The role of stochasticity in an information-optimal neural population code

N G Stocks¹, A P Nikitin¹, M D McDonnell² and R P Morse³

¹ School of Engineering, University of Warwick, Coventry CV4 7AL, UK
² Institute for Telecommunications Research, University of South Australia, SA 5095, Australia
³ School of Life and Health Sciences, Aston University, Birmingham B4 7ET, UK

E-mail: n.g.stocks@warwick.ac.uk

Abstract. In this paper we consider the optimisation of Shannon mutual information (MI) in the context of two model neural systems. The first is a stochastic pooling network (population) of McCulloch-Pitts (MP) type neurons (logical threshold units) subject to stochastic forcing; the second is (in a rate coding paradigm) a population of neurons that each displays Poisson statistics (the so called ‘Poisson neuron’). The mutual information is optimised as a function of a parameter that characterises the ‘noise level’—in the MP array this parameter is the standard deviation of the noise; in the population of Poisson neurons it is the window length used to determine the spike count. In both systems we find that the emergent neural architecture and, hence, code that maximises the MI is strongly influenced by the noise level. Low noise levels leads to a heterogeneous distribution of neural parameters (diversity), whereas, medium to high noise levels result in the clustering of neural parameters into distinct groups that can be interpreted as subpopulations. In both cases the number of subpopulations increases with a decrease in noise level. Our results suggest that subpopulations are a generic feature of an information optimal neural population.

1. Introduction

It is well known in neuroscience that repeated presentation of identical stimuli to neurons (nerve cells) result in variable timings of resultant neural action potentials (voltage spikes). Thus neuronal responses are often referred to as ‘noisy.’ This variability arises due to noise sources that occur at all stages of neural processing e.g. from noise in the sensory receptors at the sensory periphery to ‘network’ and synaptic noise in the cortex [1]. Consequently, it is an unavoidable fact that neural coding and processing is dominated by stochasticity. However, what is still unclear is what impact the noise has had on the evolution of efficient neural coding and processing strategies.

In this study we investigate this issue in the context of neural coding. We consider the optimisation of Shannon Mutual information (MI) [2, 3] in neural populations as a function of the noise level. Two systems are studied; the first is an array of McCulloch-Pitts neurons and the second is a population of Poisson neurons. Both can be classified as stochastic pooling networks [4]. The objective of the study is to establish the neural parameters that maximise the MI. This, in turn, enables us to uncover the optimal neural architecture and hence gives an insight into the form an optimal neural code may take. It also enables us to investigate how the...
emergent architecture/codes depend on the level of stochasticity. As we will demonstrate, noise appears to be a significant organising influence behind an optimal neural structure.

The use of MI as an organising principle of neural coding was first proposed by Linsker in 1988 in the context of the visual system [5]. This is the so-called infomax principle. The idea is that neural systems organise themselves (i.e. their parameters) in such a way as to maximise the MI between their inputs and outputs subject to some constraints (such as energy, connectivity etc). The infomax principle is closely related to earlier work by Barlow [6] who proposed redundancy reduction as an organisational rule. More recently, the infomax principle has been extended by Bell and Sejnowski to independent components analysis [7]. The infomax principle has also recently been used to derive optimal codes for auditory coding that replicate closely the ‘revcor’ kernels observed in physiological experiments [8]. Consequently, there is growing evidence that the infomax principle is a useful paradigm for explaining neurophysiological structure and function. However, it is unlikely that such a principle can explain the structure of the brain as a whole because one would expect that the structure is related to function - and it is less than obvious that all brain function (e.g. detection, processing, computation, cognition, memory recall etc) can all be described by an optimisation of MI.

Nevertheless, where one might expect the infomax principle to be at work is at the sensory periphery. Information theory was originally developed to quantify the ability of systems to code and transmit information, which appears to be the same task as that faced by biological sensory systems. Furthermore, it is not unreasonable to assume that this information is coded in an efficient manner, especially with respect to metabolic energy. Inefficient coding is likely to result in a greater metabolic deficit and hence a survival disadvantage. Consequently, natural selection would be expected to deselect such strategies. These considerations lead us to postulate that information is coded using an infomax principle that is probably constrained by energy.

