (5) to derive the relation $\langle (n_i/N - p_i)(n_j/N - p_j)\rangle = (p_i\delta_{ij} - p_i p_j)/N$, Eq. (6) may be written

$$\delta H^2_{\rm naive} = \frac{1}{N}\left[\sum_i p_i h'(p_i)^2 - \sum_{ij} p_i p_j h'(p_i)h'(p_j)\right] . \tag{6}$$

Finally, performing the derivatives in Eq. (6), we arrive at

$$\delta H^2_{\rm naive} = \frac{1}{N}\left[\sum_i p_i \log_2^2 p_i - \left[\sum_i p_i \log_2 p_i\right]^2\right] . \tag{7}$$

In our numerical calculations we use $n_i/N$ in place of $p_i$ in Eq. (7); that is, we use the approximate relation

$$\delta H^2_{\rm naive} = \frac{1}{N}\left[\sum_i (n_i/N) \log_2^2 (n_i/N) - \left[\sum_i (n_i/N) \log_2 (n_i/N)\right]^2\right] . \tag{8}$$

Equation (8) gives us an estimate for the variance of the total entropy. When estimating the variance of the conditional entropy, we assume that each term in the sum on $s$ (see Eqs. (3b) and (3c)) is independent. In addition, we include sampling error that arises because we only show a finite number of stimuli. The resulting expression for the error in $H_T(R_1, R_2|S)$ is given by

$$\begin{aligned}\delta H^2_T(R_1, R_2|S) = \\ \frac{1}{N_s}\sum_s P(s)\left\{\left[-\sum_{r_1,r_2} P_T(r_1,r_2|s)\log_2 P_T(r_1,r_2|s) - H_T(R_1,R_2|S)\right]^2\right. \\ \left. + \sum_{r_1,r_2} P_T(r_1,r_2|s)\log_2^2 P_T(r_1,r_2|s) - \left[\sum_{r_1,r_2} P_T(r_1,r_2|s)\log_2 P_T(r_1,r_2|s)\right]^2\right\}\end{aligned} \tag{9}$$

where we used $P(s) \equiv 1/N_s$ and, in numerical calculations, $P_T(r_1, r_2|s)$ is replaced by $n_T(r_1, r_2|s)/N_s$. The first term in Eq. (9) corresponds to sampling error; the second follows from Eq. (8). The expression for $\delta H^2_{T,IND}(R_1, R_2|S)$ is identical to the one given in

Eq. (9) except that $P_T(r_1, r_2|s)$ is replaced by $P_{T,IND}(r_1, r_2|s)$; this corresponds to replacing $n_T(r_1, r_2|s)$ by $n_{T,IND}(r_1, r_2|s)$.

As discussed above, we assume that the conditional entropy produces the dominant error in the mutual information, $I_T$. Thus, the variance in $I_T$, denoted $\delta^2 I_T$, is given by

$$\delta^2 I_T = \delta H_T^2(R_1, R_2|S)\,, \tag{10}$$

and the variance in $\Delta I_T$, $\delta^2 \Delta I_T$, is given by

$$\delta^2 \Delta I_T = \delta H_T^2(R_1, R_2|S) + \delta H_{T,IND}^2(R_1, R_2|S)\,. \tag{11}$$

In deriving Eq. (11), we assumed that $H_T(R_1, R_2|S)$ and $H_{T,IND}(R_1, R_2|S)$ are independent; this probably provides a slight overestimate of the variance.

As discussed above, to compute $I$ and $\Delta I$ we plot $I_T/T$ and $\Delta I_T/T$ versus $1/T$ and use linear regression to find the intercepts at $1/T = 0$. The variances in $I_T/T$ and $\Delta I_T/T$ at each value of $1/T$ are computed using Eqs. (10) and (11), and their inverses are used to weight $I_T/T$ and $\Delta I_T/T$ at each value of word length, $T$, when fitting the regression lines. The values of the intercepts, which correspond to information rates, are denoted $I$ and $\Delta I$. While the error in the intercepts could be computed from the variance at each word length, instead we use surrogate data, as described in the next two sections.

