Optimal Stimulus Coding by Neural Populations using Rate Codes

Don H. Johnson and Will Ray

Department of Electrical & Computer Engineering, MS 366
Rice University
6100 Main Street
Houston, Texas 77251–1892

Revised October 12, 2003

Abstract

We create a framework based on Fisher information for determining the most effective population coding scheme for representing a continuous-valued stimulus attribute over its entire range. Using this scheme, we derive optimal single- and multi-neuron rate codes for homogeneous populations using several statistical models frequently used to describe neural data. We show that each neuron's discharge rate should increase quadratically with the stimulus and that statistically independent neural outputs provides optimal coding. Only cooperative populations can achieve this condition in an informationally effective way.

Contact: Don H. Johnson
713.348.4956, 713.348.5686 (FAX)
e-mail: dhj@rice.edu web: http://www.ece.rice.edu/~dhj

1   Introduction

Recordings from populations have, in some cases, shown statistical dependence among the constituent neuron's responses (Dan et al., 1998; deCharms and Merzenich, 1996; Gerstein et al., 1989; Perkel and Bullock, 1968; Nirenberg et al., 2001) and little or no dependence in others (Gawne and Richmond, 1993; Johnson and Kiang, 1976; Reich et al., 2001; Zohary et al., 1994). Because a population code presumes that neurons jointly represent information (figure 1), statistical dependence would be expected. Statistical dependence among individual responses is usually categorized into two classes. Stimulus-induced dependence arises because each neuron receives the same input as the others (figure 1), while connection-induced correlation arises when neurons are interconnected somehow. Consequently, studies that find little dependence are surprising. Using a variety of theoretical approaches, many researchers have shown that the effectiveness of neural population coding can be enhanced by statistical dependence among the individual responses (Abbott and Dayan, 1999; Jenison, 2000; Johnson, 2004; Panzeri et al., 1999; Pola et al., 2003; Seung and Sompolinsky, 1993; Shamir and Sompolinsky, 2001; Sompolinksy et al., 2001; Wilke and Eurich, 2002; Wu et al., 2002), while some of these (Abbott and Dayan, 1999; Panzeri et al., 1999; Sompolinksy et al., 2001) and others (Gawne and Richmond, 1993; Zohary et al., 1994) have shown that dependence can hinder population coding. These studies judged the efficacy of population coding using either Fisher information or mutual information, and assumed a specific parametric form for individual and population coding (Seung and Sompolinsky, 1993; Shamir and Sompolinsky, 2001) or examined performance for a specific value of the stimulus (Johnson, 2004; Abbott and Dayan, 1999). One recent study explicitly decomposed the mutual information between stimulus and population response into independent and stimulus-induced and connection-induced dependence terms (Pola et al., 2003).

This report takes a different approach: regardless of how dependence might arise, what is the optimal population code? Here, we develop an approach to deriving the optimal code for populations faced with effectively representing a stimulus attribute over its *entire* range. A single stimulus attribute, denoted here by α, is to be encoded by neural discharge patterns as the attribute varies over the normalized range of $[0, 1]$. We ask, in the case of single neurons, how should discharge rate vary with the stimulus attribute and, in the case of a population, how should individual discharge rates and statistical dependence vary to achieve the most effective neural code. In both cases, we seek the coding function — the way discharge rate or dependence vary with the stimulus — that maximizes the ability of the optimal processing system to extract the value of the stimulus attribute throughout its range. We could not solve this problem without focusing on a particular coding mechanism — the rate code — and on specific statistical models that describe small (three or fewer neurons) populations.

2   Methods

The notation we use here is the same as in Johnson (2004). We use the symbols X and Y to denote a neuron's input and output, respectively (figure 1). The population output is designated by $\mathbf{Y} = \{Y_1, \ldots, Y_N\}$, where N denotes the number of neurons in the population. How stimuli are encoded in the population's output is expressed by the joint probability function $p_{\mathbf{Y}}(\mathbf{y}; \alpha)$, with α denoting the stimulus attribute being coded.

$$p_{\mathbf{Y}}(\mathbf{y}; \alpha) = \int p_{\mathbf{Y}|X}(\mathbf{y}|x) p_X(x; \alpha) \, dx \tag{1}$$

The quantity $p_{\mathbf{Y}|X}(\mathbf{y}|x)$ describes how the population processes its input to produce its output. Ideally, we would seek the optimal form this conditional probability function should take for a specific input probability function. We found this problem too difficult. Our approach is to find the optimal form for the output's joint probability function $p_{\mathbf{Y}}(\mathbf{y}; \alpha)$ when we assume a specific model for the individual neuron's output probability function. For the various models considered here, we found a similarity of results. Based on this similarity, we infer what form the optimal population code might be regardless of model.

