LETTER Communicated by Peter Latham

Implications of Neuronal Diversity on Population Coding

Maoz Shamir
shamir@bu.edu
Center for BioDynamics, Boston University, Boston, MA 02215, U.S.A.

Haim Sompolinsky
halm@fiz.huji.ac.il
Racah Institute of Physics and Center for Neural Computation, Hebrew University of Jerusalem, Jerusalem 91904, Israel

In many cortical and subcortical areas, neurons are known to modulate their average firing rate in response to certain external stimulus features. It is widely believed that information about the stimulus features is coded by a weighted average of the neural responses. Recent theoretical studies have shown that the information capacity of such a coding scheme is very limited in the presence of the experimentally observed pairwise correlations. However, central to the analysis of these studies was the assumption of a homogeneous population of neurons. Experimental findings show a considerable measure of heterogeneity in the response properties of different neurons.

In this study, we investigate the effect of neuronal heterogeneity on the information capacity of a correlated population of neurons. We show that information capacity of a heterogeneous network is not limited by the correlated noise, but scales linearly with the number of cells in the population. This information cannot be extracted by the population vector readout, whose accuracy is greatly suppressed by the correlated noise. On the other hand, we show that an optimal linear readout that takes into account the neuronal heterogeneity can extract most of this information. We study analytically the nature of the dependence of the optimal linear readout weights on the neuronal diversity. We show that simple online learning can generate readout weights with the appropriate dependence on the neuronal diversity, thereby yielding efficient readout.

1 Introduction

In many areas of the central nervous system, information on specific stimulus features is coded by the average firing rates of a large population of neurons (see Hubel & Wiesel, 1962; Georgopoulos Schwartz & Kettner, 1982; Razak & Fuzessery, 2002; Coltz, Johnson, & Ebner, 2000). Recently Yoon and Sompolinsky (1999) and Sompolinsky, Yoon, Kang, and Shamir (2001) have

Neural Computation **18, 1951–1986** (2006) © 2006 Massachusetts Institute of Technology

shown that information coded by the mean responses of a neural population is greatly suppressed in the presence of the experimentally observed positive correlations. Thus, in the presence of noise correlations, there exists a finite amount of noise in the neural representation of the stimulus. This noise cannot be overcome by increasing the population size. However, central to the analysis of Sompolinsky et al. was the assumption of a homogeneous population of neurons. Empirical observations show a considerable measure of heterogeneity in the response properties of different neurons (Ringach, Shapley, & Hawken, 2002). Here we study the possible role of neuronal diversity in coding for information. We address two questions. First, what is the effect of heterogeneity on the information capacity? In particular, we are interested in the scaling of the information capacity with the number of cells in the population. Second, how does neuronal diversity affect biological readout mechanisms?

We address these questions in the context of a statistical model for the responses of a population of N neurons coding for an angular variable, θ, which we term the stimulus, such as the direction of arm movement during a simple reaching task or the orientation of a grating stimulus. Below we define the statistical model of the neural responses and review the main results for a homogeneous population of neurons. We use the Fisher information (see, e.g., Thomas & Cover, 1991; Kay, 1993) and the accuracy of biologically plausible readout mechanisms as measures of information capacity. The first question is addressed in section 2, where the Fisher information of a diverse population of neurons is studied. In section 3 we investigate the efficiency of linear readout mechanisms. First, we study the population vector (see Georgopoulos, Schwartz, & Kettner, 1986) readout; then we investigate the optimal linear estimator (see Salinas & Abbott, 1994) and show a different scaling of their efficiencies with the population size in the presence of correlations. Finally we summarize our results in section 4 and discuss further extensions of our theory.

1.1 The Statistical Model. We consider a system of N neurons coding for an angle, $\theta \in [-\pi, \pi)$. Let r_i denote the activity of the ith neuron during a single trial; r_i can be thought of as the number of spikes the ith neuron has fired within a specific time interval around the onset of a stimulus, θ. Assuming the firing rates of the neurons are sufficiently high, we model the probability distribution of the neural responses during a single trial, given the stimulus θ, according to a multivariate gaussian distribution:

$$P(\mathbf{r} \mid \theta) = \frac{1}{Z} \exp\left\{-\frac{1}{2}(\mathbf{r} - \mathbf{m}(\theta))^T \mathbf{C}^{-1} (\mathbf{r} - \mathbf{m}(\theta))\right\}. \tag{1.1}$$

