Fisher and Shannon Information in Finite Neural Populations

Stuart Yarrow
s.yarrow@ed.ac.uk
Institute for Adaptive and Neural Computation
DTC in Neuroinformatics and Computational Neuroscience
School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK.

Edward Challis
e.challis@cs.ucl.ac.uk
Dept. of Computer Science, University College London, London WC1E 6BT, UK.

Peggy Seriès
pseries@inf.ed.ac.uk
Institute for Adaptive and Neural Computation
School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK.

The precision of the neural code is commonly investigated using two different families of statistical measures: (i) Shannon mutual information and derived quantities when investigating very small populations of neurons and (ii) Fisher information when studying large populations. These statistical tools are no longer the preserve of theorists, and are being applied by experimental research groups in the analysis of empirical data. Although the relationship between information theoretic and Fisher-based measures in the limit of infinite populations is relatively well understood, how these measures compare in finite size populations has not yet been systematically explored. We aim to close this gap. We are particularly interested in understanding which stimuli are best encoded by a given neuron within a population and how this depends on the chosen measure. We use a novel Monte Carlo approach to compute a stimulus-specific decomposition of the mutual information (the SSI) for populations of up to 256 neurons and show that Fisher information can be used to accurately estimate both mutual information and SSI for populations of the order of 100 neurons, even in the presence of biologically realistic variability, noise correlations and experimentally relevant integration times. According to both measures, the stimuli that are best encoded are then those falling at the flanks of the neuron's tuning curve. In populations of less than around 50 neurons, however, Fisher information can be misleading.

1 Introduction

Population coding—the transmission of information by the combined activity of many neurons—is known to be a feature of many neural systems. Both experimentalists and theorists have shown great interest in developing tools to assess the precision of population codes. Such methods can be used to help understand the relationship between neural representations and behaviour, as well as between neural activity and environmental stimuli. Informational measures are also useful for assessing the functional consequences of changes in neural response properties, such as observed in sensory adaptation.

Measuring the precision of any neural code involves quantifying how the activity of a neuron, or neurons, relates to some measurable quantity in the external world, typically a feature of a presented stimulus or an observed action. The precision of the code is essentially the degree to which the neural activity reflects the quantity of interest, but several different methods of quantifying this interdependency exist, and it is not always clear exactly what they imply or how they relate to each other. For population codes, the situation is further complicated by the number of neurons involved; in all but the simplest systems—such as the cricket cercal interneurons discussed below (Theunissen and Miller, 1991)—it is impossible to identify and record from all cells involved. This means that the measured activity can only be a small sample of the activity of the population, although the situation is improving due to the increasing use of multi-electrode arrays and two-photon calcium imaging.

This article first provides an overview of the principle measures used to assess coding precision, with emphasis upon their intuitive interpretation and practical application in experimental neuroscience. We then go on to address some previously unanswered questions regarding the relationship between Fisher information and Shannon mutual information in populations with a finite number of neurons. Following the work of Butts and Goldman (2006), we also examine in detail how the stimulus-specific precision of a neuron, and in particular the identification of the stimuli that are best encoded by a given cell, depends upon the measure used.

1.1 A probabilistic view of neural coding

Before discussing the precision measures that are the main focus of this article, it is worth clarifying, in probabilistic terms, what is being measured in a typical sensory electrophysiology experiment. Let's assume that the experiment consists of a large number of trials in which a stimulus is presented and the response of a neuron, the number of action potentials elicited, is recorded over a given time window. By repeatedly presenting a stimulus, sufficient data can be gathered to estimate the distribution of the responses; this process can then be repeated for a range of stimuli. The resulting model is a conditional distribution—the distribution $p(R|\Theta)$ of the response R conditioned upon the stimulus Θ.

Classical single electrode techniques only allow the recording of one (or very few) cells simultaneously. This means that it is not possible to measure interdependencies between the activity of cells. In this situation it is usual to assume that the activities of each cell are conditionally independent given the stimulus, i.e. that the trial to trial variability or noise is independent. However, this can lead to under or overestimation of the precision of the code (see e.g. Averbeck et al., 2006). In order to characterise inter-neuronal correlations in the variability—'noise correlations'—it is necessary to simultaneously record from multiple cells, for example through multi-electrode array or two-photon calcium imaging techniques.

1.2 Information theory

Information theory is a mathematical framework proposed in the 1940s by engineer and mathematician Claude Shannon (1948). While originally intended as a tool for analysing telecommunications systems, information theory is more generally applicable and has been widely utilised in other fields (Cover and Thomas, 2006). In contrast to many other statistical techniques, information theory does not rely on any assumptions about the form of distributions or the properties of underlying processes. It quantifies all forms of probabilistic interdependency between variables, unlike less general statistics such as the correlation coefficient.

The basic quantity of information theory is information entropy, a measure of the uncertainty or randomness of a variable. Entropy can be intuitively, but very loosely, thought of as a generalisation of variance; while variance has a special relevance to the Gaussian distribution, entropy is equally applicable to any arbitrary distribution. More correctly, entropy is the amount of information required, on average, to represent the value of a variable, and, for the purposes of this article, is measured in bits. The entropy $H(\Theta)$ of a stimulus ensemble Θ is given by: [1]

$$H(\Theta) = -\sum_{\theta \in \Theta} p(\theta) \log_2 p(\theta) \tag{1}$$

Shannon or mutual information I_{mut}, is a measure of the informativeness of one variable about another e.g. of a neural response R about a stimulus Θ. It is the portion of a variable's entropy that can be explained by the other variable;

[1] Since both the stimulus and response variables in our model are continuous, all the entropies calculated in our analyses are differential entropies. These are largely equivalent to discrete entropy as described here, but are obtained by integrating over a continuous distribution rather than summing over a discrete distribution. See Appendix B.1 for further details.

specifically, it is the total entropy minus the conditional entropy:

$$\begin{aligned} I_{mut}(\Theta, R) &= H(R) - H(R|\Theta) = H(\Theta) - H(\Theta|R) \\ &= \sum_{\theta \in \Theta} p(\theta) \sum_{r \in R} p(r|\theta) \log \frac{p(r|\theta)}{p(r)} \end{aligned} \tag{2}$$

Uppercase characters Θ and R represent the stimulus and response ensembles, while lowercase characters (θ, r) represent a single value within the ensemble.

Mutual information can be used to quantify the information provided by an entire response ensemble about an entire stimulus ensemble, but it cannot inform us about the precision with which specific stimuli within the ensemble are encoded. To address this, several decompositions of the mutual information have been proposed (see Butts, 2003, for a review), in particular the stimulus-specific surprise, specific information and stimulus-specific information.

Stimulus-specific surprise is the most widely used MI decomposition. Like all of the stimulus-specific measures described here, the average of the specific surprise over the stimulus ensemble is equal to the mutual information. Equation 3 illustrates an intuitive interpretation: the specific surprise is the reduction in surprise (log reciprocal probability) of a given stimulus, averaged over the response ensemble. The specific surprise was one of the first stimulus-specific measures to be applied to population coding (Theunissen and Miller, 1991), there referred to as local transinformation. Confusingly, specific surprise is also referred to in some articles as stimulus-specific information.

$$I_{sur}(\theta) = \sum_{r \in R} p(r|\theta) \log \frac{p(r|\theta)}{p(r)} = \sum_{r \in R} p(r|\theta) \left[\log \frac{1}{p(\theta)} - \log \frac{1}{p(\theta|r)} \right] \tag{3}$$

Specific information is a mutual information decomposition that quantifies the decrease in uncertainty about the stimulus due to the observation of a given response:

$$I_{SI}(r) = \sum_{\theta \in \Theta} p(\theta|r) \log p(\theta|r) - p(\theta) \log p(\theta) \tag{4}$$

The specific information has a unique and advantageous property in that it is additive (DeWeese and Meister, 1999): the sum over the specific information associated with a number of individual observations is equal to the specific information of the whole set considered jointly.

The stimulus specific information (SSI) is a stimulus-specific development of the specific information (Butts, 2003). The SSI is the average specific information associated with a given stimulus:

$$\begin{aligned} I_{SSI}(\theta) &= \sum_{r \in R} p(r|\theta) I_{SI}(r) \\ &= \sum_{r \in R} p(r|\theta) \left[\sum_{\theta \in \Theta} p(\theta|r) \log p(\theta|r) - p(\theta) \log p(\theta) \right] \end{aligned} \tag{5}$$

In this article we discuss both the population SSI (the SSI of the population as a whole) and the singleton SSI, which is the SSI of a single neuron considered in isolation. A closely related quantity, the marginal SSI (mSSI) for a particular neuron within the population, is defined as the difference between the population SSI and the SSI for the population of remaining neurons with the neuron of interest removed.

The SSI is a relatively recent development and has not yet been explored or applied as widely as the specific surprise. The SSI was until recently considered to be intractable for all but small populations; Butts and Goldman (2006) calculated the SSI for a maximum of four neurons. The SSI has been used to analyse experimental data from single neurons only (Sawtell and Williams, 2008; Montgomery and Wehr, 2010). In this article we demonstrate that this can be overcome through the use of Monte Carlo integration to compute the average over the high-dimensional response ensemble.

Specific surprise and SSI are both stimulus-specific decompositions of the mutual information, so how do they differ? The SSI tells us the average reduction in uncertainty—about all possible values of stimulus—that results from the presentation of a given stimulus. The specific surprise is the average amount by which the surprise of a given stimulus reduces following the presentation of that stimulus. The SSI could therefore be considered less stimulus-specific than the specific surprise, since it relates to an observer's knowledge of the full stimulus ensemble (Butts, 2003).

All information-theoretic measures have one major disadvantage in an experimental neuroscience context. In order to calculate any of these measures directly, it is necessary to establish the full joint distribution $p(\Theta, R)$. In a model this is relatively simple, but in an experimental context it is, at best, very difficult to record the number of trials necessary to establish an accurate joint distribution. One method that has been proposed to avoid the problem of constructing the joint distribution $p(\Theta, R)$ involves calculating the transmitted information using spike train metrics (Victor and Purpura, 1997; Victor, 2005). Since this method relies on stimulus-dependent clustering, it is inherently suited to assessing classification of discrete stimuli. Another approach, which is suited to assessing the discrimination of continuous-valued stimuli, is to estimate the mutual information by calculating the Fisher information, as described in the following section. One of the goals of this article is to assess the validity of this approximation.

1.3 Fisher information

Fisher information is a statistical measure of precision commonly used in both theoretical (see e.g. Paradiso, 1988; Seung and Sompolinsky, 1993; Abbott and Dayan, 1999; Wilke and Eurich, 2002; Berens et al., 2011) and experimental (e.g. Jenison and Reale, 2003; Harper and McAlpine, 2004; Durant et al., 2007; Gutnisky and Dragoi, 2008) studies of population coding. Fisher information J is

defined as:

$$J(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(r|\theta)\right)^2 \middle| \theta\right] \tag{6}$$

In a population code with Gaussian variability, mean response vector (tuning function) $\boldsymbol{f}(\theta)$ and covariance matrix $Q(\theta)$, the Fisher information about θ is given by:

$$J(\theta) = \boldsymbol{f}'(\theta)^T Q(\theta)^{-1}\boldsymbol{f}'(\theta) + \frac{1}{2}\,\mathrm{Tr}\left[Q(\theta)^{-1}\,Q'(\theta)\,Q(\theta)^{-1}\,Q'(\theta)\right] \tag{7}$$

Despite its name, it is not a measure of information in the information theoretic sense; its units are those of the reciprocal of variance (e.g. deg^{-2} for an angular stimulus). Fisher information is perhaps more intuitive than the information theoretic measures: its reciprocal defines a lower limit (the Cramér-Rao bound) on the variance of an unbiased estimator[2], and hence the smallest achievable standard error. Unfortunately, this level of precision is not necessarily achievable; the performance of an optimal estimator only approaches the Cramér-Rao bound asymptotically as the population size tends towards infinity. Predicting what population size is required for effective saturation of the bound is non-trivial, and this question has rarely been addressed in the literature (Bethge et al., 2002; Xie, 2002). Fisher information should therefore be treated with some caution, as it is not always clear whether it indicates the true coding precision of a population.

1.4 Linking Fisher and Shannon

Brunel and Nadal (1998) linked Fisher and Shannon information by proposing I_{Fisher}, a new information theoretic measure derived from Fisher information. They considered an optimal estimator $\hat{\Theta}(R)$, computed from R, with a Gaussian conditional distribution $p(\hat{\Theta}(R)|\Theta)$ and variance that saturates the Cramér-Rao bound. This is equivalent to assuming that the population size is infinite and therefore that each estimate $\hat{\theta}(\boldsymbol{r})$ is based on an infinite number of independent observations. Given these assumptions, we can determine the conditional entropy of the estimator from the variance, and hence from the Fisher information, using the following relation:

$$h(\hat{\Theta}(R)|\Theta=\theta) = \frac{1}{2}\log_2(2\pi e\sigma^2) = \frac{1}{2}\log_2\left(\frac{2\pi e}{J(\theta)}\right) \tag{8}$$

This gives the conditional entropy for a specific stimulus value θ; to obtain $h(\hat{\Theta}(R)|\Theta)$ it is necessary to take the average over the stimulus ensemble:

$$h(\hat{\Theta}(R)|\Theta) = \mathbb{E}_\Theta[h(\hat{\Theta}(R)|\Theta=\theta)] = \int_\Theta p(\theta)\frac{1}{2}\log_2\left(\frac{2\pi e}{J(\theta)}\right)d\theta \tag{9}$$

[2] A function of r that yields an estimate of θ.