2. Optimization of neural systems
2.1. Population of McCulloch-Pitts neurons

We first consider the system shown in Fig. 1. Each neuron is connected to a common random

![Figure 1](image-url)

Figure 1. Array (population) of McCulloch-Pitts neurons. Each of the $K$ neurons, with threshold values $\theta_n$, $n = 1, ..., K$, is subject to independent additive noise, $\eta_n$, $n = 1, ..., K$. The common input to each neuron is a random signal, $X$, which in the results presented here is assumed to be Gaussian. Each neuron thresholds its input to produce a binary signal, $y_n$. The overall output is the sum of each $y_n$. 
signal, $x$, but has an independent noise source $\eta_n$. The response of each neuron is governed by

$$y_n = \Theta(x + \eta_n), \quad n = 1, \ldots, K,$$

(1)

where $\Theta(\cdot)$ is the Heaviside function and the response of the population, $Y$ is

$$Y = \sum_{n=1}^{K} y_n.$$  

(2)

If the input signal has probability density function $P_X(x)$, then the mutual information between the input and output is given by [9]

$$I(X;Y) = -\sum_{i=0}^{K} P_Y(i) \log_2 P_Y(i) - \left( -\int_{-\infty}^{\infty} P_X(x) \sum_{i=0}^{K} P_{Y|X}(i|x) \log_2 P_{Y|X}(i|x) dx \right),$$

(3)

where

$$P_Y(n) = \int_{-\infty}^{\infty} P_{Y|X}(n|x) P_X(x) dx.$$

Therefore for a given signal distribution, the mutual information depends entirely on the conditional probabilities, $P_{Y|X}(n|x)$. If the noise distribution is specified, the only free variables are the population size, $K$, and the threshold values. Let $\hat{P}_n$ be the probability of neuron $n$ ‘spiking’ in response to signal value $x$. Then

$$\hat{P}_n = \int_{\theta_n-x}^{\infty} f_{\eta}(\eta) d\eta = 1 - F_{\eta}(\theta_n - x),$$

(4)

where $F_{\eta}$ is the cumulative distribution function of the noise and $n = 1, \ldots, K$.

Given a noise density and arbitrary threshold values, $\hat{P}_n$ can be calculated exactly for any value of $x$ from (4). These can then be used in an algorithm specified in [10] for calculating the conditional probabilities, $P_{Y|X}(n|x)$.

We now aim to find the threshold settings that maximize the mutual information as the input SNR varies. This goal can be formulated as the following nonlinear optimization problem:

$$\text{Find:} \quad \max_{\{\theta_n\}} I(X;Y)$$

subject to: $\quad \{\theta_n\} \in \mathbb{R}^K$.  

(5)

It is necessary to solve (5) numerically using standard unconstrained non-linear optimization methods such as the conjugate gradient method [11]. However, note that the objective function is not convex, and there exist a number of local optima. This problem can be overcome by employing random search techniques such as simulated annealing. We present here in Figs. 2-5 results for the optimal thresholds obtained by solving Problem (5) for Gaussian signal and Gaussian noise, as a function of $\sigma := \frac{\sigma_\eta}{\sigma_x}$, where $\sigma_\eta$ and $\sigma_x$ are the noise and signal standard deviation respectively; we have set $\sigma_x = 1$ for all results. In Figs. 2, 3, 4, 5 we have $K = 3, 4, 5, 15$ respectively [12].

It can be observed that the dependence of the optimal threshold values on the noise level is similar for all results. For large $\sigma$ it is optimal to set all the thresholds equal to the signal mean which is zero in this case. As the noise is reduced this optimal solution bifurcates, at a value $\sigma_1$, in favour of grouping all the thresholds at two distinct values. Further reduction in the noise level results in three distinct values becoming optimal. This pattern is repeated as the noise is
reduced further. In general it is observed that the number of distinct threshold values increases as $1 \to 2 \to 3 \to \ldots \to K$ as the noise level is reduced. It should be noted that the exact values of the noise at which the transitions occur (e.g. $\sigma_1$) depends on the size of the population size $K$. This problem is discussed in Chapter 8 of Ref. [12]. For the limit $K \to \infty$, it is observed that an asymptotic limit of $\sigma_1$ exists, but that its value is dependent on the distributions of the signal and noise. For example, for Gaussian signal and noise the limiting $\sigma_1$ is about 0.8, while for logistic signal and noise it is about 1.2. However, qualitatively, the pattern is always the same.