Here $m_i(\theta)$, the tuning curve of neuron i, is the mean activity of the ith neuron averaged over many trials with the same stimulus, θ; $\mathbf{C}$ is the firing rate covariance matrix; $\mathbf{X}^T$ denotes the transpose of $\mathbf{X}$; and Z is a normalization constant. We model the single neuron tuning curve by a unimodal bell-shaped function of θ with a maximum at the neuron's preferred direction,

$$m_i(\theta) = m(\varepsilon_i, \phi_i - \theta), \tag{1.2}$$

where ϕ_i is the preferred direction of neuron i. We take the preferred directions of the neurons to be evenly spaced on the ring: $\phi_k = -\pi \frac{N+1}{N} + \frac{2\pi}{N} k$. The ε_i is a set of parameters characterizing the tuning curve of the ith neuron and representing the deviation from homogeneity. For example, ε_i can quantify how sharper or narrower is the tuning curve of neuron i than the stereotypical tuning curve or can represent the ratio between the tuning amplitude of the ith neuron and the tuning amplitude of the stereotypical tuning curve. Here, for simplicity, we shall assume that ε_i is a scalar. It is important to note that ε_i is a number that characterizes the tuning curve of neuron i; it is a property on the ith neuron and does not change from trial to trial (we ignore effects of plasticity and learning). However, different neurons in the population may have different values for their ε parameters, reflecting the heterogeneity of the population. Different neural populations can be characterized by different realizations of their set of $\{\varepsilon_i\}_{i=1}^N$. We shall assume the $\{\varepsilon_i\}_{i=1}^N$ are independent and identically distributed random variables that are independent of the stimulus, that is, $P(\{\varepsilon_i\}) = \prod_{i=1}^N p(\varepsilon_i)$.

We distinguish between two sources of randomness in this model. One source is the "warm fluctuations," represented by the trial-to-trial variability of the neural responses, equation 1.1. The second is the "quenched disorder," represented by the heterogeneity of the tuning curves of the different neurons, reflected by the distribution of the $\{\varepsilon_i\}$. Throughout this article, we will be interested in calculating quantities that involve spatial averaging over the entire population. The value of such quantities will depend on the specific realization of the neural heterogeneity, the $\{\varepsilon_i\}$, and will fluctuate from one realization of the neuronal heterogeneity to another. We can characterize such a quantity by its statistics with respect to the quenched randomness. Although for local quantities the quenched fluctuations may be considerable, they are uncorrelated spatially; hence, the quenched fluctuations of quantities that involve spatial averaging, relative to their means, will decrease to zero in the large N limit. This property of a random variable with vanishing standard deviation to mean ratio, in the large N limit, is called *self-averaging*. Note that the value of a self-averaging quantity in a typical system will be equal to its mean across different systems. The practical implications of this property are that one can replace a self-averaging quantity by its quenched mean for large systems. Hence, instead of computing self-averaged quantities for a specific realization of the neuronal heterogeneity, we can calculate the average of this quantity over different realizations of the heterogeneity.

We denote by angular brackets averaging with respect to the trial-to-trial fluctuations of the neural responses, given a specific stimulus:

$\langle X \rangle = \int \mathbf{dr} X P(\mathbf{r} \mid \theta)$. This averaging is done with a fixed set of parameters, $\{\varepsilon_i\}$, reflecting the fact that the single-neuron tuning curves are fixed and unchanged across many different trials. Averaging over the quenched disorder is denoted by double angular brackets, $\ll X(\{\varepsilon_i\}) \gg = \int \prod_{i=1}^{N} d\varepsilon_i\, p(\varepsilon_i) X(\{\varepsilon_i\})$. Fluctuations with respect to the distribution of the neural responses in a given system are denoted by δ, that is, $\delta X \equiv X - \langle X \rangle$. We use Δ to denote quenched fluctuations: $\Delta X \equiv X - \ll X \gg$.