The mutual information of stimulus and estimator is therefore given by:

$$I_{mut}(\Theta, \hat{\Theta}(R)) = h(\hat{\Theta}(R)) - \int_{\Theta} p(\theta) \frac{1}{2} \log_2 \left(\frac{2\pi e}{J(\theta)} \right) d\theta \tag{10}$$

Using the data processing inequality to relate $I_{mut}(\Theta, \hat{\Theta}(R))$ and $I_{mut}(\Theta, R)$:

$$I_{mut}(\Theta, R) \geq h(\hat{\Theta}(R)) - \int_{\Theta} p(\theta) \frac{1}{2} \log_2 \left(\frac{2\pi e}{J(\theta)} \right) d\theta \tag{11}$$

and showing that i) this inequality becomes an equality in the limit of large N and under certain regularity conditions, and ii) that $h(\hat{\Theta}(R)) \to h(\Theta)$ in the limit where the estimator is sharply peaked around its mean value (i.e. $J(\theta) \gg 1$), Brunel and Nadal show that $I_{mut}(\Theta, R)$ can be approximated by:

$$I_{Fisher} = h(\Theta) - \int_{\Theta} p(\theta) \frac{1}{2} \log_2 \left(\frac{2\pi e}{J(\theta)} \right) d\theta \tag{12}$$

which they call I_{Fisher}, since it is defined in terms of Fisher information. To our knowledge, no assessment of how good this approximation is for finite populations has previously been made.

In summary, information theory provides us with measures that are very powerful, but which can be difficult to apply in practice. Other statistical measures, such as the Fisher information, are often easier to measure or calculate, but it is not always clear exactly what they tell us, or what the precise limits of their applicability are. I_{Fisher} goes some way towards bridging the gap between mutual information and Fisher information by allowing their absolute values to be compared, in the special case of an infinite population.

1.5 Applications of Fisher and Shannon information in neuroscience

Information measures tell us about the precision of neural representations and, through careful selection of what is being measured, can also be used to address other questions about neural codes. Here we include a few examples of the use of Fisher and Shannon information in neuroscience, to illustrate the range of possible applications. For more detailed information on applications of information measures in the field of neural coding, see reviews by Borst and Theunissen (1999), Sanger (2003), Averbeck et al. (2006), Nelken and Chechik (2007) and Quian Quiroga and Panzeri (2009).

Information measures can be used to accurately assess how precision changes when properties of the neural response change, such as through adaptation. Fairhall et al. (2001) recorded from a single motion-responsive neuron in the fly visual system and used information theory to show that the average information per spike was maintained through adaptation as the variance of the stimulus distribution was manipulated. A similar analysis of sound intensity coding in the mammalian

midbrain, this time using Fisher information, showed that intensity tuning curves adapted to changes in the stimulus statistics, allowing precision to be maintained across a wide stimulus dynamic range (Dean et al., 2005). Fisher information has also been used to measure how adaptive changes in noise correlations affected the precision of orientation representation by cells in macaque V1 (Gutnisky and Dragoi, 2008). Seriès et al. (2009) used Fisher information together with simulated decoding to analyse the reconstruction precision and bias associated with various models of neural decoding.

The nature of the neural code—which aspects of cell and population activity are information bearing—is generally unknown. By comparing the coding precision of various response properties (e.g. firing rate, spike times or inter-spike intervals), information measures can be used to address this question. An example of this type of analysis is the work of Panzeri et al. (2001) on the representation of whisker stimuli in the barrel cortex of the rat. In this study, information theory was used to examine whether spike times conveyed information about spatial aspects of the stimulus by computing the time course of information accumulation following the stimulus presentation for both spike count and spike times. In this case, spike timing was found to contribute a significant amount of information beyond that carried by the spike count alone. More generally, the inherent temporal precision of a code can be found by perturbing the spike times by introducing progressively larger amounts of temporal noise, and noting how the precision of the code degrades as a function of the amount of jitter (Quian Quiroga and Panzeri, 2009).

Informational measures can also be used to examine which aspects of the stimulus are best encoded—most precisely represented—by a cell or population. In this case, the type of code (e.g. spike count versus spike timing) is fixed, and the amount of information transmitted about various stimulus properties is compared. Machens et al. (2005) used this approach to determine the optimal stimulus ensemble—the distribution of stimuli that maximised the information transmitted by a neuron—for grasshopper peripheral auditory neurons. The optimal stimulus ensemble was found to coincide with grasshopper communication sounds and not with natural sounds in general, indicating that the communication calls and auditory system were well matched. Panzeri et al. (2001) also provide us with an example of the use of stimulus-specific surprise to identify which whiskers are most precisely represented within a given barrel.

Information measures can also be used to determine the optimal arrangement of tuning curves in order to cover a given range of stimuli. Harper and McAlpine (2004) conducted a theoretical study to determine the optimal (in terms of Fisher information) frequency tuning for populations of auditory neurons selective for interaural time difference (ITD). The study predicted that cells that responded to frequencies below a certain species-specific threshold were more likely to respond maximally to ITDs that were outside the range that occurs in nature. This arrangement leads to the flanks of the tuning curves—the regions of maximum Fisher information—coinciding with the physiological range of ITDs, and was in

agreement with experimental findings in small mammals.

The relationship between neural precision and behavioural performance is a key area of neural coding research. In order to examine this relationship, it is necessary to ensure that both measures, neural and behavioural, are addressing equivalent questions. Fisher information is rather inflexible in this respect, as it only tells us about the precision of fine discrimination or stimulus reconstruction, and not about coarser discrimination, classification or detection tasks. Information theoretic measures are more flexible as they can be tailored to suit a particular task by changing the stimulus ensemble. An alternative approach is to explicitly model a decoder that mimics the decision required by the task; in this case the performance of the decoder can be directly compared to behavioural performance. See Oram et al. (1998) and Quian Quiroga and Panzeri (2009) for reviews that cover this approach.

1.6 Outline

Both Fisher information and information theoretic measures are now widely used for the study of neural codes. These tools are no longer the preserve of theorists, and are being applied by experimental research groups in the analysis of empirical data. Fisher information is a particularly accessible tool for experimentalists, as it is generally easier to calculate than information theoretic measures, in terms of both data requirements and computational complexity.

While both measures are widely used, studies almost invariably make use of either Fisher information (when measuring whole populations) or information theory (for studying single neurons). This leads to difficulties in comparing the findings of studies based on different measures, since they are rarely applied to the same cases. Are the two families of measure interchangeable; do they ultimately provide the same results as to which stimuli are best encoded? The answer to this question is: sometimes (Butts and Goldman, 2006), but to date this issue has only been examined for very small populations (number of neurons $N = 4$). For most biologically relevant population codes, the relationship between Fisher and Shannon information is unclear. Resolving this ambiguity is of crucial importance to bridge the gap between the Fisher information and information theoretic strands of the literature.

In the remainder of this article we employ numerical models of simplified, but broadly biologically realistic, populations to clarify the link between Fisher and Shannon information. We also examine in detail the limits of applicability of Fisher information. How many neurons are required before I_{Fisher} provides a good working estimate of I_{mut}? How does Fisher information relate to information theoretic measures? We go on to show, through numerical simulation, that Fisher information can be used to obtain the asymptotic value of SSI, in the same way that it can provide the asymptotic value of I_{mut}.

2 Model Framework

We consider here a population of N sensory neurons encoding a unidimensional circular stimulus variable θ, which represents a direction e.g. of a moving bar. Each experiment consists of a number of virtual trials, in which the spike count r_i of each neuron over a time interval τ is computed. Each presentation of a stimulus θ is therefore associated with a response vector $\boldsymbol{r} = [r_1 \dots r_N]$. For the purposes of this study, information is assumed to be encoded exclusively by the spike counts; the timing of individual spikes within the measurement window is disregarded. Although this represents a simplification, the rate coding model is frequently employed for its tractability and has been shown to be valid in a number of contexts (Heller et al., 1995; Tovée et al., 1993).

The response of each neuron can be represented by a deterministic component (the tuning curve) that defines the mean response over many trials, and a random component that models the trial-to-trial variability or noise; these are described in the following sections. The model framework described below (sections 2.1–2.2) was used in all experiments described in this article, except for those based on the cricket cercal interneuron model described by Theunissen and Miller (1991), which is covered here in section 2.4.

2.1 Tuning curves

The mean firing rates of each neuron were modelled by a circular Gaussian function, given here for the ith neuron:

$$f_i(\theta) = f_{bg} + f_{max} \ \exp\left[-\frac{1-\cos(\theta-\phi_i)}{(\frac{\pi}{180}\sigma_f)^2}\right] \tag{13}$$

Where f_{max} and f_{bg} are the peak firing rate and stimulus-independent background firing rate respectively, both measured in spikes/s; ϕ_i is the preferred stimulus of the ith neuron; θ is the stimulus angle and σ_f a width parameter. Unless otherwise stated, the following parameter values were used in all simulations involving this tuning function: $f_{max} = 50\,\text{spikes/s}$, $\sigma_f = 30°$. In all simulations the neurons' preferred stimuli were uniformly distributed around the 360°range of the stimulus angle.

2.2 Trial-to-trial variability

Trial-to-trial variability was modelled by a multivariate Gaussian distribution:

$$\boldsymbol{r} \sim \mathcal{N}[\tau \boldsymbol{f}(\theta), Q(\theta)] \tag{14}$$

Where $\boldsymbol{r}$ is the vector of spike counts recorded in response to stimulus θ, $\boldsymbol{f}(\theta)$ is the vector of mean neuronal responses defined in the preceding section and τ is the integration time over which spike counts are recorded in each trial. In order

to construct the inter-neuronal covariance matrix $Q(\theta)$, it is first necessary to establish the variance of each individual neuron and any correlations in trial to trial variability.

A multiplicative model of neuronal variability was used:

$$\sigma_i^2(\theta) = F\tau f_i(\theta) \tag{15}$$

Where F is the Fano factor, the ratio of the spike count variance σ_i^2 to the mean spike count $\tau f_i(\theta)$ over the time interval τ. This type of model can be viewed as a generalisation of Poisson noise. The Poisson distribution is rather inflexible as it has only a single parameter, with the Fano factor fixed at unity. By using a Gaussian noise model, we gain an extra parameter and with it the flexibility to adjust the Fano factor. In addition to this, the Fisher information can be found analytically, without having to resort to time consuming numerical methods. For this reason, negative spike counts have not been rectified to zero, as this would render the variability non-Gaussian. Using a non-zero background firing rate helps to prevent the occurrence of negative spike counts, and a value of $f_{bg} = 10\,\text{spikes/s}$ has been used in most simulations.

Correlations in the trial-to-trial variability are defined by a correlation matrix C. Three forms of the correlation matrix are examined in this article:

- Independent trial-to-trial variability i.e. uncorrelated noise: C is the identity matrix:

  $$C_{ij} = \delta_{ij} \tag{16}$$

  Where δ is the Kroeneker delta function.

- Localised correlations, specifically correlations that decay exponentially as a function of the difference in preferred stimuli. In this case C is given by:

  $$C_{ij} = \delta_{ij} + (1 - \delta_{ij})\, c\, \exp\left(-\frac{|\phi_i - \phi_j|}{\rho}\right) \tag{17}$$

  Where c is a correlation scaling coefficient and ρ is a correlation range coefficient. The correlation scales examined here ($0 \leq c \leq 0.3$) cover most biologically realistic scenarios. Unless otherwise stated, a range value of $\rho = 30^\circ$ was used, meaning that the extent (in stimulus space) of the noise correlations and tuning curves were matched.

- Uniform correlations, where every pair of neurons has a trial-to-trial variability correlation coefficient of c:

  $$C_{ij} = \delta_{ij} + (1 - \delta_{ij})\, c \tag{18}$$

Once the correlation matrix has been defined, the covariance matrix is given by:

$$Q_{ij}(\theta) = F\, [\tau f_i(\theta)]^{0.5}\, C_{ij}\, [\tau f_j(\theta)]^{0.5} \tag{19}$$

A number of factors determine the coding precision of a population, principally: the maximum and minimum (background) mean firing rates, the level of trial-to-trial variability, and the integration time over which spike counts are recorded. The Fano factor variability model was used as it allowed a convenient simplification to be made; increasing F clearly increases the variability of the response, while increasing τ means that we average the response over a longer time window and hence reduce the effective level of variability. In the case of Gaussian noise where the variance is determined by a Fano factor, F and τ have exactly equal and opposite effects, so we can fully capture the effect of both parameters by considering only their ratio F/τ, which has units of spikes/s^2. Further details of this simplification are given in Appendix A.1.

We explore F/τ initially in the interval $[10^{-4}, 10^3]$ spikes/s^2, but most of our analyses extend only up to $F/\tau = 100$ spikes/s^2. Since F is commonly thought to be in the range $[1, 3]$, the highest values of F/τ can be thought of as corresponding to recording time windows in the region of 10–30 ms. Some caution is required when applying our model with high values of F/τ. Very short integration times lead to low mean spike counts, and bring the model into a regime where the Gaussian distribution is no longer a good approximation of the Poisson-like distribution of real neuronal responses. For this reason we have restricted most of our analyses to $F/\tau \leq 100$ spikes/s^2.

2.3 Gain modulation

To examine the effect of adaptation-like gain changes on precision, we used the following gain modulation model (Seriès et al., 2009):

$$f^i_{max} = f_{max}\left[1 - \beta \exp\left(-\frac{1-\cos(\phi_i - \phi_{mod})}{(\frac{\pi}{180}\sigma_{mod})^2}\right)\right] \tag{20}$$

Where f^i_{max} is the post-modulation peak firing rate of the ith neuron and f_{max} is the original peak firing rate common to all neurons. The 'adapting stimulus' and extent of adaptation (centre and width of the modulation profile) are defined by ϕ_{mod} and σ_{mod} respectively, while $\beta = [0, 1]$ is the modulation depth.

2.4 Cricket cercal system model

We also re-implemented a model of the cercal interneurons in the cricket first described by Theunissen and Miller (1991). The formulation given here is that used by Butts and Goldman (2006).