These results raise a number of interesting points. First, the structure of the optimal threshold distribution is discrete. Numerical experiments indicate that this bifurcational structure is not changed even in the limit $K \to \infty$. Consequently, the clustering of thresholds into distinct groups is not a consequence of the finite number of neurons but is rather a property of the system (channel) itself. Second, the existence of these distinct groups indicate an optimal coding strategy is to group neurons with similar thresholds together into subpopulations; the number of distinct subpopulations depends on the level of noise. Third, the optimal design solution is strongly dependent on the level of the noise. Indeed, this point is graphically highlighted in
Figure 6. Plot of $I(X, Y)$ against noise intensity, $\sigma$, for $K = 15$, the optimal thresholds (dashed line) and $\theta_i = 0 \ \forall \ i$ (solid line). The solid line shows the suprathreshold stochastic resonance (SSR) effect [9, 10]. Reproduced from Ref. [12].

In Fig. 6 the MI is plotted as a function of the noise level for two different cases i) for the optimal threshold values (dashed line) shown in Fig. 5 and ii) for the optimal threshold solution for $\sigma \geq \sigma_1$ (solid line); i.e. the thresholds are held constant at $\theta_i = 0 \ \forall \ i$ as the noise level is varied. The two curves coalesce at $\sigma_1$. It can be observed that while the solution $\theta_i = 0 \ \forall \ i$ is optimal for $\sigma \geq \sigma_1$ this solution performs very badly in the limit $\sigma \to 0$. Indeed, it achieves a MI of only 1-bit compared to the optimal solution at $\sigma = 0$ which achieves 4-bits. Consequently, the correct choice of neural parameters (and hence the resulting neural code) is seen to strongly depend on the level of noise.

An interesting consequence of the sub-optimal nature of the solution $\theta_i = 0 \ \forall \ i$ at small $\sigma$ is that it gives rise to a suprathreshold stochastic resonance [9, 10] i.e. the MI is maximised at a non-zero level of noise. Such counter-intuitive ‘noise-benefits’ rely on some aspect of a system being ‘sub-optimal’ [9, 12].

Some theoretical progress in calculating the optimal information can be made by utilising a connection between mutual information and Fisher information. Following [16, 18] we can write,

$$Y(x) = T(x) + \sqrt{V(x)} \xi$$  \hspace{1cm} \text{(6)}

where $T(x)$ is the mean output for a give input $x$ and $V(x)$ is the variance of the output conditioned on the input, and $\xi$ is a random variable with zero mean and unit variance. For not too large $V(x)/N^2$ we can write

$$I(X; Y) \approx H(x) - \frac{1}{2} \int_x f_x(x) \log_2 \left( \frac{2\pi e}{J(x)} \right) dx,$$  \hspace{1cm} \text{(7)}

where the Fisher information $J(x)$ is given by

$$J(x) \approx \frac{(dT(x)/dx)^2}{V(x)}.$$  \hspace{1cm} \text{(8)}
Given that the input entropy \( H(x) \) only depends on the statistics of the signal and not on the properties of the channel, maximisation of (7) can be seen to be equivalent to a minimisation of the function

\[
F = -\langle \log_2 Z \rangle_x,
\]

where \( Z = J(x) \) and the average, \( \langle . \rangle_x \) is taken over the input signal distribution. It is interesting to note that this looks similar to the minimisation of a Free Energy with a partition function \( Z \).