It is convenient to write the tuning curve of neuron i as the sum of its quenched average, $\ll m_i(\theta) \gg$, plus a fluctuation $\Delta m_i(\theta)$:

$$m_i(\theta) = f(\phi_i - \theta) + \Delta m_i(\theta) \tag{1.3}$$

$$f(\phi_i - \theta) \equiv \ll m_i(\theta) \gg . \tag{1.4}$$

Note that in the last equation, we used equation 1.2 and the fact that the statistics of the $\{\varepsilon_i\}$ are independent of the neuronal preferred directions and the stimulus. Similar to the single-cell tuning curve, we model $f(\theta)$ by a smooth bell-shaped function that peaks at $\theta = 0$. Specifically, in our numerical simulations, we used the following average tuning curve,

$$f(\theta) = (f_{\max} - f_{ref}) \exp\left(\frac{\cos(\theta) - 1}{\sigma^2}\right) + f_{ref}, \tag{1.5}$$

where σ, $(f_{\max} - f_{ref})$ and f_{ref} are the tuning width, the tuning amplitude, and a reference value for the stereotypical average tuning curve $f(\theta)$, respectively.

For a given stimulus, the quenched fluctuations of the tuning curves, $\{\Delta m_i(\theta)\}$ (see equation 1.3), is a set of zero mean independent random variables with respect to the statistics of the quenched disorder. An important quantity for the calculation of the Fisher information is the derivative of the tuning curve with respect to θ, $m_i' = \frac{\partial m_i}{\partial \theta}$. The quenched fluctuations of the tuning curve, $\{\Delta m_i(\theta)\}$, are also smooth periodic functions of θ as a difference of such functions. Using the independence of the $\{\varepsilon_i\}$ and equation 1.2, one obtains

$$m_i'(\theta) = f'(\phi_i - \theta) + \Delta m_i'(\theta) \tag{1.6}$$

$$\ll \Delta m_i'(\theta) \gg = \int p(\varepsilon_i) \frac{\partial \Delta m_i(\varepsilon_i, \theta)}{\partial \theta} d\varepsilon_i = \frac{d}{d\theta} \int p(\varepsilon_i) \Delta m_i(\varepsilon_i, \theta) d\varepsilon_i = 0 \tag{1.7}$$

$$\ll \Delta m_i'(\theta) \Delta m_j'(\theta) \gg = \delta_{ij} K(\phi_i - \theta), \tag{1.8}$$

where $K(\phi_i - \theta)$ is the variance of the tuning curve derivative of a neuron with preferred direction ϕ_i, given a stimulus θ, with respect to the quenched disorder. We further assume that the quenched fluctuations of the tuning curve derivatives, $\{\Delta m_i'\}$, follow gaussian statistics. In section 2, where we study the Fisher information of a heterogeneous population of neurons, we use equations 1.6 to 1.8 to define the gaussian statistics of the quenched fluctuations of tuning curves.

For small, quenched fluctuations, one can expand the tuning curves, equation 1.2, in powers of ε_i and approximate

$$m_i(\theta) = f(\phi_i - \theta) + \varepsilon_i g(\phi_i - \theta) \tag{1.9}$$

where $g(\theta) = \partial m(\varepsilon = 0, \theta)/\partial\varepsilon$. A simple example, where the approximation of equation 1.9 is exact, is in the case where

$$m_i(\theta) = f(\phi_i - \theta)(1 + \varepsilon_i). \tag{1.10}$$

We term this model the *amplitude diversity* model. In section 3, where we address the question of readout, we use the specific form of equation 1.10 for the tuning curves in order to make the analysis analytically tractable. This model, equation 1.10, is also used for all of the numerical results presented in this article. We further assume, in section 3 and in all of the numerical results, that the $\{\varepsilon_i\}$ are independent and identically distributed (i.i.d.) gaussian random variables with zero mean and variance κ,

$$\ll \varepsilon_i \gg = 0 \quad \forall i \tag{1.11}$$

$$\ll \varepsilon_i \varepsilon_j \gg = \delta_{ij}\kappa \quad \forall i,\, j. \tag{1.12}$$

In this case, $\Delta m_i' = -\varepsilon_i f'(\phi_i - \theta)$ and

$$\ll (\Delta m_i'(\theta))^2 \gg = K(\phi_i - \theta) = \kappa |f'(\phi_i - \theta)|^2. \tag{1.13}$$