The stimulus model is as described above, and the population consists of four neurons with preferred directions evenly spaced at 90°intervals around the 360°stimulus space. The mean response is given by a rectified cosine tuning curve:

$$f_i(\theta) = \frac{\cos(\theta - \phi_i) - 0.14}{0.86} \tag{21}$$

The standard deviation of the neuronal response is defined as a linear function of the mean response, hence the variance is a quadratic function of the mean (c.f. Equation 15, where the variance is a linear function of the mean). The parameter A is a variability scaling factor.

$$\sigma_i = A[0.048 + 0.052 f_i(\theta)] \tag{22}$$

The cosine tuning curve and noise are added together and negative values are rectified to zero, yielding the response spike count:

$$r_i(\theta) = [f_i(\theta) + \eta]_+$$
$$\eta \sim \mathcal{N}(0, \sigma_i^2)$$

The rectification has the effect that the variability becomes non-Gaussian; note that this is in contrast to the other simulations described in this article, where negative spike counts are not rectified in order to preserve Gaussianity. For the cricket cercal system model (Figure 4) only, Fisher information is calculated by Monte Carlo integration in order to take into account the non-Gaussian response distribution. The SSI calculations for this model assume a Gaussian response distribution and are therefore an approximation.

3 Mutual information and I_{Fisher}

While it is known that mutual information and I_{Fisher} are equal for infinite populations, how they are related in finite populations is less clear. As discussed in section 1.4, it has been shown (Brunel and Nadal, 1998) that I_{Fisher} forms an upper bound on the mutual information, and that the mutual information approaches this bound asymptotically as N tends to infinity. To verify this numerically, and to establish the population size required for I_{Fisher} to provide an accurate estimate of the mutual information, a series of population models were examined. In addition, a four-neuron population model was used to assess the effect of trial-to-trial variability and background activity in very small populations.

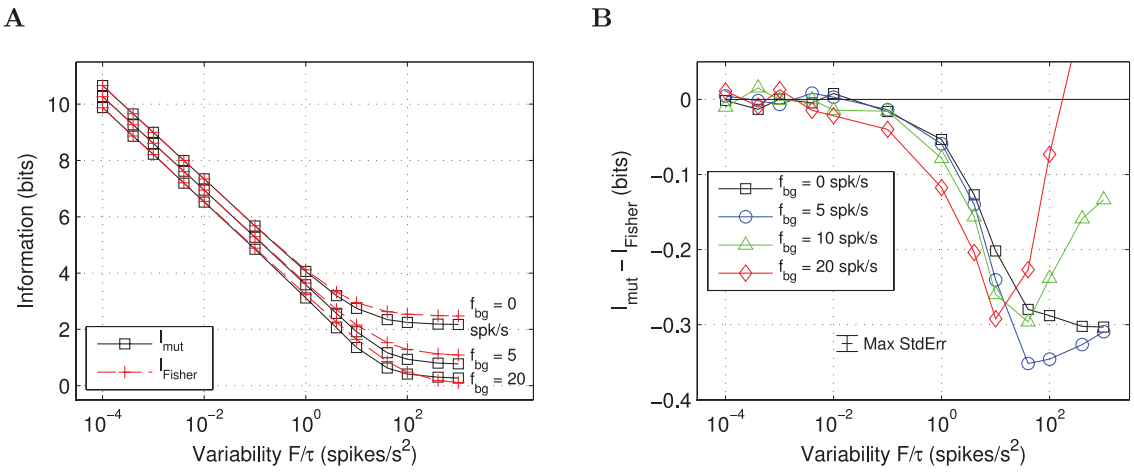

Figure 1: A comparison of mutual information and I_{Fisher} in a four-neuron population, showing the effect of trial-to-trial variability (noise) and background activity.
(A) Mutual information and I_{Fisher} as a function of variability, for several levels of background activity. Increasing the background activity increases the signal to noise ratio and reduces information. I_{Fisher} diverges from mutual information with increasing F/τ; background activity accelerates this divergence. At high levels of variability, both I_{mut} and I_{Fisher} flatten out and do not reduce with further increases in F/τ. Error bars too small to plot; see B.
(B) Difference between I_{mut} and I_{Fisher} for the same data shown in A, together with an additional case $f_{bg} = 10$ spikes/s. The maximum standard error across all points is shown on the plot.

Figure 1 shows how mutual information and I_{Fisher} for a very small population (four neurons) vary as a function of the trial-to-trial variability. Mutual information is almost equal to I_{Fisher} when the noise level is low, even in a population of only four cells, and the difference between the two measures increases as the variability increases (see Figure 1B). Both mutual information and I_{Fisher} decrease with increasing variability, and are almost logarithmically proportional to F/τ (see Figure 1A).

Background activity has a similar effect to variability. Since background activity is uniformly present and gives no information about the stimulus it is essentially noise, therefore increasing the background activity reduces the signal to noise ratio, and this drop in SNR results in lower information values. Increased background activity also contributes to the divergence of MI and I_{Fisher}, leading to greater differences between the two measures for a given level of trial-to-trial variability.

For large values of F/τ (roughly corresponding to integration times of less than 10 ms with a Fano factor of 1), both I_{mut} and I_{Fisher} flatten out and do not reduce with further increase in variability. In terms of Fisher information, this can be understood as the regime within which the trace term in Equation 7 is dominant (Shamir and Sompolinsky, 2004). In this regime, information is encoded primarily by the stimulus-dependent response variances, as opposed to the mean responses.

In general, I_{Fisher} forms an upper bound upon the mutual information, as shown by Brunel and Nadal (1998). However, for very high levels of variability combined with background activity, I_{Fisher} can be less than the mutual information (e.g. when $f_{bg} = 20\,\text{spikes/s}$ in Figure 1), and can even become negative (unlike the Fisher information, which is inherently non-negative). This occurs because the amount of noise in the system is such that the overall entropy of the response becomes significantly greater than the stimulus entropy, while the derivation of I_{Fisher} relies on the assumption that the entropies of stimulus and response are approximately equal. I_{Fisher} is therefore best at predicting the mutual information when I_{Fisher} is non-negative and within the logarithmically proportional regime with respect to F/τ.

Figure 2 shows the effect of population size upon I_{mut} and I_{Fisher} under a number of different variability regimes. Figure 2A shows that I_{Fisher} is essentially proportional to $\log N$ over the range of population sizes examined. The asymptotic approach of the mutual information to the bound formed by I_{Fisher} is evident in Figure 2B, which shows the difference between the two measures plotted against N. Increasing the variability F/τ increases the difference between I_{mut} and I_{Fisher} for a given population size. Despite this, even for the highest level of variability modelled (e.g. equivalent to supra-Poisson variability $F = 3$ with a time window of 30 ms), there is a difference of only 3.5% between the two measures for a population of 50 neurons. For $\tau = 300\,\text{ms}$, $F = 3$, the same relative error is achieved with less than 20 neurons.

From a decoding perspective, increasing the population size means that there are more parallel 'channels' carrying information about the stimulus. With a greater number of channels, a decoder can better average out the variability of these channels, hence coding precision is increased. The information carried by each channel becomes increasingly redundant as N increases, so the gain in coding precision diminishes; this is why we observe that information is approximately proportional to $\log N$ rather than N.

The relationship between I_{mut} and I_{Fisher} is complicated slightly when there are inter-neuronal correlations in the trial-to-trial variability. Figures 2C and 2D show the effect of uniform correlations. The presence of uniform correlations slightly increases the information conveyed by the population, but the effect is much less than that of altering the level of variability. The information increase due to uniform correlations is effectively independent of population size. The reason for this increase in coding precision can be understood by considering the extreme case of $c = 1$. In this scenario, the noise correlation coefficient for every pair of neurons is 1, therefore every cell in the population exhibits exactly the same random noise. The relative firing rates of the neurons (the profile of activity across the whole population, determined by the tuning curves) are thus perfectly preserved, allowing very accurate decoding (see Averbeck et al., 2006, for further explanation of how noise correlations affect the precision of population codes).

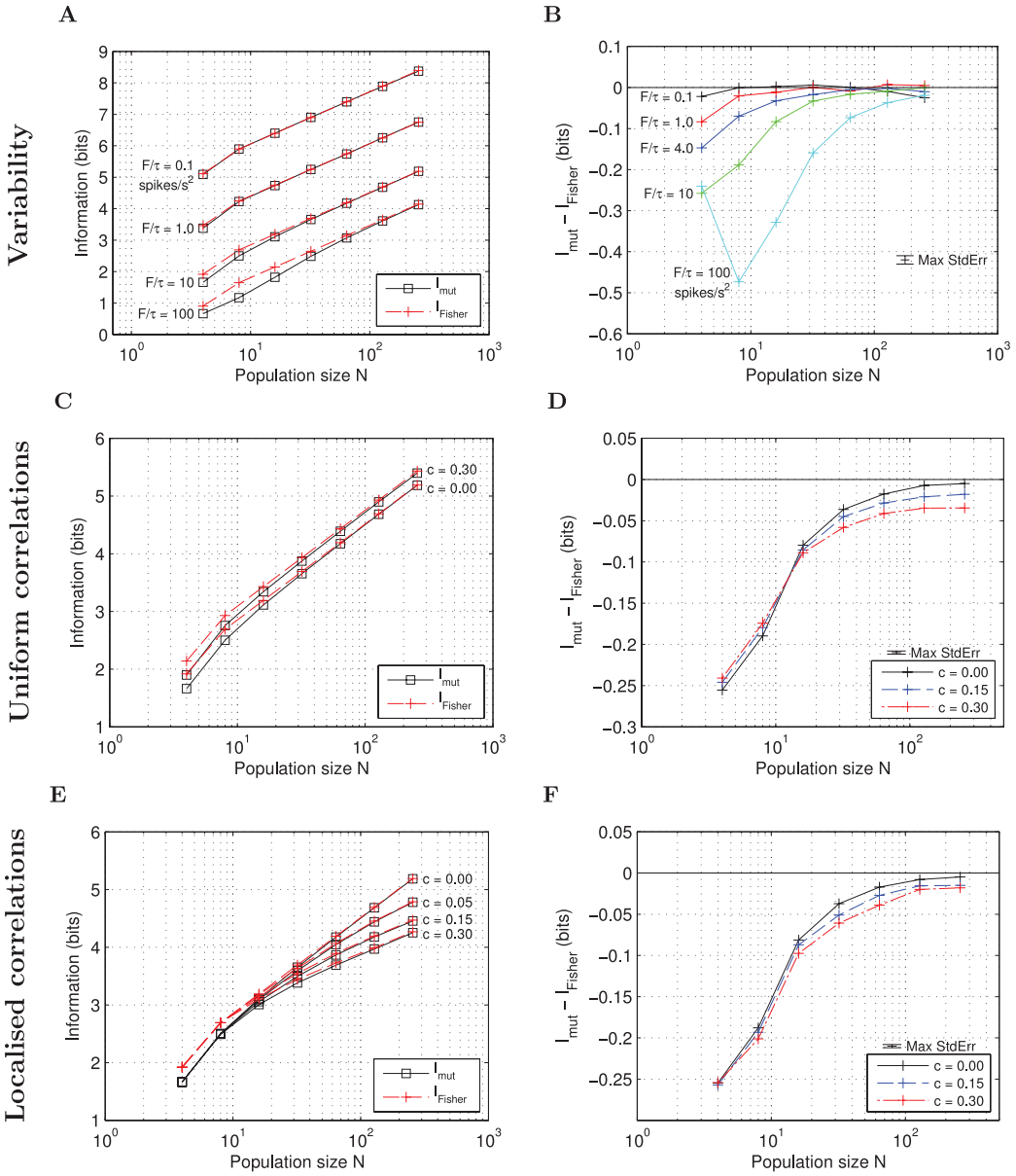

Figure 2: Relationship between mutual information and I_{Fisher} in populations varying in size from 4 to 320 neurons. I_{Fisher} is, in most cases, a good approximation of I_{mut}. The two measures diverge only for small populations ($N < 100$). Errors are shown as in Figure 1.
(A) I_{mut} and I_{Fisher} for various levels of independent variability. I_{mut} converges towards I_{Fisher} from below with increasing N. Parameters: $f_{bg} = 10\,\text{spikes/s}$.
(B) This plot shows the difference between I_{mut} and I_{Fisher} for the same cases as A.
(C, D) Absolute values, and difference between, I_{mut} and I_{Fisher} for various values of c with uniform correlation structure. Uniform correlations increase coding precision, but delay convergence of I_{mut} and I_{Fisher}. Parameters: $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$.
(E, F) As per C and D, but with a localised correlation structure. Localised correlations reduce coding precision and delay convergence between I_{mut} and I_{Fisher}. Parameters: $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$.

Figures 2E and 2F illustrate the effect of localised correlations. In contrast to uniform correlations, these act to reduce coding precision, although this effect is again small in comparison to that of variability. In large populations, the presence of localised correlations can have a marked effect, as it greatly reduces the rate with which both Fisher and mutual information increase with $\log N$ (Wilke and Eurich, 2002).

In general, both uniform and localised correlations act to increase the difference between the mutual information and I_{Fisher}, although this effect is small in comparison to that of changing the level of variability. For very small populations ($N < 10$), however, uniform correlations actually reduce the difference. The effects of localised correlations vanish as the population size decreases, as the increasing spacing between tuning curves leads to a general reduction in pairwise correlation coefficients across the population. The effect of correlations, both uniform and localised, on the difference between I_{mut} and I_{Fisher} is greatest for large populations, in contrast to the effect of variability, which diminishes with increasing N. As a result of this, correlations reduce the rate of convergence of the two measures, whereas variability itself does not.

The three noise correlation scenarios examined here can be seen as lying on a single continuum, where uniform correlations are equivalent to localised correlations with infinite range, and independent variability corresponds to zero range. The correlation range parameter ρ can be varied continuously, allowing the change in coding precision across this continuum to be explored. It is most useful to consider the correlation range relative to the width of the tuning curves, as the tuning curve width determines the extent of activity and the range of signal (as opposed to noise) correlations present in the population. Figure 3 shows how coding precision varies across the correlation range continuum. The worst case scenario in terms of precision is when the correlation range matches the tuning curve width. It has been shown previously how signal and noise correlations with the same sign, as is the case with localised noise correlations, degrade coding precision (Latham and Nirenberg, 2005; Averbeck et al., 2006). Our results can be seen as a logical extension of that principle: maximal degradation of precision occurs when the signal correlations (tuning curves) and noise correlations have not only the same sign, but the same extent and shape.