These expressions can be simplified by noting that the responses of the neurons, \( y_n(x) \), are (conditionally) independent. Consequently, the mean response and variance can be obtained by summing the individual means, \( T_n(x) \), and variances, \( \text{Var}(y_n|x) \) [12]. Hence,

\[
T(x) = \mathbb{E}(Y|x) = \sum_{n=1}^{N} T_n(x),
\]

and

\[
V(x) = \text{Var}(Y|x) = \sum_{n=1}^{N} \text{Var}(y_n|x) = \sum_{n=1}^{N} T_n(1 - T_n).
\]

The final expression for the variance comes from the fact that \( \hat{P}_n = T_n \) and the variance of each neuron is governed by Bernoulli statistics, which gives \( \text{Var}(y_n|x) = \hat{P}_n(1 - \hat{P}_n) \). Consequently, substituting these expressions into Eq. (8) yields

\[
Z = J(x) = \left( \sum_{n=1}^{N} \frac{dT_n(x)}{dx} \right)^2 \frac{1}{\sum_{n=1}^{N} T_n(1 - T_n)}.
\]

It is interesting to note that Eq. (12) only requires specification of the mean response \( T_n(x) \), i.e. the tuning curve of each neuron. Furthermore, \( T_n(x) \) is, itself, simply related to the cumulative noise distribution of each neuron as \( T_n(x) = 1 - F_\eta(\theta_n - x) \). Consequently, it is largely the cumulative noise distribution that ultimately specifies the information flow in these systems.

A particularly simple expression for the mutual information can be derived under the assumption that the noise has a logistic distribution. In this case the cumulative distribution function of the noise is \( F_\eta(\theta_n - x) = [1 + \exp(-\beta(\theta_n - x))]^{-1} \) where \( \beta \) is the inverse noise level. With this assumption it can be shown that \( T_n \), via \( F_\eta \), satisfies the differential equation associated with the logistic function, that is

\[
\frac{dT_n}{dx} = \beta T_n(1 - T_n).
\]

Noting that the expression for the Fisher information simplifies to \( J(x) = \beta^2 \sum_{n=1}^{N} T_n(1 - T_n) \), Eq. (9) can be written

\[
F(\{\theta_n\}, \beta) = - \left( \log_2 \sum \frac{\beta^2}{4} \left[ \cosh^2 \left( \frac{\beta}{2}(\theta_n - x) \right) \right]^{-1} \right). \tag{14}
\]

To obtain the optimal information it is now necessary to find the set \( \{\theta_n\} \) that minimises \( F(\{\theta_n\}, \beta) \) for a given \( \beta \). Although we have used a similar approach to find the location of the first bifurcation point for large \( N \) we have not been successful in describing the entire sequence of bifurcations. We speculate that further progress might be made by using a replica approach, but we have been unable to verify this.
2.2. Population of Poisson neurons
In this section we replace each MP neuron in Fig. 1 with a ‘Poisson’ neuron [13, 14]. Spike timing variability is often closely modelled by Poisson statistics and, hence, the Poisson neuron has become an archetypal model for neural rate coding. In this model the input signal $x$ is coded in the mean firing rate $\nu = g(x)$ where $g(x)$ is known as the tuning curve (or the stimulus-response curve, gain function or rate-level function). Pooling (summing) the outputs of Poisson neurons results in a Poisson process, so without loss of generality we can treat the tuning function $g(x)$ as the tuning function for the whole population. While several definitions of rate exist [15], following related studies [16, 17], here we assume the observable output when the mean rate is $\nu$ is the number of spikes, $k$, that occur in a time window $T$. The input $x$ is assumed to be a continuous variable, such as an external sensory stimulus.

To make connection with the studies presented in the preceding section we need to quantify the stochasticity in the system. Unlike the MP neuron, where we simply introduced additive noise and used the standard deviation of the noise as a measure of stochasticity, we cannot introduce noise in this fashion for a Poisson neuron. In a Poisson neuron the statistics of the action potentials (neural spikes) are defined through the Poisson distribution—see Eq. (16). Consequently, the level of stochasticity in some sense is fixed. However, in a rate coding paradigm the signal is coded in a spike count over a time window $T$, where it is further assumed that the rate is constant over this window. Consequently, the accuracy of the code will depend on the length of the window; the longer the window length the more accurate the estimate of the rate and hence, in principle, the more accurate our estimate of $x$.