We assume the response covariance of two neurons i and j, C_{ij}, is independent of the stimulus, θ, and only depends on the functional distance between the two neurons, that is, the difference in their preferred directions. A decrease in the response covariance of different neurons with the increase in their functional distance has been reported in cortical areas (see, e.g., Zohary, Shadlen, & Newsome, 1994; Lee, Port, Kruse, & Georgopoulos, 1998; van Kan, Scobey, & Gabor, 1985; Mastronarde, 1983). Specifically, in all of our numerical results, we have used exponentially decaying correlations,

$$C_{ij} = C(\phi_i - \phi_j) = a\left[\delta_{ij} + c(1 - \delta_{ij})\exp\left(-\frac{|\phi_i - \phi_j|}{\rho}\right)\right], \tag{1.14}$$

where a is the variance of the single-neuron response, c and ρ are the correlation strength and correlation length, respectively, and $|\phi_i - \phi_j| \in [0, \pi]$ is the functional distance between neurons i and j. Note that in this simple model, we did not incorporate any measure of diversity in the higher-order statistics of the neuronal responses. Hence, the Fourier modes of the system are eigenvectors of the covariance matrix. We assume that in the biologically relevant regime, every neuron is correlated with a substantial fraction of the entire population. Mathematically this means that as we consider the limit of large N, both ρ and c remain finite. In this regime, the eigenvalues of the covariance matrix, **C**, scale linearly with the size of the system, N.

For all numerical results presented in this article, we used the amplitude diversity model for the neuronal tuning curves, equation 1.10, with the following parameters: $\rho = 1$, $c = 0.4$, $a = 40$ [sec^{-1}], $\sigma = 1/\sqrt{2}$, $f_{\max} = 60$ [sec^{-1}], $f_{ref} = 20$ [sec^{-1}], and $\kappa = 0.25$, unless stated otherwise. Note that these parameters are are given in rates, that is, the number per 1 second. In order to obtain the spike count statistics in a given time interval T, the tuning curves and correlation strength were scaled by a factor of T; unless stated otherwise, we used $T = 0.5$ [sec].

1.2 The Fisher Information. Throughout this article, we will be interested in studying the efficiency of different estimators, $\hat{\theta}(\mathbf{r})$, of the stimulus θ. We define the efficiency of an estimator by the inverse of its average quadratic estimation error, $1/\langle(\hat{\theta} - \theta)^2\rangle$. It is convenient to distinguish between two sources of estimation error: the bias, $b = \langle\hat{\theta}\rangle - \theta$, and the trial-to-trial variability, $\langle(\delta\hat{\theta})^2\rangle$.

The Fisher information (see, e.g., Thomas & Cover, 1991; Kay, 1993) is given by

$$J = \left\langle \left(\frac{\partial \log P(\mathbf{r}|\theta)}{\partial\theta} \right)^2 \right\rangle. \tag{1.15}$$

From the Cramér-Rao inequality, the square estimation error of any readout $\hat{\theta}(\mathbf{r})$ is bounded by

$$\langle(\hat{\theta} - \theta)^2\rangle = \langle(\delta\hat{\theta})^2\rangle + b(\theta)^2 \tag{1.16}$$

$$\langle(\delta\hat{\theta})^2\rangle \geq \frac{(1 + b^{'})^2}{J}, \tag{1.17}$$

where $b' = \frac{db}{d\theta}$. The Fisher information of this model is given by

$$J = \mathbf{m}'^{T}\mathbf{C}^{-1}\mathbf{m}', \tag{1.18}$$

where, $\mathbf{m}' = \frac{d\mathbf{m}}{d\theta}$. Note that the Fisher information has the form of a squared signal-to-noise ratio, where the signal is the sensitivity of the neural