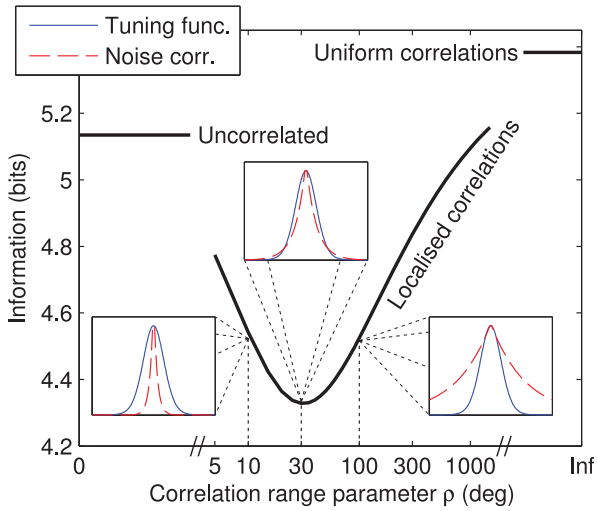

Figure 3: The effect of localised correlation range on precision. The main plot shows how, given a constant correlation strength, I_{Fisher} is dependent on the correlation range parameter ρ. I_{mut} is very similar and has been omitted for clarity. Introducing short-range correlations reduces coding precision relative to the uncorrelated noise case (top left), which is equivalent to $\rho = 0$. Precision decreases as the noise correlation range increases until it reaches a minimum, before increasing and converging towards the precision of the uniform correlation case as $\rho \to \infty$. The insets show the normalised tuning curve (solid line) and the correlation coefficients (i.e. a slice through the correlation matrix C; dashed line) for one neuron. Minimum information occurs where the noise correlation profile is most closely matched to the tuning curve i.e. where $\rho = \sigma_f = 30°$.

4 Which stimuli are most precisely represented by a neuron?

Tuning curves are commonly used to characterise the selectivity of neurons, but it is not always clear how they should be interpreted. Which stimuli does a neuron represent; which does it encode most precisely? Those at the peak of the tuning curve, where the activity of the neuron is most prominent? Or those at the steep flanks of the tuning curve, where the level of activity is most strongly modulated by small changes in the stimulus? To address these questions, Butts and Goldman (2006) calculated the stimulus-specific information for small populations of model neurons ($N \leq 4$) and showed that the stimuli that are best encoded by a neuron depend upon the level of variability. For neurons operating within a low noise regime (see Figure 4B), the best encoded stimuli lie at the flanks of the tuning curve ('flank coding'), while those operating in the high noise regime (see Figure 4D) have a single best encoded stimulus coinciding with the peak of the tuning curve ('peak coding'). This property is not unique to the SSI; the specific surprise also gives similar predictions. This is in contrast to Fisher information, which always predicts that the best encoded stimuli lie at the flanks of the tuning curve. This finding is potentially troublesome to the field as it suggests that the interpretation of the tuning curve depends on the measure used to determine the stimulus-specific precision. To investigate the extent of this issue and its implications for the analysis of experimental data, we use the SSI to further investigate how trial-to-trial variability, and also population size, affect which stimuli are most precisely encoded.

When determining the best-encoded stimuli for a neuron within a population, both the marginal SSI and singleton SSI are relevant. The meaning of the singleton and marginal SSI can be intuitively understood by considering a scenario where a population is constructed progressively by introducing one neuron at a time. The singleton SSI and marginal SSI are the contributions to the population SSI from the first and last neurons respectively. Because there is redundancy in the information encoded by each neuron, the actual informational contribution from a single neuron within a population lies somewhere between these bounds (e.g. the shaded regions in Figure 4).

4.1 The effect of variability and integration time in small populations

As reported by Butts and Goldman (2006), the stimuli most precisely represented by a neuron, according to the SSI, can lie at either the peak or flanks of the tuning curve, depending on the amount of noise present. Figure 4 illustrates this by showing the marginal and singleton SSI of the cricket cercal interneuron model for three different noise levels. The tuning curves and variability are shown

in Figure 4A.[3] Figures 4B–D show how the best encoded stimuli shift from the flanks of the tuning curve to the peak of the tuning curve as the noise level is increased. Both the singleton and marginal SSI undergo this transition, with the marginal SSI transitioning between peak and flank regimes at a higher noise level (thus, if the singleton SSI is known to be in the flank coding regime, we can infer that the marginal SSI, which is more difficult to calculate, is also in the flank regime). The difference in the peak/flank transition point is due to the fact that the marginal SSI relates to a four-neuron population, while the singleton SSI is based only upon a single neuron. The presence of more neurons in the population increases the coding precision, reducing the effective noise level of the code; this will be examined further in the following section on the effect of population size.

[3]Figure 4 is based on results obtained from our re-implementation of the cricket cercal interneuron model used by Butts and Goldman (2006), and its layout is based on that of Figure 3 from their article.

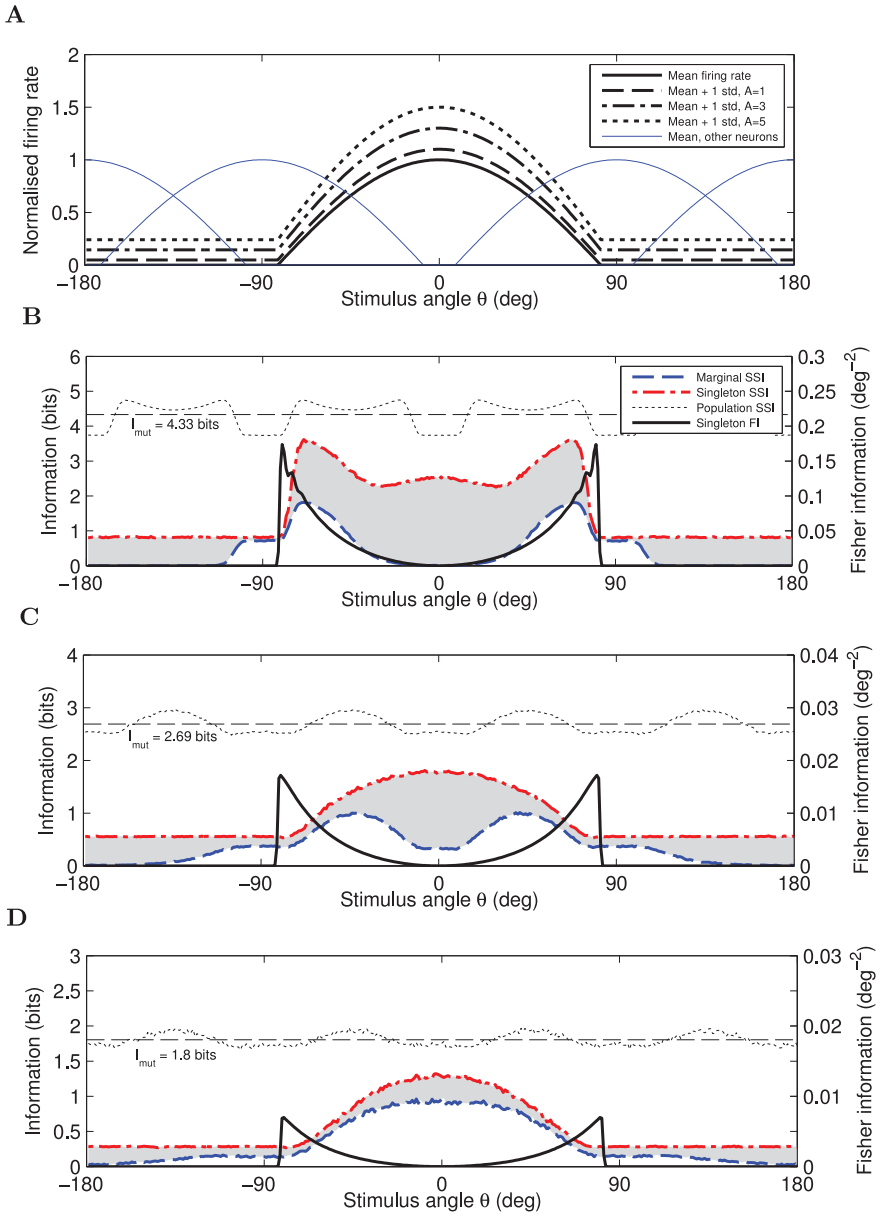

Figure 4: Model four-neuron population of cricket cercal interneurons as described by Theunissen and Miller (1991). This is a re-implementation, using our model, of the simulations shown in Butts and Goldman (2006), Figure 3.
(A) Tuning curves (mean responses) and three levels of trial to trial variability (illustrated as curves of mean + one standard deviation).
(B) Stimulus-specific information for the low noise case ($A = 1$). The dotted line shows the SSI of the whole population, while the shaded region shows the potential range of the information contribution of a single neuron, bounded from above by the singleton SSI and from below by the marginal SSI. Fisher information for a single neuron is shown for comparison. Both the singleton and marginal SSI indicate that the best-encoded stimuli lie at the flanks of the tuning curve.
(C) Intermediate noise case ($A = 3$). Here the singleton SSI is greatest at the peak of the tuning curve while the mSSI is greatest at the flanks.
(D) High noise case ($A = 5$). In this case both the singleton and marginal SSI are greatest at the peak of the tuning curve.

Under the flank coding regime, stimulus values can be read out by matching the firing rate of the neuron with the flanks of the tuning curve. Under the peak coding regime, it is not the precise level of activity, but the fact that the neuron's activity stands out from the background noise that conveys most of the information. This is a more robust, but coarser, indicator of the stimulus value—we know only that it is somewhere close to the neuron's preferred stimulus—and this is reflected in the lower absolute SSI values.

It is important to note that in all three cases the predictions of Fisher information and SSI differ; the shapes of the curves are different, and indicate that different stimuli are most precisely encoded. This is a consequence of the small number of neurons involved: the performance of an optimal decoder would not saturate the Cramér-Rao bound, so in this case the Fisher information is uninformative.

To further investigate the peak/flank transition in small populations, we used a four-neuron population model with circular Gaussian tuning curves and Fano factor variability, as described in sections 2.1–2.2. We calculated both the marginal SSI and marginal specific surprise (I_{sur}) for several levels of variability (F/τ), so that the predictions of these closely-related measures could be compared (see Figure 5A). Both measures have similar shapes and absolute values, and both exhibit a transition from the flank regime to the peak regime with increasing F/τ, although for I_{sur} the transition occurs at a higher value of F/τ i.e. its flank regime is more extensive. It is important to note that the quantity F/τ represents both noise level (Fano factor) and integration time; a transition from peak to flank regime could be caused by an decrease in the Fano factor, or equivalently by an increase in the time over which spikes are counted in each trial. At low levels of variability (probably unrealistically low in biological terms), SSI and I_{sur} are practically indistinguishable. Although the two measures differ more at higher F/τ values, their shapes are qualitatively similar. Fisher information differs from both SSI and I_{sur} in all four cases. While the shape of the singleton Fisher information, and hence its indication of best-encoded stimulus, remains identical across the four levels of variability, its absolute value varies by two orders of magnitude (not shown). Even in the lowest variability case ($F/\tau = 0.1\,\text{spikes/s}^2$), where all three measures indicate flank coding, the best encoded stimuli predicted by Fisher information and the Shannon information measures differ. Again, this is due to the small size of the population; four neurons is insufficient for the Fisher information to accurately predict the shape of the SSI or specific surprise.

Figure 5B shows the effect of altering the background firing rate upon the level of variability at which the peak/flank transition occurs. The shape of the marginal SSI is summarised by its peak to flank ratio (PFR); this is defined as the ratio of the SSI at the preferred stimulus (tuning curve peak) to its value at the maxima of the Fisher information (flanks of the tuning curve).[4] A PFR value of one indicates the point at which the SSI has three peaks of approximately equal value, and is therefore at the transition between the peak and flank regimes. PFR values of less than one correspond to the flank coding regime and values greater than one indicate the peak coding regime.

[4]The SSI_{flank} value does not necessarily correspond to the local maximum of the SSI, as the peaks of the SSI and Fisher information only become aligned as $N \to \infty$. The PFR can therefore be subject to fluctuations as parameter sweeps cause local SSI features to move across the stimulus value at which SSI_{flank} is calculated. The advantage of calculating the PFR in this way is that it is only necessary to compute the SSI at two predetermined points, as opposed to over the entire range of the stimulus variable.