Optimal response criterion.   We assume the stimulus attribute is a scalar, and that one population code is better than another if it could lead to a smaller error in estimating the attribute from the population response. Our optimization criterion is to minimize the mean-squared estimation error $\mathcal{E}[(\widehat{\alpha} - \alpha)^2]$, where $\widehat{\alpha}$ denotes the estimate of α. To approach this problem, we use Fisher information. Fisher information is important because of the Cramér-Rao bound, which states that the mean-squared error for any unbiased estimator of a parameter's value cannot be smaller than the reciprocal of the Fisher information evaluated at that value (Johnson et al., 2001): $\mathcal{E}[(\widehat{\alpha} - \alpha)^2] \geq 1/F_{\mathbf{Y}}(\alpha)$, where $F_{\mathbf{Y}}(\alpha) = -\mathcal{E}\left[\frac{d^2}{d\alpha^2} \ln p_{\mathbf{Y}}(\mathbf{y}; \alpha)\right]$. $\mathcal{E}[\cdot]$ denotes expected value with respect to the probability function $p_{\mathbf{Y}}(\mathbf{y}; \alpha)$. By using the Cramér-Rao bound, we need not specify how the attribute is decoded from the population response to determine how well the decoder could work. Information theoretic methods boil down to Fisher information when attribute values must be estimated (Sinanović and Johnson, 2003). Note that, in general, the Fisher information varies with the value of α, which means that the best possible mean-squared error that any estimator can achieve depends on the stimulus attribute's value. By maximizing the Fisher information, we minimize the smallest value the mean-squared error can achieve for a given parameter value. As the amount of data increases, the Cramér-Rao bound can be achieved with equality with a maximum-likelihood estimator (Cramér, 1946). Consequently, by calculating the Fisher information, we can assess how well a stimulus attribute can be measured from a population response. Even though reliance on mean-squared error is limiting, it is well-understood

and gives us more analytic freedom. An important property of Fisher information exploited here is additivity: when $\mathbf{Y}$ consists of statistically independent components, the population's Fisher information equals the sum of the individual neuron's Fisher information. Stated mathematically, when $p_{\mathbf{Y}}(\mathbf{y}; \alpha) = p_{Y_1}(y_1; \alpha) \cdot p_{Y_2}(y_2; \alpha) \cdots p_{Y_N}(y_N; \alpha)$, $F_{\mathbf{Y}}(\alpha) = \sum_n F_{Y_n}(\alpha)$.

We assume that the probability function governing the homogeneous population's collective response can be described by the parameter vector $\boldsymbol{\theta}$. For example, consider a Bernoulli model for the number of spikes occurring in a time bin. This model merely says that no more than one spike can occur in a bin and that the probability of a spike in a bin is p, which equals the product of discharge rate and binwidth.

$$\Pr[m \text{ spikes in a bin}] = \begin{cases} p & m = 1 \\ 1 - p & m = 0 \\ 0 & m > 1 \end{cases} \tag{2}$$

This model is parameterized by the spike probability p; thus, $\boldsymbol{\theta} = \{p\}$. How a probability function's parameters vary with the stimulus define what we call *coding functions* $\boldsymbol{\theta}(\alpha)$: an analytic description of how a one-dimensional stimulus attribute is encoded by a single neuron or by a population as a whole. Using its definition and the chain rule, the Fisher information that results when all of the components of $\boldsymbol{\theta}$ varying in concert with a single stimulus attribute α is given by the quadratic form

$$F_{\mathbf{Y}}(\alpha) = \frac{d\boldsymbol{\theta}^t}{d\alpha} \mathbf{F}_{\mathbf{Y}}\big(\boldsymbol{\theta}(\alpha)\big) \frac{d\boldsymbol{\theta}}{d\alpha} \, , \tag{3}$$

where $\boldsymbol{\theta}^t$ denotes the transpose of $\boldsymbol{\theta}$. The Fisher information $\mathbf{F}_{\mathbf{Y}}$ is a matrix that depends solely on the population's statistical model and how this model depends on its parameters. $\mathbf{F}_{\mathbf{Y}}$ does *not* depend on the stimulus; rather, it depends only on the probabilistic model and how the parameters influence the model. For example, in a two-neuron Bernoulli model, which will be detailed later, this matrix varies with two parameters: spike discharge probability and inter-neuron correlation coefficient. The quantity $F_{\mathbf{Y}}(\alpha)$ indicates how the population's Fisher information varies with the stimulus attribute, which means, in the context of the Bernoulli model, the result of how spike probability and inter-neuron correlation vary together with the stimulus according to some coding function $\boldsymbol{\theta}(\alpha)$. From an information coding viewpoint, it is $F_{\mathbf{Y}}(\alpha)$ that determines how well the stimulus attribute is coded. We seek the "best" coding function by establishing a criterion for how the mean-squared error should vary with the stimulus and solving this differential equation.

Many choices could be made for how the mean-squared error varies with the stimulus. The most general specification is to demand the mean-squared error vary as $K\epsilon^2(\alpha)$, with $\epsilon^2(\alpha)$ a specification of how the optimal estimator's mean-squared error should vary with the attribute's value and with the proportionality constant K chosen as small as possible for optimality. We equate $K\epsilon^2(\alpha)$ with the

reciprocal of the Fisher information. The Fisher information transformation equation becomes

$$\frac{d\boldsymbol{\theta}^t}{d\alpha}\mathbf{F_Y}\left(\boldsymbol{\theta}(\alpha)\right)\frac{d\boldsymbol{\theta}}{d\alpha} = \frac{1}{K\epsilon^2(\alpha)}, \quad \text{with } K \text{ as small as possible.} \tag{4}$$

The population size directly affects the Fisher information because it depends on the population's joint response statistics. We want to apply this criterion for a given population and investigate the optimal coding strategy for it. By merging the square-root of the mean-squared error criterion with the derivative $\frac{d\boldsymbol{\theta}}{d\alpha}$, we obtain a much simpler way of finding the optimal coding function.

$$\left(\sqrt{\epsilon^2(\alpha)}\frac{d\boldsymbol{\theta}}{d\alpha}\right)^t \mathbf{F_Y}\left(\boldsymbol{\theta}(\alpha)\right)\left(\sqrt{\epsilon^2(\alpha)}\frac{d\boldsymbol{\theta}}{d\alpha}\right) = \frac{1}{K}, \quad \text{with } K \text{ as small as possible.}$$