We can quantify this accuracy in terms of a signal-to-noise ratio (SNR). The information received at the output about the input signal is based on a measurement of the output rate $z = k/T$, where $z$ is a random variable. The ‘signal’ can be identified as $E[z]$ and the uncertainty, or ‘noise’, is given by the variance, $\text{var}[z] = E[(z - E[z])^2]$. We can now define an SNR in the standard way as the ratio of the signal power to the noise power i.e. $\text{SNR} = E[z]^2/\text{var}[z]$. Noting that $E[z] = E[k]/T$ and $\text{var}[z] = \text{var}[k]/T^2$, and furthermore that for a Poisson process we have $\text{var}[k] = E[k]$, the SNR simplifies to $\text{SNR} = E[k]$. Given also we have $E[k] \sim T$ this implies SNR $\sim T$. Therefore, the SNR ratio improves proportionally with increasing $T$.

The $g(x)$ that maximizes MI has only previously been obtained in the limit $T \rightarrow \infty$ [16, 18]. Arguably, this limit is not relevant to a large number of biological sensory systems where it is well established that behavioural responses occur on timescales that imply short coding windows [14]. Here we present the optimal tuning curve for finite $T$ and outline a proof that the tuning curve is discrete; full details of the proof can be found in [19].

To derive the optimal tuning curve, we make use of a known result from the photonics literature. Both the generation of neural action potentials, and the emission of photons can be modeled as Poisson point processes [20].

A classical problem in communication theory is that of finding the signal distribution that maximizes the mutual information for a channel. The resultant optimal code is said to achieve channel capacity [2]. The optimal input distribution for the direct detection photon channel has been proven to be discrete [21, 22, 23]. Indeed, the discreteness of optimal input distributions is the norm, regardless of whether the output distribution is discrete or continuous [24, 25] and is a property of the channel and not the output measure—see [25, 26].

The discreteness of the optimal signal for the optical Poisson channel implies that the optimal stimulus for a Poisson neural system is also discrete. However, the signal distribution is not something that a neural channel can control directly. However, instead, it is plausible that a neural system may have been optimized by evolution so that the tuning curve discretises its input to match the theoretical optimal source distribution and, indeed, our results suggest this is the case.

We now proceed to formulate the problem of optimising the MI. The MI between the
(continuous) input random variable $x \in \mathcal{X} = [x_{\text{min}}, x_{\text{max}}]$, and the (discrete) output random variable, $k$, is

$$I(x; k) = \sum_{k=0}^{\infty} \int_{x \in \mathcal{X}} dx P_x(x) P[k|x] \log_2 \frac{P[k|x]}{P_k(k)},$$

(15)

where $P_k(k) = \int_{x \in \mathcal{X}} dx P_x(x) P[k|x]$. Here $P_x(x)$, $P_k(k)$ and $P[k|x]$ are the distributions of the stimulus, the response and the conditional distribution respectively.

For Poisson statistics, the conditional distribution is

$$Q[k|\nu] = \frac{[T\nu]^k}{k!} \exp(-T\nu), \quad k = 0, \ldots, \infty.$$  

(16)

The mean firing rate is restricted to $\nu \in [\nu_{\text{min}}, \nu_{\text{max}}]$, where the upper bound $\nu_{\text{max}}$ is due to physiological limits (metabolic and refractory), and we set $\nu_{\text{min}} = 0$. Later, we use the notation $N = T\nu_{\text{max}}$ to denote the maximum mean spike count.

The conversion of a signal follows the Markov chain $x \rightarrow \nu \rightarrow k$. We refer to $x \rightarrow \nu$ and $\nu \rightarrow k$ as separate ‘subchannels.’ To find the optimal channel, we maximize the MI by variation of the distribution $P_\nu(\nu)$ for given $P_x(x)$ and $Q[k|\nu]$. Since the distribution of $\nu$ is $P_\nu(\nu) = \int_{x \in \mathcal{X}} dx P_x(x) \delta(\nu - g(x))$, where $\delta(.)$ is the Dirac delta function, variation of $P_\nu(\nu)$ means variation of the tuning curve $g(x)$.