**Are we correctly estimating $\Delta I/I$?**

The values of $\Delta I/I$ reported in Fig. 3 of the main text were generated using an *estimator* – an algorithm whose input is experimental data and whose output is $\Delta I/I$. To evaluate the quality of this estimator, we need to answer two questions. First, is it biased? That is, does it consistently under-estimate or over-estimate $\Delta I/I$? Second, what is its variance?

We will attempt to answer these questions using surrogate data. Specifically, we will 1) generate pairs of model spike trains with statistics that resemble as closely as possible the statistics of the spike trains observed in our experiments (see next paragraph), 2) calculate the true value of $\Delta I/I$ for those spike trains, either analytically or numerically (the latter by

generating extremely large data sets), and 3) calculate $\Delta I/I$ using our estimator. We will then compare the estimated value of $\Delta I/I$ to the true value, paying particular attention to the bias and variance.

To make the surrogate data, we divided time into bins of 1 ms and generated spike patterns probabilistically in each bin. We denoted the probability of a particular spike pattern occurring in bin $k$ as $P(\rho_1, \rho_2|k)$, where $\rho_1$ and $\rho_2$, the responses of neurons 1 and 2, respectively, could take on the values 0 (no spike) or 1 (spike). (We use $\rho_1$ and $\rho_2$ to denote 1-bin responses.) Spike trains were generated in 7-s segments (referred to loosely as movies), so $k$ ranged from 1 to 7,000, and the movies were repeated multiple times.

The raw probabilities, denoted $P_{\rm raw}(\rho_1, \rho_2|k)$, were taken directly from the data: a pair of spike trains was binned at 1 ms, and in each bin $P_{\rm raw}(\rho_1, \rho_2|k)$ was computed from the 300 repeats of the stimulus. Temporal correlations were then introduced by generating a correlated noise source,

$$x_k = \lambda x_{k-1} + \sigma(1-\lambda^2)^{1/2}\eta_k$$

where the $\eta_k$ are a set of independent Gaussian random variables with zero mean and unit variance, and $\lambda < 1$. The $x_k$ exhibit correlations that decay exponentially,

$$\langle x_{k+m} x_k \rangle = \sigma^2 e^{-|m|\log(1/\lambda)} \,. \tag{12}$$

We combined the raw probabilities with the noise source to generate $P(\rho_1, \rho_2|k)$,

$$\begin{aligned}
P(1,1|k) &= (1+x_k)P_{\rm raw}(1,1|k) && \text{(13a)}\\
P(1,0|k) &= (1+x_k)P_{\rm raw}(1,0|k) && \text{(13b)}\\
P(0,1|k) &= (1+x_k)P_{\rm raw}(0,1|k) && \text{(13c)}\\
P(0,0|k) &= P_{\rm raw}(0,0|k) - x_k[P_{\rm raw}(1,1|k) + P_{\rm raw}(1,0|k) + P_{\rm raw}(0,1|k)] \,. && \text{(13d)}
\end{aligned}$$

The correlation time, $\tau$, of spike trains derived from $P(\rho_1, \rho_2|k)$ can be computed by combining Eq. (12) with the 1 ms time step. This yields

$$\tau = \frac{\delta t}{\log(1/\lambda)} \tag{14}$$

with $\delta t = 1$ ms.

Using the above method, we generated surrogate data based on the probabilities given in Eq. (13), with underlying raw probabilities taken from the 15 most correlated cell pairs. "Most correlated" was measured by the excess fraction of correlated spikes (ECF). The ECFs ranged from 9.9% to 34%, and the loss in information, $\Delta I/I$, ranged from 1% to 11%. For each cell pair, and thus each underlying probability distribution $P_{\rm raw}(\rho_1, \rho_2|k)$, we generated 10 different spike trains (using 10 different random number generator seeds).