For each value of α, the left side is a quadratic form we seek to maximize. The maximum value of $\mathbf{v}^t\mathbf{F_Y}\mathbf{v}$ occurs when $\mathbf{v}$ is proportional to $\mathbf{v}_{\max}$, the eigenvector of $\mathbf{F_Y}$ having the largest eigenvalue $\mu_{\max}$. Our optimality criterion means that $\sqrt{\epsilon^2(\alpha)}\frac{d\boldsymbol{\theta}}{d\alpha}$ is proportional to the eigenvector $\mathbf{v}_{\max}(\alpha)$ corresponding to largest eigenvalue $\mu_{\max}(\alpha)$ for each value of α. The differential equation that would need to be solved is

$$\frac{d\boldsymbol{\theta}}{d\alpha} = \pm\sqrt{\frac{1}{K\epsilon^2(\alpha)\mu_{\max}(\alpha)}}\mathbf{v}_{\max}(\alpha) \tag{5}$$

Although complicated, this equation can be solved numerically. To be concrete, all eigenvectors are normalized to have unit length; consequently, $\mathbf{v}^t_{\max}\mathbf{v}_{\max} = 1$. We solved this equation for the plus and minus signs separately, and took the one yielding the smallest meaningful value of K. [1]

Perhaps the simplest choice for the mean-squared error criterion $\epsilon^2(\alpha)$ derives from the assumptions that the population strives to represent the attribute uniformly well over the attribute's entire range. With this choice, optimal decoders observing the population's response would yield constant mean-squared error for any attribute value. We thus seek the coding function that yields the largest possible *constant* Fisher information with respect to the stimulus parameter α: $F_\mathbf{Y}(\alpha) = \frac{1}{K} = F_\mathbf{Y}^0$. Note that $F_\mathbf{Y}(\boldsymbol{\theta})$, the Fisher information with respect to the model's parameters $\boldsymbol{\theta}$ is not a constant and will vary according to how population members interact and to what aspect of the response encodes information. We seek to determine how these coding mechanisms should be related to the stimulus so as to achieve constant Fisher information in estimating the stimulus parameter from the population response. Estimators achieving constant mean-squared error are termed equivariance estimators and have the property that they minimize the maximal estimation error (Lehmann and Casella, 1998). Assuming that the stimulus attribute α is defined over the normalized range $[0, 1]$, the differential equation (5) becomes

$$\frac{d\boldsymbol{\theta}}{d\alpha} = \pm\sqrt{\frac{F_\mathbf{Y}^0}{\mu_{\max}(\alpha)}}\mathbf{v}_{\max}(\alpha), \quad 0 \le \alpha \le 1, F_\mathbf{Y}^0 \text{ as large as possible} \tag{6}$$

---

[1] We found that if K is chosen too small, negative probabilities or probabilities greater than one resulted.

Solving this differential equation yields equivariance coding functions for each assumed value of $F_\mathbf{Y}^0$; boundary conditions described below determine $F_\mathbf{Y}^0$.

The standard statistical model for a single neuron's response is the point process (Johnson, 1996). Because correlating point processes in general presents analytic difficulties, we focus on two simpler models frequently used to describe neural data: Bernoulli and Poisson descriptions. Using these essentially means we focus on rate codes. For each model, we use θ to denote all of the parameters governing the joint probability distribution that describes the population's output. Because sequences of Bernoulli random variables are statistically independent from bin to bin, this model does not take into account temporal statistical correlations that may be present in the response. Because of this independence, the Fisher information accumulated over a time interval is simply the sum of each bin's contribution and we only need to consider a single bin. We also derived results for the Poisson and deadtime-modified Poisson counting models, the latter of which incorporates a simple description of refractory phenomena.

3   Results

Optimal single-neuron rate codes.   We first derive the optimal coding function for the single-neuron Bernoulli model (equation (2)). With the probability function's parameter vector consisting only of the discharge probability p, the differential equation (6) becomes

$$\frac{dp}{d\alpha} = \sqrt{F_Y^0 p(1-p)}$$

because in a one-dimensional problem the largest eigenvalue corresponds to the Fisher information of the neuron's statistical model, which in this case equals $1/p(1-p)$. This differential equation's solution describes how the discharge probability should vary with the stimulus attribute so that the smallest possible constant estimation error would result. In other words, the solution is the optimal way (according to the equivariance criterion) to encode the stimulus in a single neuron's discharge probability. With the initial condition that $p(0) = 0$, this equation has the closed form solution

$$p(\alpha) = \frac{1}{2}\left(1 - \cos\sqrt{F_Y^0}\,\alpha\right) \ .$$

Solving for F_Y^0, we find that this quantity depends on the discharge probability at the largest value of the stimulus attribute: $F_Y^0 = \left[\cos^{-1}\left(1 - 2p(1)\right)\right]^2$. The specification of the discharge probability at the extreme values of the stimulus attribute comprise the boundary conditions mentioned after equation (6). Requiring that $p(1) = \frac{1}{2}$, for example, results in a Fisher information value of $F_Y^0 = (\pi/2)^2 = 2.47$, which is equivalent to a root-mean-squared error bound of 0.64/bin. Figure 2 compares the Fisher information for this equivariance coding with several suboptimal (i.e., non-constant Fisher information) choices. By considering the mean-squared error of these alternative rate coding schemes, we found that the equivariance criterion yields a well-behaved variation of the error with the stimulus.