We have recently been able to prove the following theorem [19]: The mutual information in the neural channel, $x \rightarrow \nu \rightarrow k$, is maximized when the distribution $P_\nu(\nu)$ is discrete. The full proof can be found in [19] but, briefly, is as follows:

First we prove that $I(x; k) = I(\nu; k)$ using Theorem 5.2.8 of [27]. We then only need to consider the optimisation of the noisy neural subchannel $\nu \xrightarrow{Q[k|\nu]} k$. Here we make use of a result from [23] on the direct detection photon channel. The photon channel and our noisy neural subchannel are mathematically equivalent, by which we mean that they are both described by Eq. (2) except that for the photon channel $k$ denotes the number of emitted photons rather than the number of action potentials. See [20, 28] for further discussion on the connection between these channels. Note that [23] assumes the rate is a continuous function of time, as is the case in our study, and yet Theorem 1 of [23] states that the optimal distribution of $\nu$, $P_\nu(\nu)$, has a discrete form. This means that since $I(k; x) = I(k; \nu)$, the mutual information in the neural channel is maximized when the information in the noisy subchannel is maximized. By Theorem 1 of [23], this occurs when the distribution $P_\nu(\nu)$ is discrete.

Unfortunately, the theorem does not provide any means for finding a closed-form solution for the optimal discrete distribution, $P_\nu(\nu)$. However, its utility is that it allows a reduction in the set of functions we need to consider when optimizing $P_\nu(\nu)$ and/or the tuning curve $g(x)$.

Without loss of generality we can now introduce the following simplifying restriction for the function $g(x)$. Let $g(x)$ be a non-decreasing multi-step function

$$g(x) = \sum_{i=0}^{M-1} \gamma_i \sigma(x - \theta_i),$$  

(17)

where $M$ is the number of levels and $\sigma(.)$ is the Heaviside step function. Letting $\beta_i = \sum_{n=0}^{i} \gamma_n$ we have $\theta_{i+1}$ as the value of $x$ at which $g(x)$ jumps from value $\beta_i$ to $\beta_{i+1}$. Since we assume $x_{\text{min}} = \theta_0 < \theta_1 < \theta_2 < \ldots < \theta_{M-1} < \theta_M = x_{\text{max}}$, the optimal $g(x)$ is unique. This latter requirement means that we consider only the case of monotonically non-decreasing (sigmoidal) tuning curves. Without this restriction it is not possible to find a unique solution and hence this study does not generalize to non-monotonic tuning curves. This is not highly restrictive,
since sigmoidal tuning curves are widely observed in many sensory modalities [29]. The mutual information of the neural channel can be written as

$$ I(x; k) = \sum_{k=0}^{\infty} \sum_{i=0}^{M-1} \alpha_i Q[k|\beta_i] \log_2 \frac{Q[k|\beta_i]}{\sum_{n=0}^{M-1} \alpha_n Q[k|\beta_n]}, $$

where $\alpha_i = \int_{\theta_i}^{\theta_i+1} dx P_z(x)$. The optimal function $g(x)$ cannot be easily found in an analytical form using variational principles, because it leads to a set of transcendental equations. Therefore we use stochastic gradient descent methods to solve for the optimal $P_r(\nu)$.

Fig. 7 shows the results. The upper insets display the normalized optimal tuning curve, $f(x) \equiv g(x)/N$, for four different values of maximum mean spike count, $N$. Note that $I(x; k)$ in Eq. (18) is parameterized entirely by the set $\alpha_i$, $\beta_i$, $i = 0, \ldots, M - 1$, and it is these parameters that are optimized. The set of $\theta_i$-s required for the optimal $g(x)$ can be obtained for any given $P_z(x)$ from the $\alpha_i$. Hence, in Fig 7, without loss of generality we have assumed that the stimulus is uniformly distributed on $[0, 1]$. Similarly, the $\gamma_i$-s follow from $\beta_i$.

For small $N < 3$, only two firing rates are observed; for values of $x < \theta_1$, $f(x) = 0$ (the absence of firing) while for larger values of $x$, $f(x) = 1$ (firing at the maximum allowable spike rate). This form of optimal binary coding has been predicted previously for Poisson neurons using estimation theory [17]. It also agrees with the well known result that a binary source maximizes information through a Poisson channel when the input can switch instantaneously between states [21, 28].