We first generated spike trains with no temporal correlations ($\sigma = 0$). This allowed us to calculate $\Delta I/I$ directly from the probabilities $P_{\rm raw}(\rho_1, \rho_1|k)$, and thus to determine how much data was necessary to obtain good estimates of the true value. The calculation of $\Delta I/I$ from the raw probabilities proceeded as follows. The conditional entropies per unit time – both correlated and independent – were calculated using the relations

$$\begin{aligned} H(R_1, R_2|S) &= -\frac{1}{N\delta t}\sum_{k=1}^{N}\sum_{\rho_1,\rho_2} P_{\rm raw}(\rho_1, \rho_2|k)\log_2 P_{\rm raw}(\rho_1, \rho_2|k) \\ H_{IND}(R_1, R_2|S) &= -\frac{1}{N\delta t}\sum_{k=1}^{N}\left[\sum_{\rho_1} P_{\rm raw}(\rho_1|k)\log_2 P_{\rm raw}(\rho_1|k)\right. \\ &\qquad \left. + \sum_{\rho_2} P_{\rm raw}(\rho_2|k)\log_2 P_{\rm raw}(\rho_2|k)\right] \end{aligned}$$

where $P_{\rm raw}(\rho_1|k) \equiv \sum_{\rho_2} P_{\rm raw}(\rho_1, \rho_2|k)$ and $P_{\rm raw}(\rho_2|k) \equiv \sum_{\rho_1} P_{\rm raw}(\rho_1, \rho_2|k)$. The total entropy and the Kullback-Leibler distance (Eqs. (3a) and (3d)) are more difficult to calculate, since they depend on word length. (The dependence on word length arises because the non-uniform firing rate introduces temporal correlations: a high firing rate at the beginning of a word increases the probability that there will be spikes at the end of the word.) These two quantities thus have to be calculated numerically at each word length, and their true values estimated by extrapolating to infinite word length. The numerical calculation at each

word length, however, can be done exactly. To do this, we first constructed $P_T(r_1, r_2)$, the response probability distribution at word length $T$, by averaging over words,

$$P_T(r_1, r_2) = \sum_s P_T(r_1, r_2|s)P(s)$$

where

$$P_T(r_1, r_2|s) = \prod_k P_{\rm raw}(\rho_1, \rho_2|k)\,. \tag{15}$$

In Eq. (15), there are $T/\delta t$ terms in the product over $k$ (as above, $T$ denotes word length, so $T/\delta t$ is the number of bins in a word). We then inserted $P_T(r_1, r_2)$ into Eq. (3a) to compute the total entropy. An essentially identical calculation yielded $P_{T,IND}(r_1, r_2)$, which was combined with $P_T(r_1, r_2)$ to compute the Kullback-Leibler distance, Eq. (3d).

The number of possible responses as a function of word size is $4^{T/\delta t}$, which increases rapidly with word size. We thus computed $P_T(r_1, r_2)$ only out to 11 bins. This corresponded to about 4 million terms for each stimulus; combined with over 600 stimuli, we were stretching the limits of our computational capabilities. Fortunately, the relative loss in information, $\Delta I/I$, changed very little from 1-bin to 11-bin words: the mean change averaged over the 15 cell pairs was .04% $\pm$ 0.15%. Thus, our calculation of $\Delta I/I$ is relatively accurate.

Given that we know the true value of $\Delta I/I$, we can determine how much data our estimator needs by increasing the number of repeats of the movie until the estimated value of $\Delta I/I$ is close to the true value. Figure 1a shows the error between the estimated value of $\Delta I/I$ and the true value versus the number of movie repeats. The error for $n$ repeats, denoted $\epsilon_n$, is taken to be the standard deviation between the estimated and true value of $\Delta I/I$,

$$\epsilon_n^2 \equiv \left\langle \left[ (\overline{\Delta I/I})_n - (\Delta I/I)_{\rm true} \right]^2 \right\rangle$$

where the bar represents an average over 10 different spike trains drawn from the same underlying distribution, $P_{\rm raw}(\rho_1, \rho_2|k)$, and the angle brackets represent an average over 15 different underlying distributions. As can be seen from Fig. 1a, the error decreases rather