6

For the Poisson case, wherein the discharge rate λ is constant and the number of spikes occurring in an interval of duration T has a Poisson probability function, the Fisher information is $F_Y(\lambda) = T/\lambda$. The optimal coding function found by solving (6) is

$$\lambda(\alpha) = \left[\left(\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}\right)\alpha + \sqrt{\lambda_{\min}}\right]^2 ,$$

where $\lambda_{\max}$, $\lambda_{\min}$ are the maximum and minimum rates, respectively, achieved as α varies. This coding function differs mathematically from the Bernoulli coding function found previously, but both have similar forms (see figure 3). The Fisher information for the optimal coding function is $F_Y^0 = 4T\left(\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}\right)^2$. This result was also derived by Bethge et al. (2002).

We were also able to find the optimal coding function when an absolute refractory interval Δ was included in the Poisson model with rate being the only parameter optimized. The Fisher information is $F_Y(\lambda) = \frac{T}{\lambda(1+\lambda\Delta)}$, and solving (6) for the optimal coding function yields

$$\lambda(\alpha) = \sqrt{\lambda_{\min}^2 + \frac{\lambda_{\min}}{\Delta}}\,\sinh\left(\sqrt{\frac{F_Y^0\Delta}{T}}\alpha\right) + \left(\lambda_{\min} + \frac{1}{2\Delta}\right)\cosh\left(\sqrt{\frac{F_Y^0\Delta}{T}}\alpha\right) - \frac{1}{2\Delta}$$

$$F_Y^0 = \frac{T}{\Delta}\left[\ln\left(\frac{\lambda_{\max} + \frac{1}{2\Delta} + \sqrt{\lambda_{\max}^2 + \frac{\lambda_{\max}}{\Delta}}}{\lambda_{\min} + \frac{1}{2\Delta} + \sqrt{\lambda_{\min}^2 + \frac{\lambda_{\min}}{\Delta}}}\right)\right]^2$$

Despite its complexity, this coding function differed little from that found in the Poisson case when $\Delta\lambda_{\max} \ll 1$. The expression for the equivariance Fisher information is also complicated, but in the small refractory-interval case, $F_Y^0 \approx 4T\left(\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}\right)^2\left[1 - \frac{\Delta}{3}\left(\lambda_{\max} + \sqrt{\lambda_{\max}\lambda_{\min}} + \lambda_{\min}\right)\right]$. Thus, refractoriness reduces the Fisher information.

Optimal population rate codes.   We elaborated these statistical models to describe small populations by correlating the occurrence of events in each model neuron. In addition to finding the optimal value of Fisher information according to the equivariance principle, we need compare that value with that of an independent neural population, which equals the sum of Fisher information values contributed by each neuron. For example, this baseline value for two independent neurons described by the Bernoulli model described above will be $2 \times 2.47 = 4.94$.

Bernoulli models for the two-neuron population are specified by the correlation coefficient ρ and the spike probabilities for each neuron.[2]

$$\Pr[Y_1 = m_1; Y_2 = m_2] = \begin{cases} \rho\sqrt{p_1(1-p_1)p_2(1-p_2)} + p_1 p_2 & m_1 = 1, m_2 = 1 \\ p_1 - \Pr[Y_1 = 1; Y_2 = 1] & m_1 = 1, m_2 = 0 \\ p_2 - \Pr[Y_1 = 1; Y_2 = 1] & m_1 = 0, m_2 = 1 \\ 1 - p_1 - p_2 + \Pr[Y_1 = 1; Y_2 = 1] & m_1 = 0, m_2 = 0 \end{cases} \quad (7)$$

---

[2]Note that some combinations of correlation coefficient ρ and spike probabilities p_1, p_2 are not permitted as they would yield negative probabilities for the population's probability function.

We assume that the spike probabilities and the correlation coefficient ρ, defined in the usual way to be $(\mathcal{E}[Y_1 Y_2] - \mathcal{E}[Y_1]\mathcal{E}[Y_2])/\sigma_{Y_1}\sigma_{Y_2}$, could vary with stimulus attribute. An independent population occurs when $\rho = 0$ (which corresponds to the baseline situation). A noncooperative population, such as shown in figure 1 and expresses stimulus-induced dependence, can only yield positive values for ρ (Johnson, 2004). Negative and positive values of ρ can arise from a cooperative population, which expresses both stimulus- and connection-induced dependence. Consequently, we can only generally determine population structure from the correlation coefficient. If ρ is negative, the structure must be cooperative; if ρ is zero, the optimal structure could be the independent structure or a cooperative one; and if positive, the structure could be cooperative or noncooperative. To make the population homogeneous, we set the spike probabilities equal: $p_1 = p_2 = p$. Thus, the two-neuron Bernoulli model is parameterized as $\boldsymbol{\theta} = \{p, \rho\}$. For three-neuron populations, $\boldsymbol{\theta} = \{p, \rho, \rho_3\}$, with ρ_3 being a correlation-like quantity related to the joint probability of all neurons responding during a bin. No analytic results for the population models could be derived; we relied on numerical solution of the differential equation (6).

Before considering the general case wherein both spike probability and correlation parameters vary with the stimulus, we can examine the special case of pure correlation coding (deCharms and Merzenich, 1996) for a two-neuron population, in which the population encodes α by varying the correlation ρ between neurons while keeping spike probability p fixed. Allowing the correlation to range between 0 and 1 in solving (6) yields a Fisher information of 1.76 with $p = 0.2$ and 2.6 with $p = 0.5$. Thus, the Fisher information resulting from pure correlation coding depends on the sustained discharge probability. Negative correlation coding generally yields smaller Fisher information: 0.18 for $p = 0.2$ and 2.6 for $p = 0.5$. Thus, correlation coding is maximally efficient when high spike probabilities occur, and negative and positive correlations are equally effective at this extreme. Even so, correlation coding yields a Fisher information roughly a factor of two smaller than the baseline value of 4.94.