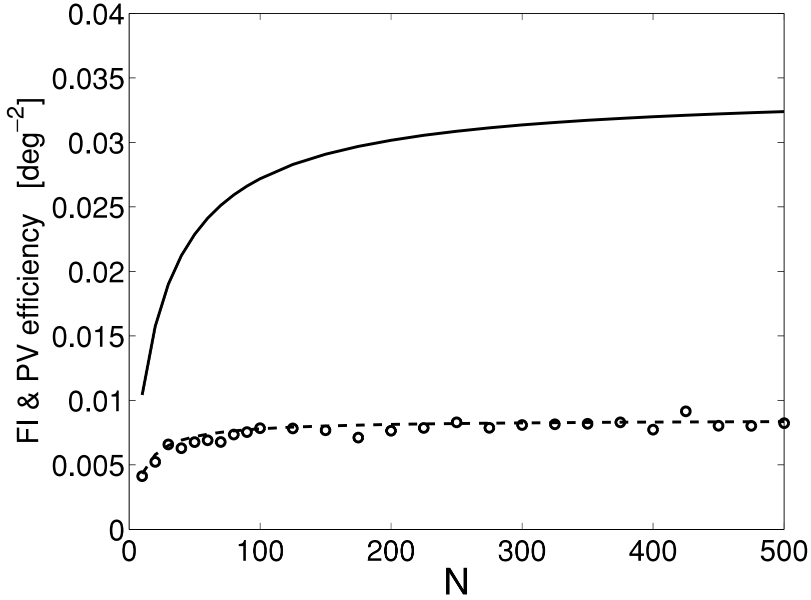

Figure 1: The Fisher information and population vector efficiency of a homogeneous population of neurons. The Fisher information is shown by the solid line as a function of the number of cells in the population, N. The analytical result for the population vector efficiency, equation 1.26, is shown by the dashed line. The open circles show the results of numerical calculation of averaging the population vector estimation error over 2000 trials estimating $\theta = 0$. In this figure, $\kappa = 0$ was used for a homogeneous population.

responses to small changes in the stimulus, $\mathbf{m}'$, and the squared noise is represented by the correlation matrix, $\mathbf{C}$.

1.2.1 Fisher Information of an Isotropic Population. In the limit of an isotropic population, $K(\theta) = 0$ (see equation 1.8), the signal, $\mathbf{m}' = \mathbf{f}'$, resides in a low-dimensional subspace of the neural responses, spanned by the slowly varying collective modes of the system. However, due to the correlations, both the squared signal and noise will scale linearly with the size of the system, yielding $J = \mathcal{O}(1)$ even in the large N limit (see Sompolinsky et al., 2001, for further discussion). Figure 1 shows the Fisher information of an isotropic system, $\kappa = 0$ in the amplitude diversity model (solid line), as a function of the population size, N. As can be seen from the figure, the Fisher information of an isotropic system saturates to a finite value in the limit of large N. In contrast, in a diverse population of neurons, the signal, $\mathbf{m}'$, will have an $\mathcal{O}(\sqrt{N})$ projection on a subspace spanned by eigenvectors of **C** corresponding to eigenvalues of $\mathcal{O}(1)$. This effect of the neuronal diversity is studied in section 2.

1.3 Linear Readout. A linear readout is an estimator of the form $\hat{z} = \sum_i w_i r_i$, where **w** is a fixed weights vector that defines the readout. It is convenient to adopt complex notation for the stimulus. We denote by $z = e^{i\theta}$ a two-dimensional unit vector in the complex plane pointing in the direction of the stimulus, θ. Similarly, the estimator, $\hat{z} = \hat{x} + i\hat{y}$, will represent a two-dimensional vector in the complex plane with $\hat{\theta} = \arg(\hat{z})$. One can measure the performance of such a readout by either the efficiency of angular estimation, $\langle(\hat{\theta} - \theta)^2\rangle^{-1}$, or by the Euclidean distance between z and $\hat{z}$. In this work, we employ both measures. We shall call $\langle(\hat{\theta} - \theta)^2\rangle^{-1}$ the *efficiency* of the estimator and $\langle|\hat{z} - z|^2\rangle$ the *Euclidean error*. Let $E(\mathbf{w})$ be the Euclidean error of a linear readout with a weights vector **w**,

$$E(\mathbf{w}) = \int \frac{d\theta}{2\pi} \left\langle |\hat{z} - z|^2 \right\rangle = \mathbf{w}^\dagger \mathbf{Q} \mathbf{w} - \mathbf{w}^\dagger \mathbf{U} - \mathbf{U}^\dagger \mathbf{w} + 1 \tag{1.19}$$

$$\mathbf{Q} = \int \frac{d\theta}{2\pi} \langle \mathbf{r}\mathbf{r}^T \rangle \tag{1.20}$$

$$\mathbf{U} = \int \frac{d\theta}{2\pi} \langle \mathbf{r} \rangle e^{i\theta}, \tag{1.21}$$