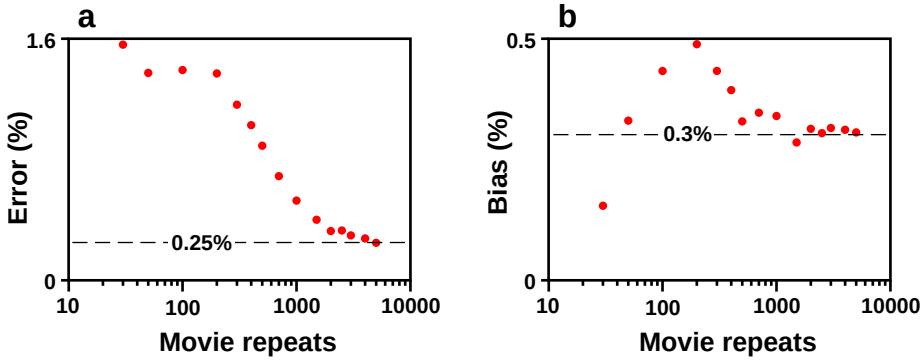


Figure 1: Mean and standard deviation of the relative loss in information. **a**) Error, $\left[\langle[(\overline{\Delta I/I})_n - (\Delta I/I)_{\rm true}]^2\rangle\right]^{1/2}$, versus number of movie repeats, $n$. **b**) Bias, $\langle(\overline{\Delta I/I})_n - (\Delta I/I)_{\rm true}\rangle$, versus number of stimulus repeats, $n$.

slowly. We thus decided that "close" corresponded to an error less than 0.25%, which occurred at 5000 repeats.

Figure 1b shows the bias in our estimator. The bias is slightly positive, even after 5000 repeats. It is small, however, only about 0.3%. We will thus ignore it, and take $(\overline{\Delta I/I})_{5000}$ to be the true value of $\Delta I/I$. This is conservative, since it will *underestimate* the bias in $(\Delta I/I)_{300}$ (300 was the number of repeats used in the experiment).

The histogram of $(\Delta I/I)_{300} - (\overline{\Delta I/I})_{5000}$ is shown in Fig. 2a. Although there is a small upward bias in $(\Delta I/I)_{300}$ compared to $(\overline{\Delta I/I})_{5000}$ of 0.13%, it is not statistically significant, ($p > 0.2$, Student's $t$-test).

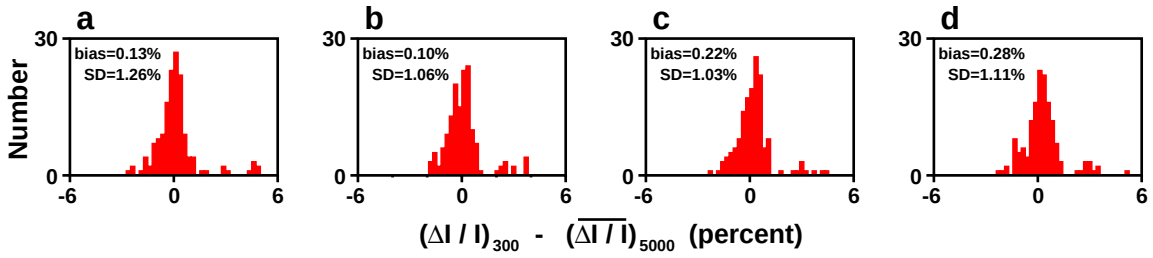


Figure 2: Histograms of $(\Delta I/I)_{300} - (\overline{\Delta I/I})_{5000}$, where $(\Delta I/I)_{300}$ is the relative loss in information computed from 300 repeats of the stimulus and $(\overline{\Delta I/I})_{5000}$ is the mean relative loss computed from 5000 repeats. Each plot contained 150 data points: 15 data sets combined with 10 realizations of spike trains from each set. **a**) $\sigma = 0$ (no temporal correlations in the spike train). **b**) $\sigma = 1$, $\tau = 1$ ms. **c**) $\sigma = 1$, $\tau = 10$ ms. **d**) $\sigma = 1$, $\tau = 100$ ms.

We performed the same calculation for spike trains with temporal correlations. We used $\sigma = 1$, corresponding to full modulation of the underlying probability of observing a spike (see Eqs. (12) and (13)), and we considered correlation times, $\tau$, of 1, 10 and 100 ms (see Eq. (14)). The histograms of $(\Delta I/I)_{300} - (\overline{\Delta I/I})_{5000}$ are shown in Figs. 2b-d for $\tau$ 1, 10 and 100 ms, respectively. There was no statistically significant difference among the four distributions shown in Fig. 2 (Kolmogorov-Smirnov test among the six pairs, $p > 0.05$ for all pairs).