Figure 3 shows optimal coding functions for the general case when both spike discharge probability and inter-neuron correlation vary with the stimulus for both two- and three-neuron populations described by a Bernoulli model. The most striking aspect of these results is that optimal population encoding required very small correlation values. The greatest value of ρ achieved was 0.016 in the two-neuron case and zero in the three-neuron case. Furthermore, positive correlation resulted in a larger Fisher information (4.97) than negative correlation (4.90) for the two-neuron case. Only positive correlation coding resulted in a Fisher information greater than the baseline, but the Fisher information increase is small (0.04 or 0.8%). Thus, *statistically independent population outputs are essentially optimal.* For the three-neuron model, optimal choices for ρ and ρ_3 were identically zero. Based on these two- and three-neuron Bernoulli models, *we infer that independent populations are optimal according*

to the equivariance criterion. Furthermore, pure correlation coding is inferior, at least in the context of Bernoulli models.

In the Poisson model for correlated activity between two neurons, the output spike count Y_n over some time interval T equals $C_n + C$, where C_n is the Poisson-distributed spike count in the nth neuron and C is a common spike count that each neuron produces (Holgate, 1964). These spike counts are statistically independent of each other and of C. All counts C_n are Poisson random variables having the same rate parameter λ and the common count C is also Poisson with rate λ_0. The common count C among all neurons means the outputs are statistically dependent with positive correlation unless the common-count rate λ_0 equals zero. For the two-neuron case, solving equation (6) resulted in a decreasing rate for the common count component (figure 3). Larger values of the Fisher information $F_{\mathbf{Y}}^0$ were found here (61.8 compared to the baseline value of 48.2). However, if the common rate was constrained to be zero at $\alpha = 0$, which corresponds to a nominally independent population, the common-count rate was zero over the entire range. Again, the independent population is found to be optimal.

Population structures exhibiting optimal rate coding.   These results create a quandary. From the viewpoint of a population endeavoring to code a continuous stimulus attribute as well as possible, the most effective rate coding scheme, as quantified by the smallest Fisher information, occurs when the population outputs are statistically independent. A population having a common input should result in stimulus-induced correlation, which is always positive (Johnson, 2004). An independent population, wherein each neuron in the population receives input statistically independent of the others, results in statistically independent outputs but provides *no* information gain beyond that of a single neuron (Johnson, 2004).

One possible role for cooperation among neurons is to create connection-induced correlation that *decorrelates* the stimulus-induced correlation produced by a noncooperative structure. We envision the structure shown in figure 4, where a lateral connection network interacts the outputs of a noncooperative population to produce a cooperative one. If this kind of cooperation has the right properties, it will be lossless from an information processing viewpoint, and the output would have theoretically ideal properties: statistically independent outputs (thereby producing optimal coding properties from an equivariance viewpoint) while exhibiting the information processing properties of a noncooperative population. We showed in our companion paper (Johnson, 2004) that noncooperative populations always yield maximal processing fidelity as the population size increases.

Complicating the creation of the decorrelation transformation is that it would need to apply over the range of possible joint probability functions $p_{\mathbf{Y}}(\mathbf{y}; \alpha)$ produced by the cooperative population *without*

depending on α. An adaptive cooperative structure could be envisioned wherein the network changes the way the population's outputs interact as stimulus conditions vary. In this paper, we take the analytically simpler tact of exploring a fixed structure that would make the noncooperative outputs as uncorrelated as possible and would exploit population size to create a transformation that would not depend on discharge rates or inter-neuron correlation. We constructed an interconnection network function $\mathbf{Z} = \mathcal{N}[\mathbf{Y}]$ that has the following properties.

- *The transformation is invertible.*

  Mathematically, $\mathbf{Y} = \mathcal{N}^{-1}[\mathbf{Z}]$ must be defined for all possible inputs $\mathbf{Y}$ and outputs $\mathbf{Z}$ of the interconnection network. Invertibility requires that the number of outputs be no smaller than the population's size. The reason for this condition is that invertible transformations are, from an information processing viewpoint, lossless (Sinanović and Johnson, 2003). Hence the cooperative population's information transfer ratio equals that of the noncooperative population and would thus have all of its asymptotic properties.

- *The transformation decorrelates the population's output regardless of ρ.*

  A transformation that yields statistically independent outputs would depend greatly on the joint probability function $p_{\mathbf{Y}}(\mathbf{y}; \alpha)$. However, by decorrelating, which makes the pairwise correlation coefficients zero, we can find a interconnection network without needing to know the details of the output probability function. In addition, we need the transformation to not depend on the stimulus-induced correlation that the noncooperative structure yields.

As shown in the appendix, if the population outputs are scalar quantities (spike counts, for example), a transformation having these properties has the form

$$Z_n \propto Y_n - \left(1 - \frac{1}{\sqrt{1 + \frac{N\rho}{1-\rho}}} \right) \overline{\mathbf{Y}} \tag{8}$$

Here, $\overline{\mathbf{Y}}$ is the average of the population outputs: $\overline{\mathbf{Y}} = \frac{1}{N} \sum_{n=1}^{N} Y_n$. In words, each output Z_n equals an input Y_n minus a constant times the average value of all inputs. Note that as the population size grows, the weighting constant tends to be independent of N and ρ. The threshold population size needed to achieve this insensitivity is $N \gg (1 - \rho)/\rho$. Curiously, insensitivity is easier to achieve when the correlation among population outputs is strong (when $\rho \to 1$); weaker correlation demands larger populations. In this way, uncorrelated population outputs can be obtained so that optimal encoding can be achieved while maintaining the information transfer properties of a noncooperative population.