As $N$ is increased, the number of steps in the optimal tuning curve increases; e.g. for $N = 7$, two steps are observed giving rise to a ternary coding scheme, for $N = 15$ three steps are observed giving a 4-ary (quaternary) coding. In general, an $M$-ary code will be optimal with increasing $N$. As $N \to \infty$ we predict that the optimal tuning curve will converge to a continuous function [16, 18]. Fig. 7 shows how the partition boundaries, $\theta_i$, vary as $N$ is increased; new boundaries can be seen to emerge via phase transitions. These appear to be continuous and hence are akin to second order phase transitions of the optimal tuning curve.

We have shown that $T \sim \text{SNR}$ and hence, given $N \sim T$, we have $N \sim \text{SNR}$; hence, increasing $N$ can be interpreted as a reduction in noise. Consequently, the bifurcation diagram in Fig. 7 can be seen to be similar to those presented in the previous section and others previously reported in the context of information optimisation [30, 12]. This similarity is even more striking when one notes that the steps in the optimal tuning curve can be interpreted as thresholds at which the input signal is mapped to a different output rate and, hence, are akin to the bifurcations of the thresholds observed in Figs. 2–5.

However, the analogy between the bifurcation diagram displayed in Fig. 7 and those in the Figs. 2–5 can be made even closer. The steps in the optimal $f(x)$ can be viewed as partitioning the stimulus into regions associated with neural subpopulations where each neuron in a subpopulation can only fire with two rates. For example, suppose an overall population consists of $K$ neurons and $M - 1$ sub-populations, within which each neuron is identical, and binary with rates 0 and $\gamma_i$. Since the neurons are Poisson, the sum of the $K$ individual normalized firing rates is equal to $f(x)$. For overall binary coding, the only way of achieving $f(x)$ would be a single sub-population, where each neuron is identical, and able to fire at two rates, $\phi_0/K = 0$ and $\phi_1/K = 1/K$, where rate $\phi_1/K$ is activated when $x \geq \theta_1$. For the ternary case, there would be two subpopulations, of sizes $J$ and $K - J$, with individual normalized firing rates $\phi_1/J$ and $(1 - \phi_1)/(K - J)$, so that the overall population has 3 rates: 0, $\phi_1$ and 1. The first subpopulation would only be activated when $x > \theta_1$ and the second when $x > \theta_2$.

We can estimate the sizes of the subpopulations in our example as follows. Since the sizes of the subpopulations are proportional to the integrated firing rates, the neurons for ternary coding are distributed with probabilities $P_1 = \phi_1 = \gamma_1/N$ and $P_2 = 1 - \phi_1$ respectively. The
Figure 7. The partition boundaries $\theta_i$ against the maximum mean spike count, $N = T\nu_{\text{max}}$. Also shown for $N = 2, 7, 15$ and 22, are the optimal $f(x)$ (top insets) and the population distributions (bottom insets). The parameters are $x_{\text{min}} = 0$, $x_{\text{max}} = 1$, $\nu_{\text{min}} = 0$, and $x$ uniformly distributed. Reproduced from Ref. [19]

quaternary coding scheme for $N = 15$ has three subpopulations with optimal individual firing rates proportional to $\gamma_1$, $\gamma_2$ and $1 - \gamma_1 - \gamma_2$, and overall rates $0$, $\phi_1$, $\phi_2$ and 1. The sizes of the subpopulations are therefore $P_1 \propto \gamma_1 = N\phi_1$, $P_2 \propto \gamma_2 = N(\phi_2 - \phi_1)$ and $P_3 = 1 - P_1 - P_2$, as shown in the lower insets in Fig. 7.

3. Discussion and Conclusions

We have investigated the neural parameters that maximise the mutual information in two very different neural populations. In the first we used McCulloch-Pitts neurons and introduced noise by adding noise to each neuron’s input. In the second system we considered each neuron to have Poisson statistics and quantified the amount of ‘noise’ in the system as the time window used to define the rate. In both these studies we came to identical conclusions. To optimise the mutual information it is necessary to choose the neural parameters so as to form subpopulations. More specifically, each subpopulation has identical neurons that have the same threshold. However, this threshold varies between subpopulations and the relationship between thresholds is chosen so as to maximise the information flow in the population as a whole. Furthermore, the number of subpopulations is strongly linked to the degree of stochasticity; the stronger the noise the fewer the number of subpopulations.
The fact that two seemingly very different systems maximise mutual information in the same manner suggests that our results are potentially fundamental and generic. The generic nature of the results appears to stem from the fact that the majority of channels known in engineering achieve channel capacity for discrete input signal distributions [21, 22, 23, 24, 25]. Whilst neural channels cannot control directly the signal distribution, it appears that they organise their thresholds to effect directly a quantization of the signal; thus approximating the capacity achieving input distribution. Consequently, if neural channels maximise Shannon mutual information we expect to observe subpopulations.