The results in Fig. 2 indicate that, when averaged over all cell pairs, our estimator is biased slightly upward and its standard deviation is small, at worst about 1.3%. The possibility remains, however, that the bias and error might depend on the observed value of $\Delta I/I$ – for example, larger values of $\Delta I/I$ might have smaller, or even negative, biases. To test this, we computed linear regression fits between the bias and $(\overline{\Delta I/I})_{300}$, and between the error and $(\overline{\Delta I/I})_{300}$, where $(\overline{\Delta I/I})_{300}$ is the loss in information for 300 repeats of the movies averaged over the 10 realizations of the spike trains. For the bias (Fig. 3a), neither the intercept nor the slope were statistically significantly different than zero ($p > 0.4$ and $p > 0.7$, respectively). For the error (Fig. 3b), we found that the slope was statistically significant ($p < 10^{-4}$), while the mean was not ($p > 0.2$). To be conservative, though, when computing the error for a particular observation, we included both the slope and intercept, and estimated the standard deviation as

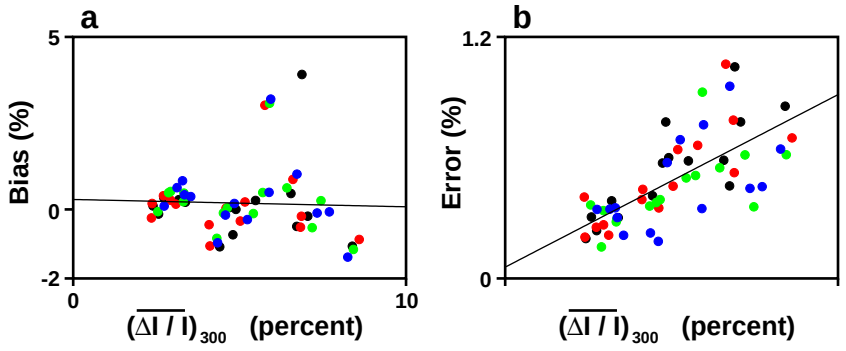


Figure 3: Bias (**a**) and error (**b**) versus $(\overline{\Delta I/I})_{300}$ for the 15 data sets. Black circles: $\sigma = 0$ (no temporal correlations in the spike train). Red: $\sigma = 1$, $\tau = 1$ ms. Green: $\sigma = 1$, $\tau = 10$ ms. Blue: $\sigma = 1$, $\tau = 100$ ms. The lines in each panel are the linear regression fits to the data.

$$\epsilon = a + b(\Delta I/I)_{300} \tag{16}$$

where $a = .079\%$, $b = 0.086$ (the values from the regression line in Fig. 3b), and $\epsilon$ and $(\Delta I/I)_{300}$ are in units of percent.

The absence of a statistically significant bias in our estimator is a key result: barring anomalously large fluctuations, the true maximum loss in information, $\Delta I/I$, for our data set should be close to the observed 11% (Fig. 3 of the main text).

**Computing the error in each measurement**

The raw error in each measurement of $\Delta I/I$ is given in Eq. (16). However, $\Delta I$ is constrained to be positive (it can be expressed as a relative conditional entropy; see *Methods* in the main text). Thus, assuming that $\Delta I/I$ is distributed according to a truncated Gaussian, its mean and variance are given by

$$\text{mean}(\Delta I/I) = \frac{1}{Z}\int_0^\infty dx\, x \exp\left[\frac{-(x-\mu)^2}{2\epsilon^2}\right] \tag{17a}$$

$$\text{var}(\Delta I/I) = \frac{1}{Z}\int_0^\infty dx\, [x - \text{mean}(\Delta I/I)]^2 \exp\left[\frac{-(x-\mu)^2}{2\epsilon^2}\right] \tag{17b}$$

where $\mu \equiv \Delta I/I$ with $\Delta I$ and $I$ computed by extrapolating to infinite word length (see the discussion following Eq. (11)), $\epsilon$ is given in terms of $\Delta I/I$ in Eq. (16), and $Z$ is the normalization; $Z = \int_0^\infty dx \exp[-(x-\mu)^2/2\epsilon^2]$. The values of the mean and variance computed in Eq. (17) were the ones reported in Fig. 3 of the main text.