4   Conclusions

Our theory for finding the optimal coding strategy can be generalized to other statistical models, including ones other than rate codes, and to mean-squared error criteria other than the equivariance principle. To use this approach, the statistical model and the mean-squared error criterion would need to be defined, model parameterized, the Fisher information with respect to that parameter set computed, and equation (5) solved. The relation between the quantity we optimize, Fisher information, and the mean-squared error that must be incurred in estimating the stimulus parameter from the population response is an inequality. Equality in the Cramér-Rao bound occurs in two different situations. On one hand, the estimator could yield an error that equals the bound; such estimators, known as efficient estimators (Lehmann and Casella, 1998), are rare and occur only for specific probabilistic models. Almost certainly, minimum mean-square estimators for population codes are not efficient. On the other hand, Cramér (1946) showed that the observation time increases, the maximum likelihood estimate yields a mean-squared error that approaches the reciprocal Fisher information. In other words, an estimator for *any* problem exists that is asymptotically efficient. Consequently, this paper's goal —determining coding functions that produce a specific mean-squared error variation with the stimulus —is an indirect one because we specify the Fisher information's variation with the stimulus (Bethge et al., 2002). Our results more accurately describe optimal coding functions the longer the observation interval.

Our choice of constant mean-squared error for the performance criterion was not without consideration. Figure 2 shows the optimal mean-square error profiles that various rate functions would yield. For example, a linear increase in discharge rate will result in an error in determining the stimulus attribute that increases as the attribute's value increases. While choosing the equivariance principle is arbitrary, it seems reasonable. We do not know how robust our results are to other choices for the variation of mean-squared error with the stimulus.

We found that the discharge rate should increase as the square of the stimulus attribute and that the population outputs should be statistically independent. The attribute α is intentionally abstract, does not refer to any specific stimulus parameter, such as amplitude. Note that α need not be proportional to stimulus parameters. Instead, α could be proportional to that stimulus aspect —a feature —the sensory system is trying to code efficiently. For example, α could be related to the logarithm of stimulus amplitude. Unfortunately, because these results are tied to specific statistical models, our optimal coding results may not be a general principle, even for all rate codes. That said, the parabolic behavior we found mimics what was found from other coding viewpoints (Miller and Troyer, 2002; Stein, 1967). Although our derivations were limited to three models —Bernoulli, Poisson, and Poisson with deadtime —and to small populations, this parabolic input-output behavior seems to be an efficient way to encode a scalar stimulus feature in a neuron's discharge rate from an equivariance viewpoint (constant mean-squared

error regardless of the stimulus attribute's value). The similarity of the results for the various models is striking, and leads to the speculation that the square-law/independent-output population behavior applies generally for individual and population rate codes.

Our conclusion that independent coding is optimal would seem to mean that no collective population coding should occur because stimulus-induced dependence is not present. We have showed that regardless of the neural code employed, noncooperative populations can represent the stimulus increasingly well as the population size increases, with no information loss occurring asymptotically (Johnson, 2004). Connection-induced dependence resulting from cooperation among the neurons can enhance, diminish, or leave unchanged the efficacy of the population's code relative to the baseline performance of the noncooperative population. We have shown here that cooperation, which causes connection-induced dependence, can cancel stimulus-induced correlation to yield a population output having optimal coding properties from the equivariance viewpoint. One consequence of this structure is that inferring population structure from population response statistics can't be done uniquely: an uncorrelated output results from both the structure shown in figure 4 and the independent population. Also, the cooperative decorrelation structure is relatively insensitive to neuron loss. Two factors make it less robust than the noncooperative structure. First of all, achieving perfect decorrelation without knowing the stimulus-induced correlation requires a sufficiently large population. Losing neurons would eventually comprise the large-population assumption. Secondly, the average of responses also depends on the population size; a robust average would need to be obtained.

Many researchers have recorded population responses, looking for population codes that collectively express the stimulus better than the individual neurons do. Several have concluded that dependence can enhance coding efficacy (Abbott and Dayan, 1999; deCharms and Merzenich, 1996; Jenison, 2000; Wu et al., 2002) while others have concluded it impairs population coding (Gawne and Richmond, 1993). This study finds that independent rate codes are optimal for a population that must represent a stimulus parameter well over its entire range. In the companion paper (Johnson, 2004), we found using the Bernoulli model that if a binary-valued stimulus attribute is to be encoded (for example, coding whether the light is on or off), correlated population responses provide optimal coding. Thus, the findings that dependence can be detrimental or beneficial are both correct: *no single population coding strategy exists that applies universally*. Optimal population coding strategies depend at least on what is to be encoded.

References

Abbott, L. and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation*, 11:91–101.

Bethge, M., Rotermund, D., and Pawelzik, K. (2002). Optimal short-term population coding: When Fisher information fails. *Neural Computation*, 14:2317–2351.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Dan, Y., Alonso, J.-M., Usrey, W., and Reid, R. (1998). Coding of visual information by the precisely correlated spikes in the lateral geniculate nucleus. *Nature Neuroscience*, 1:501–507.

deCharms, R. and Merzenich, M. (1996). Primary cortical representation of sounds by the coordination of action-potential timing. *Nature*, 381:610–613.

Gawne, T. and Richmond, B. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *Neuroscience*, 13:2758–2771.

Gerstein, G., Bedenbaugh, P., and Aertsen, A. (1989). Neuronal assemblies. *IEEE Trans. Biomed. Engineering*, 36:4–14.

Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrika*, 51:241–245.