Furthermore, we have shown that the level of noise strongly influences the structure of the neural system. Small noise leads to a diversity in the selection of neural parameters whereas large noise results in a small number of subpopulations and hence a large degree of commonality in the neural parameters. This indicates that the level of noise intrinsic to the sensory system is a significant influence on the form that an optimal code will take.

Acknowledgments
We thank Steven Holmes, Riccardo Mannella and Jan A. Freund for fruitful discussions. This work was funded by the EPSRC (grant GR/R35650/01 and EP/D05/1894/1(P)). Mark D. McDonnell is funded by the Australian Research Council, Grant No. DP0770747, and Australia’s ACoRN (ARC Communications Research Network).

References
[1] Faisal A A, Selen L P J and Wolpert D M 2008 Nat. Rev. Neurosci. 9 292
[2] Shannon C E and Weaver W 1949 The mathematical theory of communication (Urbana: University of Illinois Press)
[3] Quiroga R Q and Panzeri S 2009 Nat. Rev. Neurosci. 10 173
[4] McDonnell M D, Amblard P O and Stocks N G 2009 J. Stat. Mech. P01012
[5] Linsker R 1988 IEEE Computer 21 105
[6] Barlow H 1961 Possible principles underlying the transformations of sensory messages Sensory Communication (Cambridge, MA: MIT press)
[7] Bell A J and Sejnowski T J 1995 Neural Comput. 7 1129
[8] Smith E C and Lewicki M 2006 Nature 439 978
[9] Stocks N G 2000 Phys. Rev. Lett. 84 2310
[10] McDonnell M D, Abbott D and Pearce C E M 2002 Microelectron. J. 33 1079
[11] Nocedal J and Wright S J 1999 Numerical Optimization (New York: Springer-Verlag)
[12] McDonnell M D, Stocks N G, Pearce C E M and Abbott D 2008 Stochastic Resonance: From Suprathreshold Stochastic Resonance to Stochastic Signal Quantisation (Cambridge: Cambridge University Press)
[13] Gerstner W and Kistler W 2002 Spiking Neuron Models (Cambridge: Cambridge University Press)
[14] Dayan P and Abbott L F 2001 Theoretical Neuroscience (Cambridge: MIT Press)
[15] Lansky P, Rodriguez R and Sacerdote L 2004 Neural Comput. 16 477
[16] Brunel N and Nadal J 1998 Neural Comput. 10 1731
[17] Bethge M, Rotermund D and Pawelzik K 2003 Phys. Rev. Lett. 90 088104
[18] McDonnell M D and Stocks N G 2008 Phys. Rev. Lett. 101 058103
[19] Nikitin A P, Stocks N G, Morse R P and McDonnell M D 2009 Phys. Rev. Lett. 103 138101
[20] Teich M C and McGill W J 1976 Phys. Rev. Lett. 36 754
[21] Kabanov Y M 1978 Theory Probab. Appl. 23 143
[22] Davis M H A 1980 IEEE Trans. Inform. Theory IT-26 710
[23] Shamai S 1990 IEEE Proc. I Commun. Speech Vis. 137 424
[24] Smith J G 1971 Information and Control 18 203
[25] Huang J and Meyn S P 2005 IEEE Trans. Inform. Theory 51 2336
[26] Ikeda S and Manton J H 2009 Neural Comput. 21 1714
[27] Blahut R E 1987 Principles and practice of information theory (Reading, Massachusetts: Addison-Wesley)
[28] Johnson D H and Goodman I N 2008 Network: Computation in Neural Systems 19 13
[29] Salinas E 2006 PLoS Biology 4 2383
[30] Rose K 1994 IEEE Trans. Inform. Theory 40 1939