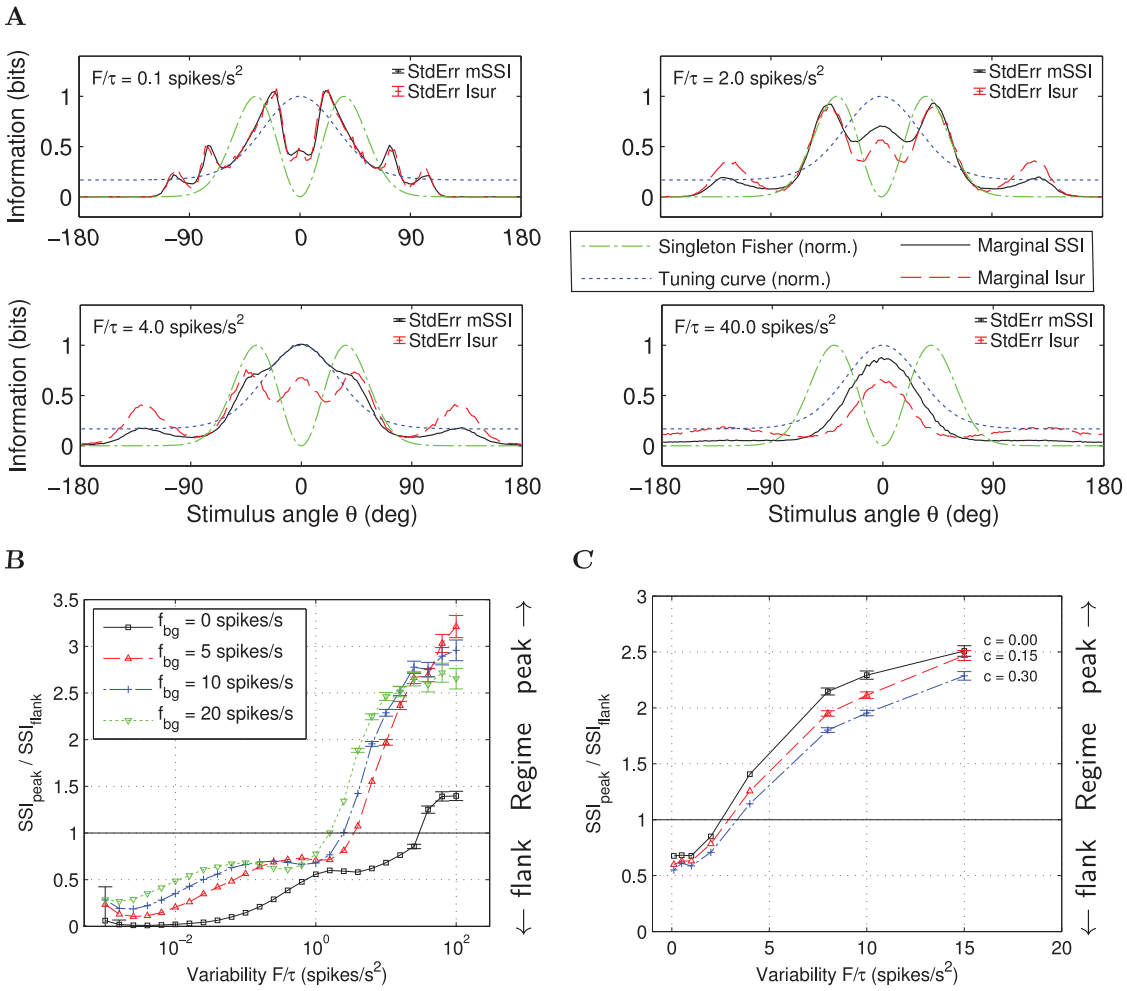

Figure 5: In small populations (here $N = 4$), SSI and specific surprise (I_{sur}) can be either unimodal or bimodal, depending on the level of trial-to-trial variability. (A) Marginal SSI and marginal I_{sur} for several levels of variability. SSI and I_{sur} are very similar at low F/τ values, but diverge to some extent as the variability increases. Both measures undergo a transition from bimodal (greatest on the flanks of the tuning curve) to unimodal (greatest at the peak of the tuning curve) as the variability increases; this occurs slightly earlier for the SSI. Parameters: independent variability, $f_{bg} = 10\,\text{spikes/s}$. Error bars show the worst case standard error for each measure.

(B) Marginal SSI peak to flank ratio (PFR) for several levels of background activity f_{bg}, with independent variability. Altering the level of background activity has a pronounced effect on the transition between low noise ($I_{peak}/I_{flank} < 1$) and high noise (> 1) regimes, with higher f_{bg} causing the transition to occur at a lower level of variability. For clarity, error bars have been omitted where the standard error is less than 0.02 bits.

(C) Marginal SSI PFR for various values of c, with uniform correlation structure. Uniform correlations improve coding precision, delaying the transition from flank to peak regime to greater levels of variability compared to the independent case. Parameters: $f_{bg} = 10\,\text{spikes/s}$. Error bars omitted when StdErr < 0.02 bits.

In the absence of background activity ($f_{bg} = 0$) the population remains within the flank coding regime up to $F/\tau \approx 30\,\text{spikes/s}^2$ (equivalent to $\tau \approx 33\,\text{ms}$ for $F = 1$). Introducing a small amount of background activity ($f_{bg} = 5\,\text{spikes/s}$, 10% of f_{max}) has a pronounced effect, with a transition to the peak coding regime now occurring at $F/\tau \approx 3.5\,\text{spikes/s}^2$ (equivalent to $\tau \approx 285\,\text{ms}, F = 1$). Further increases in baseline activity continue to shift the peak/flank transition to lower F/τ values. This is line with the findings of Wilke and Eurich (2002), who noted a rapid decrease in Fisher information at low levels of background activity. When $f_{bg} = 0$, neurons with preferred stimuli that differ from θ by more than about $3\sigma_f$ have essentially zero activity and hence zero variance. Increasing f_{bg} causes these neurons—approximately half of the population in this case—to fire at $f_{bg}\,\text{spikes/s}$ and to have a rate variance of $F f_{bg}\,(\text{spikes/s})^2$, thus substantially increasing the variability of the population as a whole.

Figure 5C illustrates the effect of uniform correlations in trial to trial variability upon the peak/flank transition. Uniform correlations improve coding precision and hence shift the regime transition to greater F/τ values relative to the uncorrelated case, although this is less pronounced than the shift caused by small levels of background activity.

4.2 The effect of population size

As demonstrated by Butts and Goldman and described in the preceding section, in very small populations the SSI can predict either flank or peak coding, depending on the amount of noise, noise correlation and the time over which spikes are counted. To date, this has not been investigated in populations larger than four neurons. Here we address the effect of population size upon the stimuli that are best encoded by a neuron. Do both peak and flank regimes occur in larger populations? How does population size affect the transition between regimes?

By using Monte Carlo integration (Metropolis and Ulam, 1949) to compute the SSI and specific surprise (see Appendix B.2), we were able to extend the analysis of Butts and Goldman to populations of up to 256 neurons. Such a sampling approach is necessary because the dimensionality of the response distribution is equal to the number of neurons, so any algorithm that exhaustively integrates over this distribution quickly becomes intractable as the population size increases. In order to validate our Monte Carlo approach we first replicated (see Figure 4) the results shown in Figure 3 of Butts and Goldman (2006), which describe the SSI for the cricket cercal interneurons, and were obtained via quadrature integration. Because of the similarity between the SSI and specific surprise, the unique advantages of the specific information (on which the SSI is based), and because the Monte Carlo estimate converges more rapidly for the SSI than for the specific surprise (due to its averaging over the stimulus ensemble), we leave aside the specific surprise and focus on the SSI for the remainder of the article.

We used the SSI to examine how the best-encoded stimulus of a neuron is affected by the size of the population that it exists within. Figure 6A shows the marginal

SSI for populations of various sizes; all curves are normalised to allow comparison. The mSSI shows a transition from the peak coding to the flank coding regime with increasing N, and the shape of the mSSI approaches the shape of the Fisher information as N becomes larger, although the units and absolute values of the two measures are different.

The transition between peak and flank regimes with N is shown in Figure 6B for several levels of variability, using the peak to flank ratio to summarise the shape of the marginal SSI. For very low levels of variability ($F/\tau = 0.1\,\text{spikes/s}^2$), the population operates in the flank regime at all population sizes, but at more realistic noise levels a transition occurs. The population size at which this happens depends on the level of variability: more noise means that a larger population size is required before the population moves into the flank coding regime.

For sufficiently large populations (approximately $N > 50$), variability no longer determines the coding regime and no (qualitative) discrepancy exists between the measures; both predict that neurons operate in the flank coding regime (for τ in the range $[10, 30]\,\text{ms}$ given F in the range [1,3]). Population size, along with trial-to-trial variability, is therefore an important determinant of the coding properties of individual neurons within a population.

Figures 6C and 6D show the effect of correlations. In line with other findings, uniform correlations increase precision and hence drive the population towards the flank coding regime, while localised correlations have the opposite effect. Figure 6D shows the effect on the PFR and transition point; localised correlations decrease the PFR and shift the peak/flank transition to lower N, while localised correlations have the opposite effect. The effect of localised transitions on the PFR is greatest at moderate population sizes around the regime transition, while uniform correlations have the greatest effect in small populations. To understand this difference, recall that the effect of localised correlations on precision is negligible in very small populations and increases with population size (see Figure 2E), whereas the effect of uniform correlations does not vary with population size (see Figure 2C).

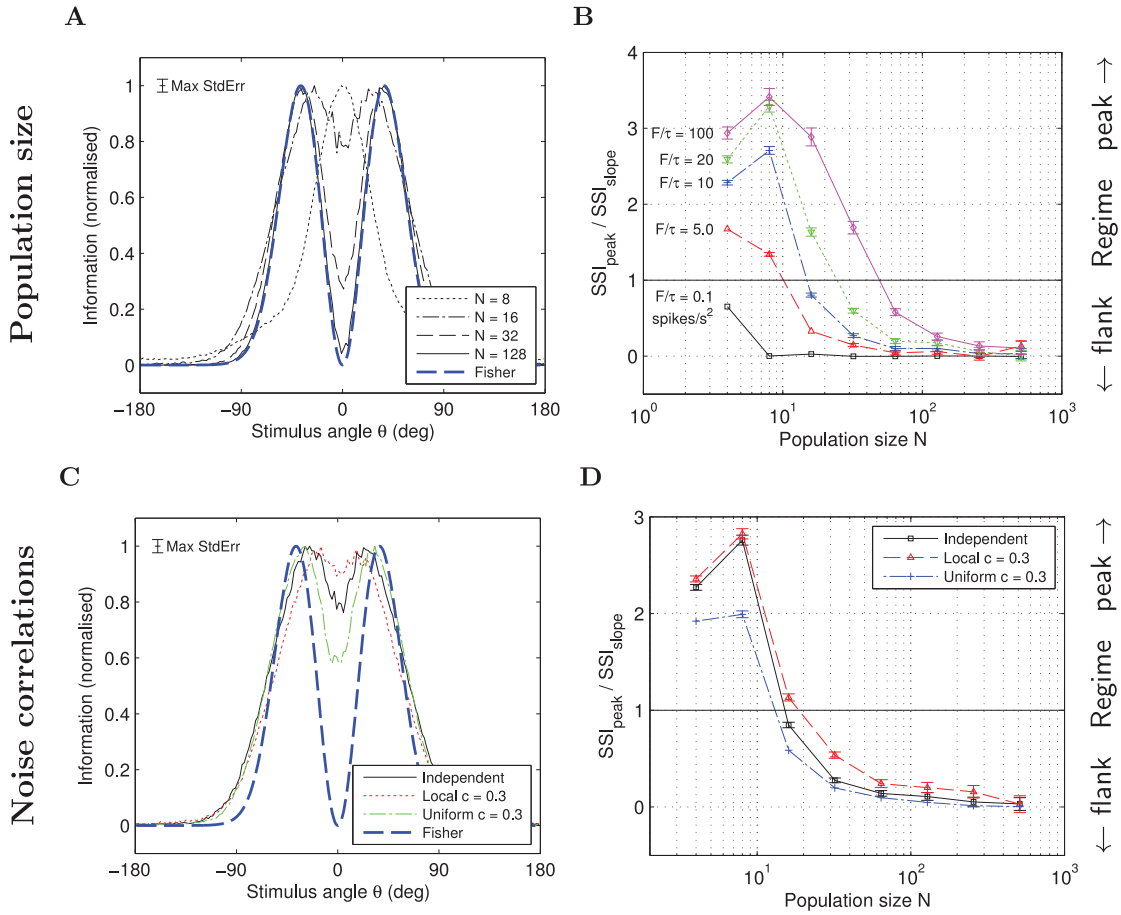

Figure 6: Marginal SSI and marginal SSI peak to flank ratio (PFR) in populations of various sizes. The stimulus value that is most precisely encoded by a neuron varies with population size. Maximum information occurs at the flanks of the tuning curve in large populations, but can occur at the peak or flanks in small populations, depending on the level of variability.
(A) Marginal SSI for various population sizes with independent variability. This plot illustrates the transition of greatest SSI from peak to flank of the tuning curve, and towards the shape of the singleton Fisher (heavy dashed line). Parameters: $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Error bar: worst case across all N, θ.
(B) PFR versus N for various levels of independent variability. The $F/\tau = 10\,\text{spikes/s}^2$ case corresponds to the SSI curves in A. Increasing variability delays the transition ($SSI_{peak}/SSI_{flank} = 1$) to greater population sizes. Parameters: $f_{bg} = 10\,\text{spikes/s}$. Error bars $< 0.02\,\text{bits}$ omitted.
(C) The effect of correlated variability on marginal SSI; localised correlations bring the neuron closer to the peak regime, while uniform correlations have the opposite effect. Singleton Fisher information shown for comparison; at this population size the mSSI has not yet converged to the shape of the Fisher information. Parameters: $N = 16$, $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Error bar: worst case across all N, c.
(D) PFR curves showing the effect of correlated variability. Localised correlations increase PFR, corresponding to reduced coding precision, while uniform correlations have the opposite effect. Parameters: $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Error bars $< 0.02\,\text{bits}$ omitted.

4.3 $\mathbf{SSI}_{Fisher}$

As described in section 1.4, I_{Fisher} allows us to make quantitative comparisons between Fisher information and Shannon mutual information when considering overall coding precision, but when dealing with stimulus-specific precision only qualitative comparisons have previously been possible. Qualitatively, Figure 6C suggests that the shape of the marginal SSI converges towards the shape of the singleton Fisher information as the population size goes to infinity. To allow this convergence to be investigated quantitatively (i.e. using the same units), we propose a new measure: SSI_{Fisher}. SSI_{Fisher} is a stimulus-specific decomposition of I_{Fisher}; more specifically it is the SSI of an optimal Gaussian-distributed estimator that saturates the Cramér-Rao bound (a formal definition is given in Appendix A.2). SSI_{Fisher} is an approximation of the SSI, in the same way that I_{Fisher} is an approximation of the mutual information. Here we consider the marginal SSI_{Fisher} (mSSI_{Fisher}), which is calculated in the same way as the marginal SSI, but is based upon SSI_{Fisher} rather than the SSI itself.

Figure 7A shows mSSI, mSSI_{Fisher}, and Fisher information together, for several population sizes. Fisher information is shown on a separate scale for ease of comparison and the scales are adjusted such that the maximum of $\mathrm{SSI}/\mathrm{SSI}_{Fisher}$ is level with the maximum Fisher information. It can be seen that the three curves converge with increasing N; in the case of SSI and SSI_{Fisher} this convergence is to the same absolute value, which equals Fisher information up to a multiplicative constant. The SSI_{Fisher}, like the SSI, undergoes a peak to flank transition with increasing N. Interestingly, the transition occurs later in SSI_{Fisher} than in the SSI itself, which is surprising as SSI_{Fisher} is derived from Fisher information, which relates to an upper bound on coding precision.