Jenison, R. (2000). Correlated cortical populations can enhance sound localization performance. *J. Acoust. Soc. Am.*, 107:414–421.

Johnson, D. (1996). Point process models of single-neuron discharges. *J. Comp. Neuroscience*, 3:275–299.

Johnson, D. (2004). Neural population structures and consequences for neural coding. *J. Comp. Neuroscience*, 16.

Johnson, D., Gruner, C., Baggerly, K., and Seshagiri, C. (2001). Information-theoretic analysis of neural coding. *J. Comp. Neuroscience*, 10:47–69.

Johnson, D. and Kiang, N. (1976). Analysis of discharges recorded simultaneously from pairs of auditory nerve fibers. *Biophysical J.*, 16:719–734.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, New York, second edition.

Miller, K. and Troyer, T. (2002). Neural noise can explain expansive, power-law nonlinearities in neural response functions. *J. Neurophysiol.*, 87:653–659.

Nirenberg, S., Carcieri, S., Jacobs, A., and Latham, P. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, 411:698–701.

Panzeri, S., Schultz, S., Treves, A., and Rolls, E. (1999). Correlations and the encoding of information in the nervous system. *Proc. Royal Soc. Lond.*, 266:1001–1012.

Perkel, D. and Bullock, G. (1968). Neuronal coding. *Neurosci. Res. Prog. Bull.*, 6:221–348.

Pola, G., Thiele, A., Hoffmann, K.-P., and Panzeri, S. (2003). An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network: Computation in Neural Systems*, 14:35–60.

Reich, D., Mechler, F., and Victor, J. (2001). Independent and redundant information in nearby cortical neurons. *Science*, 294:2566–2568.

Seung, H. and Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc. Natl. Acad.*, 90:10749–10753.

Shamir, M. and Sompolinsky, H. (2001). Correlation codes in neuronal populations. In Dietterich, T., Becker, S., and Ghahramani, Z., editors, *Advances in Neural Information Processing Systems*, volume 14, pages 277–284. MIT Press, Cambridge, MA.

Sinanović, S. and Johnson, D. (2003). Toward a theory of information processing. In preparation.

Sompolinksy, H., Yoon, H., kang, K., and Shamir, M. (2001). Population coding in neuronal systems with correlated noise. *Phys. Rev. E*, 64:051904.

Stein, R. (1967). The information capacity of nerve cells using a frequency code. *Biophysical J.*, 7:67–82.

Wilke, S. and Eurich, C. (2002). Representational accuracy of stochastic neural populations. *Neural Computation*, 14:155–189.

Wu, S., Amari, S., and Nakahara, H. (2002). Population coding and decoding in a neural field: A computational study. *Neural Computation*, 14:999–1026.

Zohary, E., Shadlen, M., and Newsome, W. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370:140–143.

Figure Captions

Figure 1. Each neuron in a noncooperative population has the same input as the others. Otherwise, the individual neurons do not interact. Here, X represents this common input and Y_n the nth neuron's output. The joint probability distribution of a homogeneous noncooperative population is given by $p_{\mathbf{Y}|X}(\mathbf{y}|x) = \prod_n p_{Y|X}(y_n|x)$, with $\mathbf{Y}$ denoting the collective output.

Figure 2. The left panel shows coding functions and the right shows the corresponding reciprocal-square root of the resulting Fisher information, which equals the smallest attainable root-mean-squared error. The dotted line in each panel corresponds to the equivariance single-neuron coding function. Arbitrarily, $p(1)$ was set equal to $\frac{1}{2}$.

Figure 3. We computed the optimal encoding (constant Fisher information) of the stimulus attribute α by analytically manageable two- and three-neuron populations. In each panel, the solid line indicates how the rate-related parameter should vary with the stimulus attribute to achieve equivariance. These quantities were indistinguishable from their noncooperative counterparts for which we have analytic formulas. The left panel shows the coding functions for spike probability p (solid line) and interneuron correlation ρ (dashed line) for a two-neuron Bernoulli model. The middle panel shows these quantities and ρ_3, a third-order correlation quantity, for the three-neuron Bernoulli model. Both of the correlation parameters are essentially zero in this plot. The right panel illustrates the optimal coding functions for the Poisson counting model, with the total rate corresponding to the solid line and the common rate to the dashed line.