**II. Measuring correlation widths**

The width of each cross-correlogram was computed by fitting the difference between the raw and shifted cross-correlograms to a Gaussian distribution[7]. Fits were only performed on correlograms in which the two distributions were different ($p < 0.005$, Kolmogorov-Smirnov test, treating the raw and shifted histograms as distributions). In addition, to avoid fitting the Gaussian to noise, only contiguous bins that were statistically significant ($p < 0.05$; binomial statistics) were included in each fit.

**III. Performing reconstructions with correlated and independent responses**

For each pair of cells, we generated two responses, correlated and independent, as follows: We presented the stimulus, which consisted of a movie repeated multiple times. To generate correlated responses, we simply paired the spike trains of the two cells when they saw the movie at the same time. To generate independent responses, we shifted the spike train of one cell relative to the other by one movie length, and then paired the two spike trains; these responses correspond to spike trains that saw the same movie, but at different times.

We then applied a linear reconstruction method[8] to each pair of responses. With this method, an estimate of the stimulus is derived by convolving the spike trains with a linear filter. The filters – one for each spike train – were chosen to minimize the root mean square (RMS) error between the stimulus and the estimate. The filters are, literally, decoders that convert the spike trains into a stimulus estimate.

We generated optimal filters for the two pairs of spike trains, correlated and independent. This gave us two decoders, one for the correlated responses and one for the independent ones. We then applied both decoders to the same pair of spike trains, the correlated spike trains. If correlations are important, then the reconstruction using the independent decoder should be much worse than the reconstruction using the correlated decoder; if correlations are not important, the reconstructions should be about the same. The quality of the reconstructions was measured using RMS error.

The spike trains were binned at 3 ms, and the filters for the reconstruction were 384 ms long. The bin size was chosen to be small enough to capture information about correlations, but large enough to keep the numerical calculation of the filters computationally tractable.

## IV. Rod- and cone-equivalent photon flux

Rod- or cone-equivalent photon flux is the photon flux produced by a light source convolved with the spectral sensitivity function associated with the rod or cone of interest. It is written

$$F_{\text{equiv}} = \int d\lambda P(\lambda)(hc/\lambda)^{-1} S(\lambda_{\max}/\lambda)$$

where $F_{\text{equiv}}$ is the equivalent photon flux in units of equivalent photons/area/time, $P(\lambda)$ is the power spectrum of the light source in units of power/area/wavelength, $S(\lambda_{\max}/\lambda)$ is the spectral sensitivity function, $h$ is Planck's constant, and $c$ is the speed of light. The term $(hc/\lambda)^{-1}$ converts energy to photons.

We evaluated rod- and cone-equivalent photon flux for the two light sources used in the experiments: the monitor that produced the movies and the dim red illuminator on the dissecting microscope. For each light source, we calculated the power incident on the retina. The total power was measured using a broad-band photometer and the spectrum using a spectrometer. The resulting power spectra are shown in Figs. 4a and b (red lines). For the spectral sensitivity function for mouse rods and cones we followed Lamb[9], who showed that $S(\lambda_{\max}/\lambda)$ could be written in the form

$$S(\lambda_{\max}/\lambda) = \frac{1}{\exp a(A - \lambda_{\max}/\lambda) + \exp b(B - \lambda_{\max}/\lambda) + \exp c(C - \lambda_{\max}/\lambda) + D}$$

where $a = 70$, $b = 28.5$, $c = -14.1$, $A = 0.880$, $B = 0.924$, $C = 1.104$, and $D = 0.655$. The spectral sensitivity function peaks at $\lambda = \lambda_{\max}$. For mouse rods, $\lambda_{\max} = 500$ nm, and for the longer wavelength cones (also called the medium wavelength cones or M-cones), $\lambda_{\max} = 511$ nm (Clint Makino, personal communication; see also ref. 10). The spectral sensitivity functions are shown in Fig. 4 (rods, solid black line; cones, dashed black line).

Equivalent photon fluxes were computed by convolving each light source with each spectral sensitivity function. For the movies, we used the temporally averaged photon flux. The results are:

Rod-equivalent photon flux for the movies: 1050 $\mu m^{-2}s^{-1}$.