The convergence of mSSI and mSSI_{Fisher} roughly parallels that of mutual information and I_{Fisher}, but the latter converge more quickly. Figures 7B and 7C show the difference between Shannon and Fisher information based measures, as proportion of the Shannon information, for stimulus-specific (SSI, SSI_{Fisher}) and overall (MI, I_{Fisher}) quantities respectively. It can be seen that convergence occurs at approximately the same rate for both sets of measures, and that the relationships between the four variability cases are similar on both plots. Note that the scales on the two plots are different; the proportional difference between MI and I_{Fisher} is less than that between SSI and SSI_{Fisher}. This is due to differences in what is being measured: the marginal measures compared by ΔmSSI relate to the rate of change of overall information with respect to N, rather than the absolute value.

4.4 Summary

For large populations (more than around 50 neurons, for integration times down to around 10–30 ms) both Fisher information and the marginal SSI indicate that neurons provide information primarily about stimuli at the flanks of their tuning

curves. Even for large populations, there is some difference between the shapes of Fisher information and marginal SSI, but this diminishes as the population size increases. Smaller populations, however, can operate in either the flank coding regime or a peak coding regime where neurons convey most information about stimuli at the peaks of their tuning curves. Here, the regime depends upon the level of trial-to-trial variability (noise), integration time window, the amount and structure of noise correlations, and the population size. Increased noise, the presence of localised noise correlations, and reduced population size all drive the system towards the peak coding regime. Conversely, decreased noise, uniform noise correlations, and larger population sizes have the opposite effect, moving the population towards the flank coding regime. For small, noisy populations—populations operating in the peak coding regime—Fisher information gives a misleading indication of which stimuli are best represented by a neuron. This discrepancy between best-encoded stimulus predictions (mSSI versus Fisher) reflects the divergence of overall coding precision (I_{mut} versus I_{Fisher}) in small populations described in section 3.

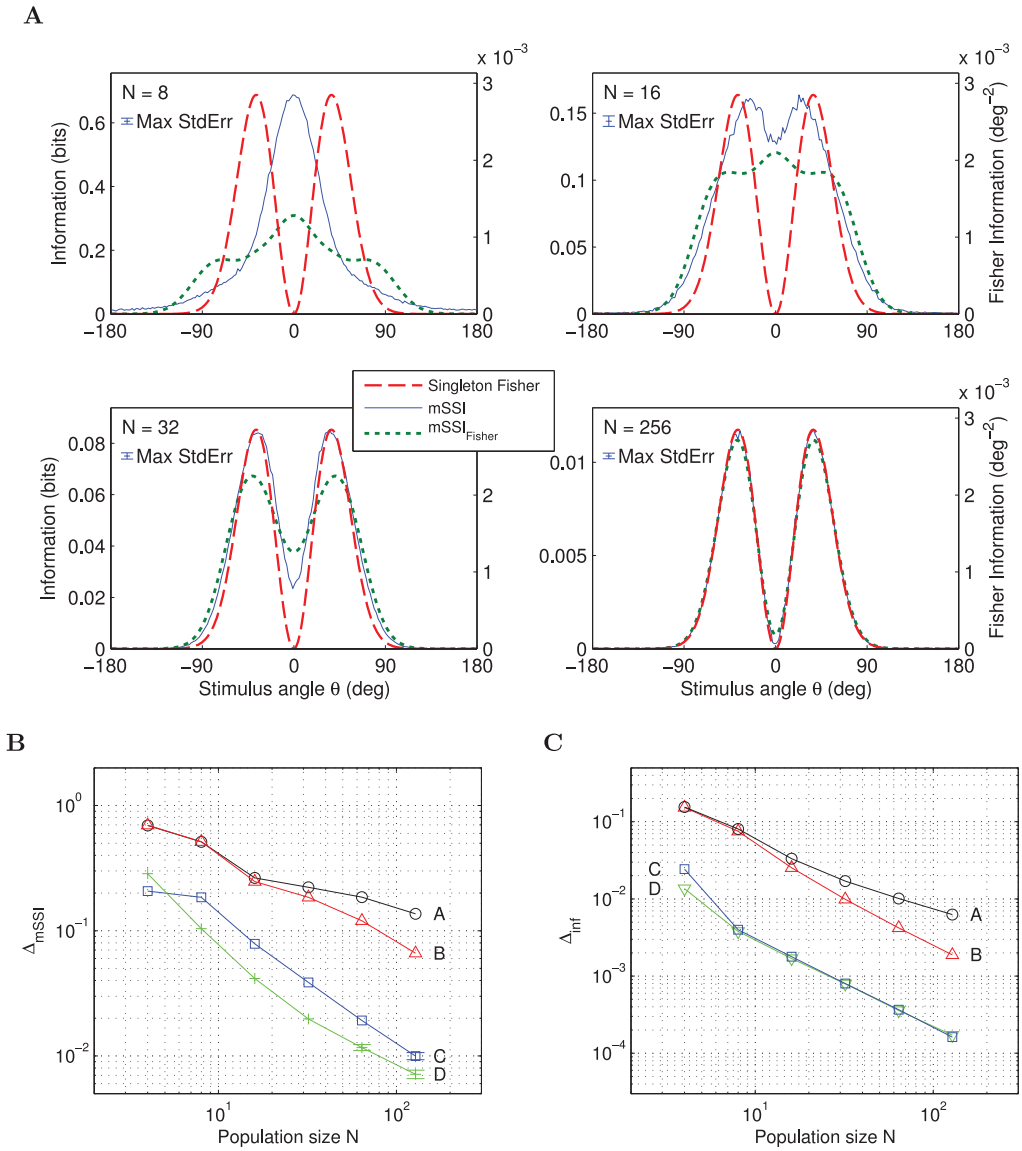

Figure 7: Marginal SSI converges towards marginal SSI_{Fisher} as population size increases.

(A) Marginal SSI, marginal SSI_{Fisher} and singleton Fisher information for populations of different sizes. The scales of the y axes are adjusted so that the maximum value of mSSI, mSSI_{Fisher} is aligned with the maximum value of the Fisher information. Parameters: independent variability, $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Error bar: worst case across θ.

(B) Convergence of marginal SSI and marginal SSI_{Fisher} for various parameter values. The y axis quantity is defined as $\Delta_{mSSI} = \frac{\text{RMS}(mSSI - mSSI_{Fisher})}{\text{RMS}(mSSI)}$ where RMS denotes the root mean square. Case A: localised correlations, $c = 0.3$, $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Case B: independent variability, $F/\tau = 10\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Case C: independent variability, $F/\tau = 1\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$. Case D: independent variability, $F/\tau = 1\,\text{spikes/s}^2$, $f_{bg} = 0\,\text{spikes/s}$. Error bars $< 5\%$ relative error omitted.

(C) Convergence of I_{mut} and I_{Fisher} roughly parallels the convergence of mSSI and mSSI_{Fisher}. $\Delta_{inf} = \frac{|I_{mut} - I_{Fisher}|}{I_{mut}}$ Parameter values are the same as in (B) for each case. Error bars $< 5\%$ relative error omitted.

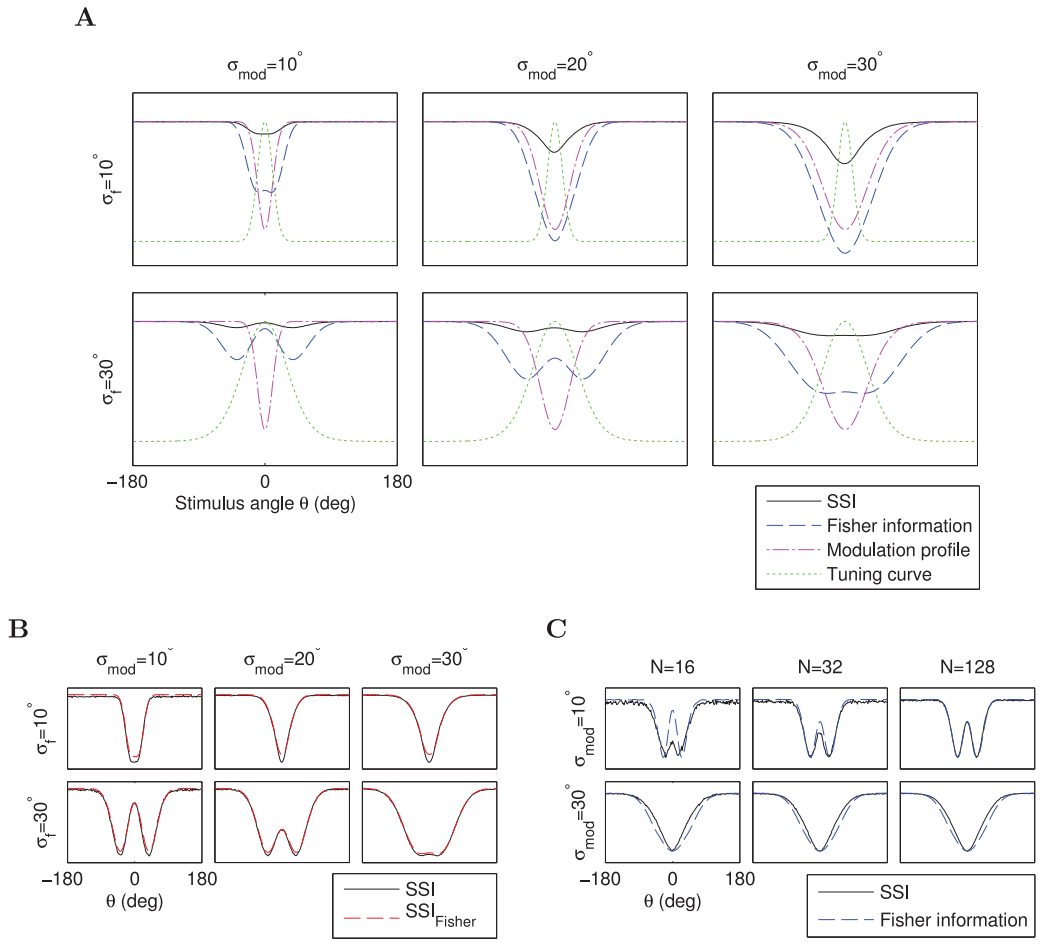

Figure 8: The effect of adaptation-like gain modulation on stimulus-specific coding precision.
(A) Population SSI and Fisher information for various combinations of tuning curve width (σ_f) and modulation profile width (σ_{mod}). SSI and Fisher are normalised so that their maximum values coincide on the plots. Similarly, zero for both measures coincide at the bottom of each plot. A single unmodulated tuning curve is shown, along with the modulation profile, to allow the widths to be visualised. The shape of both SSI and Fisher information is dependent upon the relative width of modulation and tuning curves. When $\sigma_{mod} < \sigma_f$ a double trough shape is produced. The closer the two widths, the less pronounced the central peak. For $\sigma_{mod} > \sigma_f$, there is a single trough with the least precise coding occurring at the centre of the modulation (the 'adapting' stimulus). Parameters: modulation depth 90%, independent variability, $N = 128$, $F/\tau = 5\,\text{spikes/s}^2$, $f_{bg} = 10\,\text{spikes/s}$.
(B) Population SSI and SSI_{Fisher} for the same cases shown in A. Both measures are plotted on the same scale. Although the values and shapes of SSI and Fisher shown in A differ, they indicate the same precision: SSI_{Fisher} and SSI are almost equal across all stimulus values.
(C) Trough shapes for population SSI and Fisher converge with increasing N, but the shapes of both functions are largely independent of N for large populations. Measures are normalised such that maximum and minimum values are aligned. Parameters as above, $\sigma_f = 20°$.

5 Gain Modulation

Gain modulation due to adaptation or attention-like processes is an often observed phenomenon in sensory neurons (see e.g. Wark et al., 2007). We applied the principles introduced above to examine the functional consequences of adaptation-like localised negative gain modulation, using the model described in section 2.3. Reducing the overall activity of the population is equivalent to reducing the signal to noise ratio, so the negative gain modulation causes a reduction in I_{mut} and I_{Fisher}. This is well understood, so our investigation focussed on the stimulus-specific precision: does adaptation affect the representation of the adapting stimulus itself or adjacent stimuli on the flanks of the affected tuning curves? Is the coding precision of the adapted stimulus increased, decreased or unchanged?

Population Fisher information and population SSI were calculated for 108 model populations with different combinations of population size, tuning curve width, modulation width and modulation depth. The shape of both Fisher information and SSI for the population depends mainly upon the relative width of the tuning curves and modulation profile; the function defining the height of the tuning curve peaks (see Figure 8). When the modulation profile is narrower than the tuning curves, both Fisher information and SSI have a double trough shape, with the coding precision of stimuli adjacent to the adapting stimulus reduced, while the representation of the adapting stimulus itself remains relatively unaffected (see Figure 8A, bottom left panel).[5] Hol and Treue (2001) observed a similar effect in a human psychophysical study. They observed that adaptation had no effect on the discrimination threshold at the adapting stimulus, but increased the threshold for neighbouring stimuli on both sides.

As the modulation width is increased relative to the tuning curve width, the representation of the adapting stimulus becomes progressively less precise relative to neighbouring stimuli (bottom centre panel, in this example $\sigma_{mod} = \sigma_f$). Ultimately, as the modulation width is further increased beyond the tuning curve width, the adapting stimulus becomes the least precisely represented stimulus, with coding precision increasing monotonically with distance from the adapter (top right panel).

[5]The extreme case for this scenario is an infinitesimally narrow modulation profile. In this case, the gain modulation would only affect a single neuron and the singleton Fisher information of this cell would effectively be subtracted from the population Fisher information, leading to the double trough shape with the representation precision of the adapting stimulus unchanged and that of neighbouring stimuli (those lying on the slopes of the modulated tuning curve) reduced. As the modulation width is increased, more neurons are affected and the troughs and central peaks begin to cancel out, eventually resulting in a single trough with the maximum reduction in coding precision occurring at the adapting stimulus itself.