Figure 4. One possible theoretical structure of a cooperative population is a noncooperative population followed by a lateral connection network that produces as close to a statistically independent output as possible. If this network is informationally lossless and does not depend on the details of the joint probability function $p_{\mathbf{Y}}(\mathbf{y})$, this structure would have ideal properties: it would have the information processing properties of the noncooperative structure and Fisher information would be maximized from an equivariance coding perspective.
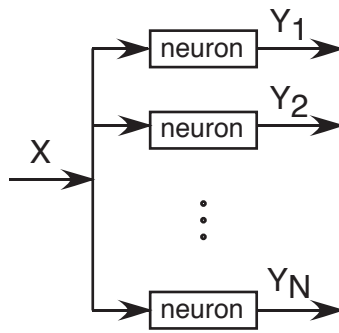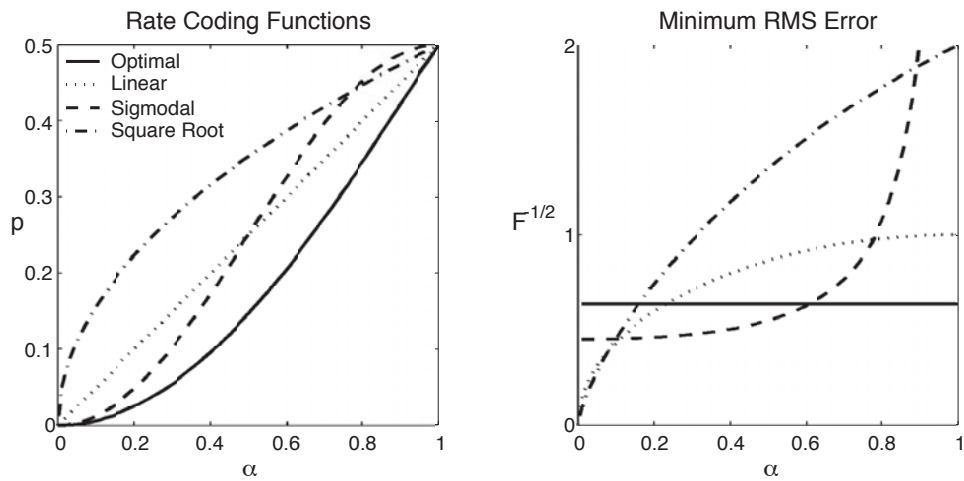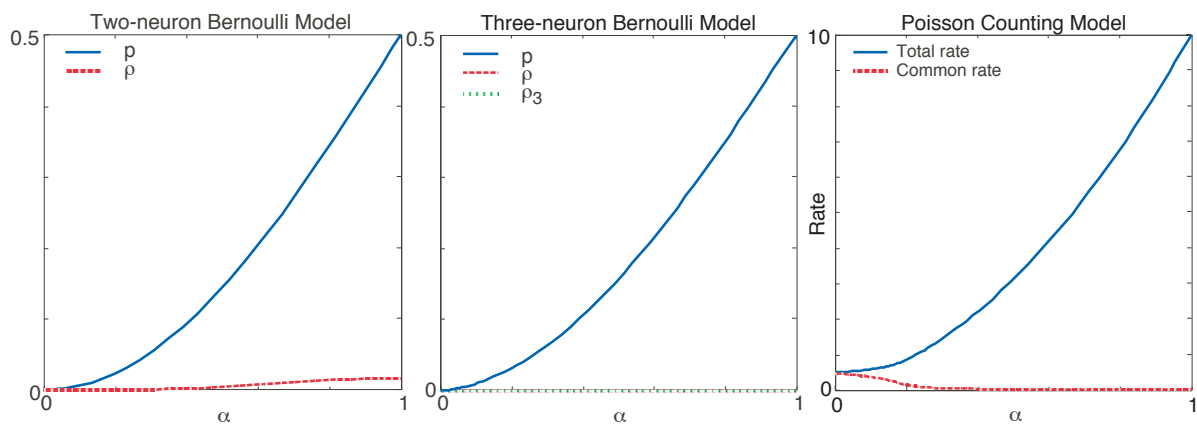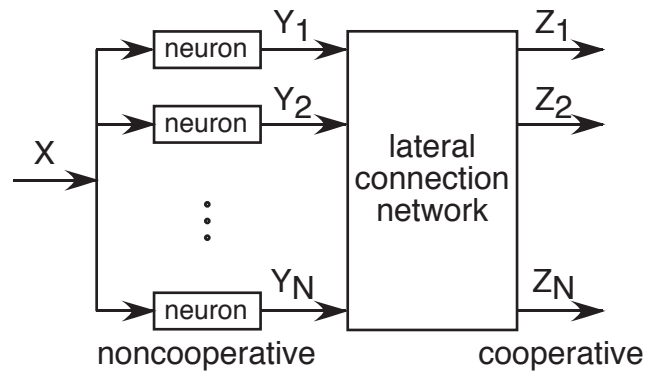
Figure 1:

Figure 2:

Figure 3:

Figure 4:

Appendix

When the neural outputs of a noncooperative population are random variables, the covariance matrix Σ of the population output must have the form

$$\Sigma = \sigma^2(1-\rho)\mathbf{I} + \sigma^2\rho\mathbf{1}\mathbf{1}^t \ ,$$

where σ^2 is the variance of each output Y_n, ρ is the correlation coefficient between any pair of outputs, $\mathbf{I}$ is the identity matrix, and $\mathbf{1} = \mathrm{col}[1, 1, \ldots, 1]$ is a vector of ones. We specialize the decorrelating transformation we seek to be multiplication by a matrix: $\mathbf{Z} = \mathcal{N}[\mathbf{Y}] = \mathbf{A}\mathbf{Y}$. The matrix $\mathbf{A}$ that decorrelates the random vector $\mathbf{Y}$ is not unique. So long as the correlation coefficient is strictly less than one, a matrix that will decorrelate and produce an output of identically distributed random variables is $\Sigma^{-1/2}$, the square root of the inverse of the covariance matrix. Thus, the covariance matrix of $\mathbf{Z} = \Sigma^{-1/2}\mathbf{Y}$ equals the identity matrix. This decorrelating matrix has the following form.

$$\Sigma^{-1/2} = \frac{1}{\sqrt{\sigma^2(1-\rho)}}\left[\mathbf{I} - \left(1 - \frac{1}{\sqrt{1 + \frac{N\rho}{1-\rho}}}\right)\frac{\mathbf{1}\mathbf{1}^t}{N}\right]$$

Consequently, each component Z_n is calculated according to

$$Z_n = \frac{1}{\sqrt{\sigma^2(1-\rho)}}\left[Y_n - \left(1 - \frac{1}{\sqrt{1 + \frac{N\rho}{1-\rho}}}\right)\frac{1}{N}\sum_{i=1}^{N} Y_i\right]$$

Furthermore, the decorrelating matrix is always invertible, which means from an information processing viewpoint that it is lossless.