where $\mathbf{X}^\dagger$ denotes the conjugate transpose of **X**. It is important to note that, being a function of the neural responses, the linear estimator, $\hat{z} = \mathbf{r}^T\mathbf{w}$, is a random variable that fluctuates from trial to trial with a probability distribution that depends on the stimulus, θ (see equation 1.1). The Euclidean error, $E(\mathbf{w})$, defined in equation 1.19, incorporates two averaging steps. First is averaging the Euclidean error that results from the trial-to-trial fluctuations of the linear readout, $\langle|\hat{z} - z|^2\rangle$, for a given stimulus angle, θ. The second is averaging the Euclidean error over the different possible stimuli by integrating over the stimulus angle, θ, assuming a uniform prior. The optimal linear estimator (Salinas & Abbott, 1994) is defined by the set of linear weights, $\mathbf{w}_{ole}$, that minimizes E. The optimal linear estimator weights are given by $\mathbf{w}_{ole} = \mathbf{Q}^{-1}\mathbf{U}$ (see also equations 3.4 to 3.6), and the average quadratic estimation error of the optimal linear estimator is given by $E(\mathbf{w}_{ole}) = 1 - \mathbf{U}^\dagger\mathbf{Q}^{-1}\mathbf{U}$. Below we present the main results for the optimal linear estimator efficiency in a correlated homogeneous population of neurons and in an uncorrelated heterogeneous population. The study of the optimal linear estimator performance in a heterogeneous population of correlated cells is the focus of section 3.

1.3.1 Linear Readout in a Correlated Homogeneous Population of Neurons. In the case of a homogeneous population of neurons, $K(\theta) = 0$ (see equation 1.8), the optimal linear estimator, $\hat{z}_{ole} = \mathbf{r}^T\mathbf{w}_{ole}$, is given by

(see Shamir & Sompolinsky, 2004)

$$\hat{z}_{pv} = \frac{\tilde{f}^{(1)^*}}{N\left(|\tilde{f}^{(1)}|^2 + \tilde{c}_1\right)} \sum_{j=1}^{N} e^{i\phi_j} r_j, \tag{1.22}$$

where we have used the following definitions for the Fourier transforms:

$$\tilde{f}^{(n)} = \frac{1}{N} \sum_{j=1}^{N} f(\phi_j) e^{in\phi_j} = \int \frac{d\phi}{2\pi} f(\phi) e^{in\phi} \tag{1.23}$$

$$\tilde{c}_n = \frac{1}{N^2} \sum_{j=1}^{N} C_{jk} e^{in(\phi_j - \phi_k)} = \int \frac{d\phi}{2\pi} C(\phi) e^{in\phi}. \tag{1.24}$$

Note that $N\tilde{c}_n$ is the eigenvalue of the correlation matrix, **C**. The asterisk in the nominator of the right-hand side of equation 1.22, $\tilde{f}^{(1)^*}$, denotes the complex conjugate of the Fourier transform $\tilde{f}^{(1)}$. However, since the average tuning curve, $f(\theta)$, is a real, even function of the stimulus, θ, its Fourier transforms are pure real. We shall omit the use of conjugate notation of real-valued terms hereafter.

This linear estimator, equation 1.22, is the population vector readout. In this case, one can show that the population vector is unbiased with respect to its argument and that its average quadratic error, $E(PV)$, is given by

$$E(PV) = \frac{1}{1 + |\tilde{f}^{(1)}|^2/\tilde{c}_1}, \tag{1.25}$$

which is of $\mathcal{O}(1)$ even in the limit of large N. Assuming small, angular estimation errors, one can expand $\hat{\theta}$ in the fluctuations of $\hat{x}$ and $\hat{y}$ and study the efficiency of the population vector angular estimation. For a homogeneous population, this error will result only from the trial-to-trial fluctuations and will be independent of the stimulus in its magnitude. The efficiency of the population vector, in this case, is given by (see Sompolinsky et al., 2001; details of the calculation of the population vector efficiency also appear in appendix B)

$$\frac{1}{\langle(\delta\hat{\theta}_{pv})^2\rangle} = \frac{2|\tilde{f}^{(1)}|^2}{\tilde{c}_1} < J. \tag{1.26}$$

Thus, in the biologically relevant regime for the correlations, $c > 0,\ c = \mathcal{O}(1),\ \rho = \mathcal{O}(1)$, both the nominator and denominator of equation 1.26 are $\mathcal{O}(1)$ in N, and the efficiency of the population vector saturates to a finite limit for large N. This can be seen in Figure 1, which shows the population vector efficiency as a function of the number of cells in the population