Although the shapes and values of the population Fisher information and SSI in Figure 8A are different, the coding precision that they imply is essentially the same. Figure 8C shows the population SSI and SSI_{Fisher} for the same six cases. In all cases the SSI and SSI_{Fisher} are very similar. This is to be expected given the size of the population ($N = 128$); with this number of neurons the performance of the code is close to saturating the Cramér-Rao bound, hence the two measures should be close to equal. The slight difference between SSI and SSI_{Fisher} in the immediate neighbourhood of the adapting stimulus (the bottom of the trough) may be due to low response firing rates, which locally increase the signal to noise level and delay convergence of the two measures.

Figure 8C illustrates the convergence of the shape of the population SSI and Fisher information as the population size is increased. In the same way that singleton Fisher information is usually in agreement with the marginal SSI, we find here that population Fisher information in most cases predicts the same pattern of stimulus-specific precision as the population SSI. For large populations ($N > 32$) the shapes of the two measures are very similar. Differences are observed only within a restricted domain where N is small and the modulation width is narrow relative to the tuning curve width. In these cases Fisher information overestimates, relative to the SSI, the representation precision for the adapting stimulus.

6 Discussion

Information theory provides a powerful and general set of tools for assessing the precision of neural codes, but information theoretic measures are difficult to compute for experimentally-characterised populations due to the large number of observations required. Fisher information is an alternative statistical measure of precision that is generally easier to compute, but specifies an upper bound on coding precision that is only achieved in infinite populations. Brunel and Nadal (1998) showed that I_{Fisher}, an information theoretic measure derived from Fisher information, could provide an estimate of the mutual information I_{mut} in infinite populations (given certain conditions). However, how these two measures are related in finite populations had not previously been investigated. By numerically simulating neural populations of various sizes and levels of variability, we found that the mutual information is well approximated by I_{Fisher} (3.5% error) in populations with more than approximately 50 neurons, even with high variability and small time windows (e.g. $F = 3$ and $\tau = 30\,\text{ms}$). For populations with fewer neurons, I_{Fisher} tends to overestimate the mutual information and this disparity is greater for smaller populations. Increasing the amount of trial-to-trial variability (noise), or reducing the time window over which spikes are counted, increases the difference between I_{mut} and I_{Fisher}, but does not change the rate at which the two measures converge as a function of population size. Noise correlations slightly increase the difference between the two measures and delay their convergence, but these effects are small in comparison to those of population size or noise level.

We next addressed a related question: which stimuli are best encoded by a neuron operating within a population? Those that elicit the maximum response, corresponding to the peak of the tuning curve, or those coinciding with the flanks? We compared the predictions of Fisher information to those of the marginal SSI, a stimulus-specific decomposition of mutual information. Butts and Goldman (2006) found that in very small populations the most precisely encoded stimulus indicated by the marginal SSI was dependent on the level of variability, and sometimes conflicted with the predictions of Fisher information. Neurons operating in the peak coding regime have also been found experimentally, using the SSI to analyse single neurons (Montgomery and Wehr, 2010). Using a novel Monte Carlo approach to computing the SSI, we extended the analysis of Butts and Goldman to populations of up to 256 neurons. We found that the shape of the mSSI converges towards that of the Fisher information as the population size increases and, consequently, both measures predict the same best-encoded stimulus in large populations. Discrepancies occur only within a restricted domain of small populations (approximately $N < 50$) combined with high levels of trial-to-trial variability or short integration times. Under these conditions, the mSSI indicates peak coding, whereas Fisher information, as in all cases, indicates flank coding. Outside this limited domain, both measures indicate that neurons operate in the flank coding regime. Decreasing variability, increasing integration time, and uniform noise correlations drive the system towards the flank coding regime, while localised correlations have the opposite effect. This dependence upon integration time means that what stimulus is best encoded by individual neurons is a dynamical process: neurons will operate in the peak coding regime immediately following stimulus presentation and transition to flank coding as time progresses.

As with any modelling study, our analysis has a number of limitations. Perhaps most importantly, the results are specific to fine discrimination tasks, as Fisher information is defined as a very local measure of precision around a particular stimulus value. This limitation also applies to the information theoretic measures, since the stimulus ensemble was constructed in such a way as to simulate a reconstruction or fine discrimination task in order for the SSI and I_{mut} to be comparable with Fisher information and I_{Fisher}. For detection tasks, and probably also for coarse discrimination tasks, neurons best encode stimuli at the peak of their tuning curves. Fisher information is not applicable to these tasks, but other measures, such as Chernoff distance (Kang et al., 2004), can be used to estimate the mutual information as a function of the discrimination 'coarseness' (the distance between stimuli). We also assume that information is carried by a rate code; in cases where this assumption does not hold, the tuning curves and rate variability do not necessarily determine the best-encoded stimuli, as additional information about other stimuli may be conveyed by spike timing. In addition, all models were based on broadly-tuned neurons; we did not investigate how tuning curve width contributes towards determining the coding regime.

Some other studies addressing the validity of Fisher information have asked: what are the properties of population codes that are optimal in terms of Fisher

information? Tuning curves optimised to give maximal Fisher information would not resemble those observed experimentally (Bethge et al., 2002). If the tuning curves are constrained to be bell-shaped, maximising the Fisher information of a population means that tuning curve width is dependent upon population size, and is narrow in large populations. This is due to the fact that Fisher information increases as the tuning function width is decreased, up to the point where the overlap of neighbouring tuning curves is insufficient to give full coverage of the stimulus space (Berens et al., 2011). These studies also found that Fisher-optimal population codes are often sub-optimal in terms of other measures. Our work does not contradict the findings of these studies, but addresses a separate question: when can Fisher information be used to assess the precision of population codes that have been characterised experimentally? Our model neurons are broadly-tuned, in line with experimental findings (see e.g. Clifford, 2002), and the width does not vary with population size. Whilst Fisher information appears to be a poor tool for assessing the optimality of population codes, our results suggest that it is a valid measure of discrimination precision, albeit with limitations.

Our findings have two main implications for the experimental characterisation of neurons. Firstly, Fisher information can be used to obtain approximations of both I_{mut} and mSSI for neurons within large populations. As such, it is a reliable indicator of both coding precision and best-encoded stimuli for discrimination or reconstruction. In cases where it appears that the population size is well above the $N \approx 50$ threshold (e.g. hundreds of cells), Fisher information can be safely used, given the limitations discussed above. Secondly, for smaller populations where the number of neurons is known or can be accurately estimated, it is feasible to compute the SSI (and even the marginal SSI) if the tuning curves, trial-to-trial variability, and pairwise correlations can be modelled. It is then possible to determine whether neurons are operating in the peak or flank coding regime. The question as to which coding regime(s) the brain operates in is an interesting one, and one that cannot yet be answered in most cases as it depends in part on the size of the population involved in the relevant computation, which is generally unknown.

There has been much interest in calculating Fisher information from experimental data, and there are several possible approaches to estimating it, depending on the data available. The simplest method of obtaining Fisher information is to compute it directly from the tabular conditional response distribution $p(\boldsymbol{r}|\theta)$ by numerically evaluating Equation 6 (as in e.g. Dean et al., 2005). Measuring $p(\boldsymbol{r}|\theta)$ directly is only feasible for single neurons or very small populations, so the population Fisher information can only be obtained by this method in the case of uncorrelated noise. Alternatively, it is possible to use experimental data to construct a model of tuning curves and variability, and then to compute Fisher information from the model (as in e.g. Durant et al., 2007). Independent noise is typically modelled as a Poisson or univariate Gaussian distribution and correlated noise by a multivariate Gaussian distribution. While the best-encoded stimuli in large populations can be identified by computing the singleton Fisher information, computing the population Fisher information under a Gaussian variability model requires knowledge of the stimulus-

dependent covariance matrix $Q(\theta)$. The measurement of $Q(\theta)$ represents the most challenging obstacle to computing the population Fisher information, as this requires many trials and simultaneous recording of multiple neurons. In addition, any inaccuracies will be amplified when $Q(\theta)$ is inverted to obtain $Q^{-1}(\theta)$ (see Equation 7). It is not yet clear what the best method of determining the covariance matrix is, or how many trials are required to measure $Q(\theta)$ with sufficient accuracy to obtain a reasonable estimate of the Fisher information; more work is required to establish the answers to these open questions. Additionally, the level of noise correlations present in the brain is a matter of active debate (Ecker et al., 2010); in cases where trial-to-trial variability is effectively uncorrelated, the process of calculating the population Fisher information is greatly simplified.

The problem of determining the covariance matrix can be avoided by using a decoding approach. This involves constructing a function that estimates the stimulus given single-trial response spike counts for each neuron in the population. The variance of this estimator $\hat{\theta}(\boldsymbol{r})$ over many trials can then be used to determine a lower bound on the Fisher information:

$$J(\theta) \geq \frac{1}{\mathrm{Var}(\hat{\theta}(\boldsymbol{r}))} \tag{23}$$

This approach has been used in theoretical studies (e.g. Seriès et al., 2004; Beck et al., 2008; Chelaru and Dragoi, 2008). With this method, the most difficult part of the analysis is constructing an efficient estimator; this can be done via a number of machine learning techniques and the quantity of data required to train the estimator will depend upon the method used.

However the Fisher information is obtained, it tends to be much less difficult to calculate than information theoretic measures such as the SSI, in terms of both data requirements for experimentalists and computational complexity for numerical modellers. Although we have focussed upon the SSI, we have also shown that specific surprise gives similar predictions as to the best-encoded stimuli. Other stimulus-specific decompositions of the mutual information are possible, in particular the local information or stimulus information density proposed by Bezzi et al. (2002). Under the uniform stimulus distribution used in our model, the latter measure approaches the specific surprise (up to a multiplicative constant), so its predictions in the cases examined here are likely to be very close to those of specific surprise.

An important direction for future research is to examine how coding accuracy and best-encoded stimuli depend on the coarseness of discrimination. We have shown that large populations probably operate in the flank coding regime for fine discrimination tasks, and it is clear that the peak coding regime is relevant for very broad discrimination, where entirely separate groups of neurons are activated in response to the stimuli. What happens between these two edge cases has yet to be investigated. The stereotypical nature of most population code models points to further open questions. Most theoretical work to date has assumed that population codes are based on regular arrays of uniform unimodal tuning curves; what effect does heterogeneity of tuning curve width or shape have on the coding regime? The

peak and flank coding regimes discussed here are specific to bell-shaped tuning curves. For monotonic tuning functions, high Fisher information corresponds to steeply sloping regions of the curve; this is equivalent to the flank coding regime. It is not clear what the best encoded stimulus is for monotonic tuning curves and tasks other than fine discrimination. In future work, it is our intention to extend the current study to monotonic tuning curves and other behavioural tasks besides fine discrimination.

7 Conclusions

We have shown that it is feasible to compute the SSI for populations consisting of hundreds of neurons via Monte Carlo integration. This means that the SSI has the potential to be used to analyse experimental results at the population level, as well as for single neurons. Although the full set of results presented in this article represents considerable computational effort, calculating the SSI for a single empirically-determined model, even with 200 neurons, requires at most a day or so of computing time on a modern desktop computer.

The predictions of the SSI and Fisher information converge rapidly as a function of the number of neurons in the population. The exact pattern of convergence depends on the parameters of the chosen model. However, we found that for populations larger than around 50 neurons, they are qualitatively identical, even with high levels of variability and/or short integration times. The stimuli that are best encoded are then always those falling at the flanks of the tuning curves. This indicates that there is no need to go to very large population sizes for the SSI and the Fisher information to lead to similar predictions. Marginal SSI and Fisher information differ only over a restricted domain (small temporal windows, small populations, high noise), which seems to roughly correspond to the range where Fisher Information 'fails' (i.e. where the Cramér-Rao Bound is not saturated by maximum-likelihood or other optimal decoders (Bethge et al., 2002; Xie, 2002)).

Correlations in the trial to trial variability (noise correlations) have a relatively minor effect upon the convergence of information theoretic and Fisher-based measures. The 50-neuron guideline threshold for qualitative convergence holds in the presence of biologically realistic levels of correlation, whether uniform or localised.

Acknowledgments

The authors would like to thank Dan Butts, Mark Goldman and Matthias Bethge for useful discussions during the course of this project and comments on an earlier version of this manuscript.

This research was supported by funding from the Engineering and Physical Sciences Research Council and the Medical Research Council of Great Britain.

This work has made use of the resources provided by the Edinburgh Compute and Data Facility (ECDF). (`http://www.ecdf.ed.ac.uk/`). The ECDF is partially supported by the eDIKT initiative (`http://www.edikt.org.uk`).