Cone-equivalent photon flux for the movies: 1200 $\mu m^{-2}s^{-1}$.

Rod-equivalent photon flux for the dissecting microscope: 250 $\mu m^{-2}s^{-1}$.

Cone-equivalent photon flux for the dissecting microscope: 720 $\mu m^{-2}s^{-1}$.

Since the visual stimulation typically lasted ~90 min (two movies at 45 min each) and the dissections took less than 4 min, the total photon dose during dissection was small compared to that during stimulation: $(4/90)\times(250/1050) = 1.1\%$ for the rods and $(4/90)\times(720/1200) = 2.7\%$ for the M-cones.

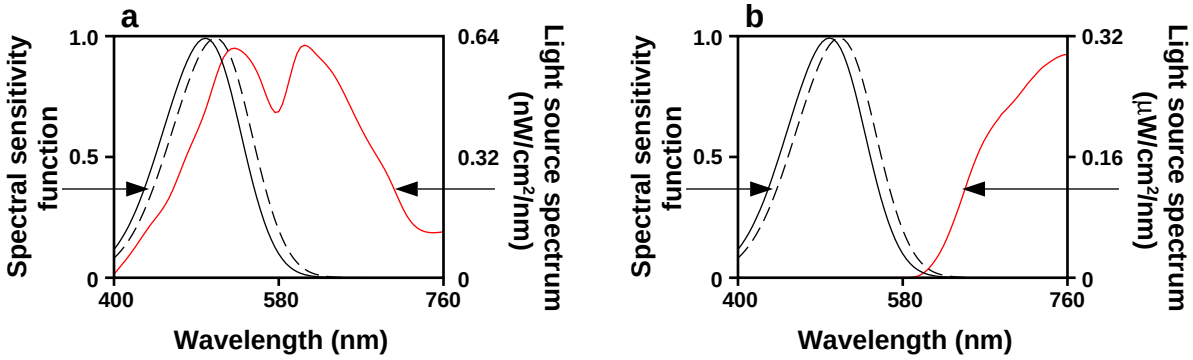


Figure 4: Light source spectra and spectral sensitivity curves. In both (**a**) and (**b**) the solid black line corresponds to the rod spectral sensitivity curve and the dashed black line to the M-cone spectral sensitivity curve. **a**) Temporally averaged spectrum of movies (red line). **b**) Spectrum of red light source on dissecting microscope (red line). The power is larger for the light source on the dissecting scope than for the movies (note units on $y$-axis), but the spectrum is shifted far into the red and well away from the region where the rods and cones are light sensitive.

## References

1. Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R. & Bialek, W. Entropy and information in neural spike trains. *Phys. Rev. Lett.* **80**, 197-200 (1998).

2. Reinagel, P. & Reid, R.C. Temporal coding of visual information in the thalamus. *J. Neurosci.* **20**, 5392-5400 (2000).

3. Buracas, G.T., Zador, A.M., DeWeese, M.R. & Albright, T.D. Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron* **20**, 959-969 (1998).

4. Reich, D.S., Mechler, F., Purpura, K.P. & Victor, J.D. Interspike intervals, receptive fields, and information encoding in primary visual cortex. *J. Neurosci.* **20**, 1964-1974 (2000).

5. Shannon, C.E. & Weaver, W. *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1949).

6. Treves, A. & Panzeri, S. The upward bias in measures of information derived from limited data samples *Neural Comput.* **7**, 399-407 (1995).

7. Brivanlou, I.H., Warland, D.K., & Meister, M. Mechanisms of concerted firing among retinal ganglion cells. *Neuron* **20**, 527-539 (1998).

8. Bialek, W., Rieke, F., de Ruyter van Steveninck, R.R. & Warland, D. Reading a neural code. *Science* **252**, 1854-1857 (1991).

9. Lamb, T.D. Photoreceptor spectral sensitivities: common shape in the long-wavelength region. *Vision Res.* **35**, 3083-3091 (1995).

10. Soucy E., Wang Y., Nirenberg S., Nathans J., Meister M. A novel signaling pathway from rod photoreceptors to ganglion cells in mammalian retina. *Neuron* **21**, 481-93 1998.