Appendix A Mathematical supplement

Appendix A.1 Justification for use of F/τ

The response spike counts are distributed as a multivariate Gaussian with mean $\tau \boldsymbol{f}(\theta)$ and covariance $Q(\theta)$:

$$\boldsymbol{r}(\theta) \sim \mathcal{N}\left[\tau \boldsymbol{f}(\theta), Q(\theta)\right]$$

Where the covariance matrix is defined as:

$$\begin{aligned} Q_{i,j}(\theta) &= F\left[\tau f_i(\theta)\right]^\alpha R_{i,j}\left[\tau f_j(\theta)\right]^\alpha \\ &= F\tau^{2\alpha} f_i(\theta)^\alpha R_{i,j} f_j(\theta)^\alpha \end{aligned}$$

Collecting the non-scalar terms as $P(\theta)$:

$$P_{i,j}(\theta) = f_i(\theta)^\alpha R_{i,j} f_j(\theta)^\alpha$$

gives the following expressions for $Q(\theta)$, its inverse, and its derivative with respect to θ:

$$\begin{aligned} Q(\theta) &= F\tau^{2\alpha}P(\theta) \\ Q(\theta)^{-1} &= \frac{1}{F\tau^{2\alpha}}P(\theta)^{-1} \\ Q'(\theta) &= F\tau^{2\alpha}P'(\theta) \end{aligned}$$

The Fisher information is given by (this is the same as Equation 7, but the integration time τ is stated explicitly rather than being included in the mean response term):

$$J(\theta) = \tau f'(\theta)^T Q(\theta)^{-1}\tau f'(\theta) + \frac{1}{2}\operatorname{Tr}\left[Q(\theta)^{-1}\, Q'(\theta)\, Q(\theta)^{-1}\, Q'(\theta)\right]$$

Separating out the scalar terms as above, we have:

$$\begin{aligned} J(\theta) &= \tau f'(\theta)^T \frac{1}{F\tau^{2\alpha}}P(\theta)^{-1}\,\tau f'(\theta) + \frac{1}{2}\operatorname{Tr}\left[\frac{1}{F\tau^{2\alpha}}\,P(\theta)^{-1}\,F\tau^{2\alpha}P'(\theta)\,\frac{1}{F\tau^{2\alpha}}P(\theta)^{-1}\,F\tau^{2\alpha}P'(\theta)\right] \\ &= \frac{\tau^{2-2\alpha}}{F} f'(\theta)^T P(\theta)^{-1} f'(\theta) + \frac{1}{2}\operatorname{Tr}\left[P(\theta)^{-1}\,P'(\theta)\,P(\theta)^{-1}\,P'(\theta)\right] \end{aligned}$$

Thus for Fano factor variability (i.e. when $\alpha = 0.5$), F and τ appear in the expression for Fisher information only in the ratio τ/F:

$$J(\theta) = \frac{\tau}{F} f'(\theta)^T P(\theta)^{-1} f'(\theta) + \frac{1}{2}\operatorname{Tr}\left[P(\theta)^{-1}\,P'(\theta)\,P(\theta)^{-1}\,P'(\theta)\right]$$

Appendix A.2 Stimulus-specific I_{Fisher}

The stimulus-specific I_{Fisher} (SSI_{Fisher}) is the SSI of an optimal Gaussian estimator $\hat{\theta}_{opt}(\boldsymbol{r})$ with variance equal to the Cramér-Rao bound:

$$I_{ssiF}(\theta) = \sum_{\hat{\theta}_{opt}(\boldsymbol{r})\in\Theta} p(\hat{\theta}_{opt}(\boldsymbol{r})|\theta) \left[\sum_{\theta\in\Theta} p(\theta|\hat{\theta}_{opt}(\boldsymbol{r})) \log p(\theta|\hat{\theta}_{opt}(\boldsymbol{r})) - p(\theta)\log p(\theta) \right]$$

$$\text{where} \quad p(\theta|\hat{\theta}_{opt}(\boldsymbol{r})) = \frac{p(\hat{\theta}_{opt}(\boldsymbol{r})|\theta)p(\theta)}{\sum_{\theta\in\Theta} p(\hat{\theta}_{opt}(\boldsymbol{r})|\theta)p(\theta)}$$

$$\text{and (due to the Cramér-Rao bound)} \quad p(\hat{\theta}_{opt}(\boldsymbol{r})|\theta) = \mathcal{N}(\theta, J(\theta)^{-1})$$

Appendix B Implementation details

Appendix B.1 Differential entropy and continuous MI

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \tag{24}$$

$$h(X) = -\int_X p(x)\log p(x)\, dx \tag{25}$$

Shannon entropy (Equation 24) can only be calculated for discrete variables. Differential entropy (Equation 25) is a generalisation of Shannon entropy to continuous-valued random variables, but unfortunately does not retain all of the useful properties of Shannon entropy. In particular, differential entropy is not invariant under a change of variables, such as a change in the units used to measure the stimulus. Also, while Shannon entropy is always positive, differential entropy can take negative values. However, mutual information computed using differential entropies (continuous mutual information) does not suffer from these problems and retains the properties of its discrete counterpart (Cover and Thomas, 2006). Since both the stimulus and response variables in our model are continuous, all the entropies discussed here in relation to our model are differential entropies. How these entropies were calculated in order to find the SSI is described in the following section.

Appendix B.2 Calculating the MI, SSI and I_{sur}

In all simulations, SSI and specific surprise were calculated simultaneously via Monte Carlo integration. The method for computing the SSI is given here as

an example; the MI and specific surprise are evaluated similarly. Referring to Equation 5, it can be seen that the SSI is an average over the entire N-dimensional response ensemble (the outer summation). Since the complexity of computing the average over the response ensemble grows exponentially with N, the calculation quickly becomes intractable as the population size increases. Monte Carlo integration enables us to avoid this problem by sampling at random from the response distribution, computing the value of the measure based on this sample, and averaging across all samples to find the final value. This process is repeated until the desired level of precision is reached.

The SSI is defined as:

$$I_{SSI}(\theta) = \sum_{r \in R} p(r|\theta) \left[\sum_{\theta \in \Theta} p(\theta|r) \log p(\theta|r) - p(\theta) \log p(\theta) \right] \tag{26}$$

To calculate the SSI for a given stimulus θ, we first sample a vector of neuronal responses $\boldsymbol{r}^k$ (where the superscript k is an index over Monte Carlo samples) from the conditional distribution.

$$\boldsymbol{r}^k \sim p(\boldsymbol{r}|\theta) = \mathcal{N}[\tau \boldsymbol{f}(\theta), Q(\theta)] \tag{27}$$

We then calculate $p(\boldsymbol{r}|\theta')$ for many values of θ' regularly spaced across the entire stimulus space Θ; this is trivial since $p(\boldsymbol{r}|\theta)$ is known. We can then apply Bayes' theorem to find $p(\theta'|\boldsymbol{r})$:

$$p^k(\theta'|\boldsymbol{r}) = \frac{p(\boldsymbol{r^k}|\theta')p(\theta')}{\int_\Theta p(\boldsymbol{r^k}|\theta')p(\theta')\ d\theta'} \tag{28}$$

Where the integral $\int_\Theta p(\boldsymbol{r}|\theta')p(\theta')d\theta'$ is evaluated by numerical quadrature. We then calculate the specific information sample:

$$I^k_{SI}(\theta) = \int_\Theta p^k(\theta'|r) \log p^k(\theta'|r) - p(\theta') \log p(\theta')\ d\theta' \tag{29}$$

This sampling process is repeated many times, and the SSI is found by averaging over the samples:

$$I_{SSI}(\theta) = \frac{1}{n} \sum_{k=1}^{n} I^k_{SI}(\theta) \tag{30}$$

Where n is the number of MC samples. The estimate of the SSI is guaranteed to converge towards the true value as $n \to \infty$. The precision of the estimate was monitored by computing the standard deviation $s_{SSI}(\theta)$ of the MC samples. This allowed the standard error of the SSI estimate to be found using the equation for the standard error of the mean:

$$SE_{SSI}(\theta) = \frac{s_{SSI}(\theta)}{\sqrt{n}} \tag{31}$$

The standard error decreases as the number of samples increases and the sampling process was halted when the standard error reached a predetermined threshold,

or when n reached a predetermined limit. In the figures, the final standard error of the MC estimates are indicated by the error bars.

The Matlab code used to obtain all the results in this article is available online from ModelDB:
`http://senselab.med.yale.edu/modeldb/ShowModel.asp?model=142990`

References

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

B. B. Averbeck, P. E. Latham, and A. Pouget. Neural correlations, population coding and computation. *Nat Rev Neurosci*, 7(5):358–366, 2006.

J. M. Beck, W. J. Ma, R. Kiani, T. Hanks, A. K. Churchland, J. Roitman, M. N. Shadlen, P. E. Latham, and A. Pouget. Probabilistic population codes for bayesian decision making. *Neuron*, 60(6):1142–52, Dec 2008.

P. Berens, A. S. Ecker, S. Gerwinn, A. S. Tolias, and M. Bethge. Reassessing optimal neural population codes with neurometric functions. *Proc Natl Acad Sci USA*, 108(11):4423–8, Mar 2011.

M. Bethge, D. Rotermund, and K. Pawelzik. Optimal short-term population coding: when Fisher information fails. *Neural Comput*, 14(10):2317–2351, 2002.

M. Bezzi, I. Samengo, S. Leutgeb, and S. J. Mizumori. Measuring information spatial densities. *Neural Comput*, 14(2):405–420, 2002.

A. Borst and F. E. Theunissen. Information theory and neural coding. *Nat Neurosci*, 2(11):947–957, 1999.

N. Brunel and J. Nadal. Mutual information, Fisher information, and population coding. *Neural Comput*, 10:1731–1757, 1998.

D. A. Butts. How much information is associated with a particular stimulus? *Network*, 14(2):177–187, 2003.

D. A. Butts and M. S. Goldman. Tuning curves, neuronal variability, and sensory coding. *PLoS Biol*, 4:639–646, 2006.

M. I. Chelaru and V. Dragoi. Efficient coding in heterogeneous neuronal populations. *Proc Natl Acad Sci U S A*, 105(42):16344–9, Oct 2008.

C. W. Clifford. Perceptual adaptation: motion parallels orientation. *Trends Cogn Sci*, 6(3):136–143, Mar 2002.

T. M. Cover and J. A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, N.J., 2nd ed edition, 2006.

I. Dean, N. S. Harper, and D. McAlpine. Neural population coding of sound level adapts to stimulus statistics. *Nature Neuroscience*, 8(12):1684–1689, 2005.

M. R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network*, 10(4):325–340, 1999.

S. Durant, C. W. G. Clifford, N. A. Crowder, N. S. C. Price, and M. R. Ibbotson. Characterizing contrast adaptation in a population of cat primary visual cortical neurons using Fisher information. *J Opt Soc Am A Opt Image Sci Vis*, 24(6): 1529–37, Jun 2007.

A. S. Ecker, P. Berens, G. A. Keliris, M. Bethge, N. K. Logothetis, and A. S. Tolias. Decorrelated neuronal firing in cortical microcircuits. *Science*, 327(5965):584–7, Jan 2010.

A. L. Fairhall, G. D. Lewen, W. Bialek, and R. R. de Ruyter Van Steveninck. Efficiency and ambiguity in an adaptive neural code. *Nature*, 412(6849):787–92, Aug 2001.

D. A. Gutnisky and V. Dragoi. Adaptive coding of visual information in neural populations. *Nature*, 452(7184):220–224, 2008.

N. S. Harper and D. McAlpine. Optimal neural population coding of an auditory spatial cue. *Nature*, 430(7000):682–686, 2004.

J. Heller, J. A. Hertz, T. W. Kjaer, and B. J. Richmond. Information flow and temporal coding in primate pattern vision. *J Comput Neurosci*, 2(3):175–93, Sep 1995.

K. Hol and S. Treue. Different populations of neurons contribute to the detection and discrimination of visual motion. *Vision Res*, 41(6):685–9, Mar 2001.

R. L. Jenison and R. A. Reale. Likelihood approaches to sensory coding in auditory cortex. *Network*, 14(1):83–102, Feb 2003.

K. Kang, R. Shapley, and H. Sompolinsky. Information tuning of populations of neurons in primary visual cortex. *J Neurosci*, 24(15):3726–3735, 2004.

P. E. Latham and S. Nirenberg. Synergy, redundancy, and independence in population codes, revisited. *J Neurosci*, 25(21):5195–5206, 2005.

C. K. Machens, T. Gollisch, O. Kolesnikova, and A. V. M. Herz. Testing the efficiency of sensory coding with optimal stimulus ensembles. *Neuron*, 47(3): 447–56, Aug 2005.

N. Metropolis and S. Ulam. The Monte Carlo method. *J Am Stat Assoc*, 44(247): 335–41, Sep 1949.

N. Montgomery and M. Wehr. Auditory cortical neurons convey maximal stimulus-specific information at their best frequency. *J Neurosci*, 30(40):13362–6, Oct 2010.

I. Nelken and G. Chechik. Information theory in auditory research. *Hear Res*, 229 (1-2):94–105, 2007.

M. W. Oram, P. Földiák, D. I. Perrett, and F. Sengpiel. The 'ideal homunculus': decoding neural population signals. *Trends Neurosci*, 21(6):259–65, Jun 1998.

S. Panzeri, R. S. Petersen, S. R. Schultz, M. Lebedev, and M. E. Diamond. The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron*, 29(3):769–77, Mar 2001.

M. A. Paradiso. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biol Cybern*, 58(1):35–49, 1988.

R. Quian Quiroga and S. Panzeri. Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci*, 10(3):173–85, Mar 2009.

T. D. Sanger. Neural population codes. *Curr Opin Neurobiol*, 13(2):238–249, 2003.

N. B. Sawtell and A. Williams. Transformations of electrosensory encoding associated with an adaptive filter. *J Neurosci*, 28(7):1598–612, Feb 2008.

P. Seriès, P. E. Latham, and A. Pouget. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat Neurosci*, 7 (10):1129–1135, 2004.

P. Seriès, A. A. Stocker, and E. P. Simoncelli. Is the homunculus "aware" of sensory adaptation? *Neural Comput*, 21(12):3271–304, Dec 2009.

H. S. Seung and H. Sompolinsky. Simple models for reading neuronal population codes. *Proc Natl Acad Sci USA*, 90(22):10749–53, Nov 1993.

M. Shamir and H. Sompolinsky. Nonlinear population codes. *Neural Comput*, 16 (6):1105–1136, 2004.

C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(7):379–423, 1948.

F. E. Theunissen and J. P. Miller. Representation of sensory information in the cricket cercal sensory system. II. Information theoretic calculation of system accuracy and optimal tuning-curve widths of four primary interneurons. *J Neurophysiol*, 66(5):1690–703, Nov 1991.

M. J. Tovée, E. T. Rolls, A. Treves, and R. P. Bellis. Information encoding and the responses of single neurons in the primate temporal visual cortex. *J Neurophysiol*, 70(2):640–54, Aug 1993.

J. Victor and K. Purpura. Metric-space analysis of spike trains: Theory, algorithms and application. *Network-Computation In Neural Systems*, 8(2):127–164, May 1997.

J. D. Victor. Spike train metrics. *Curr Opin Neurobiol*, 15(5):585–92, Oct 2005.

B. Wark, B. N. Lundstrom, and A. Fairhall. Sensory adaptation. *Curr Opin Neurobiol*, 17(4):423–9, Aug 2007.

S. D. Wilke and C. W. Eurich. Representational accuracy of stochastic neural populations. *Neural Comput*, 14(1):155–189, 2002.

X. Xie. Threshold behaviour of the maximum likelihood method in population decoding. *Network*, 13(4):447–456, 2002.