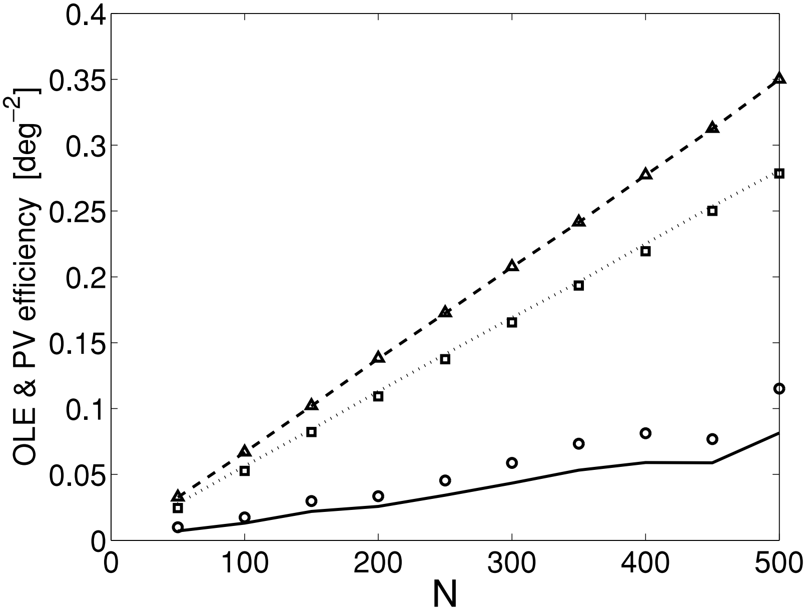

Figure 2: Linear readout efficiency in an uncorrelated population of neurons. The population vector efficiency, $\ll \langle(\hat{\theta}_{pv} - \theta)^2\rangle \gg^{-1}$ (solid line), is shown as a function of the population size, N. The inverse of the population vector bias, $\ll b_{pv}^2 \gg^{-1}$, and the population vector variance, $\ll \langle(\delta\hat{\theta}_{pv})^2\rangle \gg^{-1}$, are shown by the open circles and boxes respectively. The efficiency of the optimal linear estimator, $\ll \langle(\hat{\theta}_{ole} - \theta)^2\rangle \gg^{-1}$, is represented by the dashed line, and the inverse of its average variance, $\ll \langle(\delta\hat{\theta}_{ole})^2\rangle \gg^{-1}$, by the open triangles. The dotted line shows the population vector efficiency in an uncorrelated homogeneous neural population. The estimator's efficiency was calculated numerically by averaging over 400 different realizations of the neuronal diversity for each point. The estimation error for a given realization of the neuronal diversity was computed by averaging the angular estimation error of the readout for 500 trials of estimating $\theta = 0$. In this figure, $c = 0$ was used. For the population vector efficiency in a homogeneous neural population (dotted line), $\kappa = 0$ was used.

(dashed line and open circles). For large systems, the population vector efficiency reaches a size-independent limit.

1.3.2 Linear Readout in a Heterogeneous Population of Uncorrelated Neurons. In a diverse population of neurons, the population vector readout is no longer the optimal linear estimator. Moreover, the population vector estimator is biased. Figure 2 shows the efficiency of the population vector for angular estimation, $\ll \langle(\hat{\theta}_{pv} - \theta)^2\rangle \gg^{-1}$, in an uncorrelated ($c = 0$)

heterogeneous ($\kappa > 0$ in the amplitude diversity model, equation 1.10) population of neurons (solid line). The inverse of the population vector bias, $\ll b_{pv}^2 \gg^{-1}$ is shown by the open circles. The inverse of the population vector variance is shown by the open squares. For comparison, we plot the efficiency of the population vector in a homogeneous population ($\kappa = 0$) by the dotted line. From the figure, one can see that in a heterogeneous population of neurons, the efficiency of the population vector is decreased, relative to the homogeneous case, due to the added bias term that scales as $\ll b_{pv}^2 \gg \propto 1/N$; note that the population vector variance remains the same (compare open boxes and dotted line).

The dashed line in Figure 2 shows the efficiency of the optimal linear estimator, and the triangles show the inverse of its variance. The performance of the optimal linear estimator is superior to that of the population vector for two reasons. First, the optimal linear estimator is practically unbiased (compare the dashed line and open triangles). Second, the variance of the optimal linear estimator is smaller than the variance of the population vector. This results from the fact that the population vector extracts information only from the slowly varying collective modes of the system, whereas the optimal linear estimator can extract information from the higher-order modes as well, thus increasing its signal-to-noise ratio. However, the efficiency of both readouts scales linearly with the size of the system. Thus, in the case of an uncorrelated population, there is no qualitative difference between the two readouts. Below we show that in the presence of correlations, the neuronal diversity produces a qualitative effect on both the information capacity of the system (section 2) and the efficiency of different readouts (section 3).

2 The Fisher Information of a Diverse Correlated Population

The Fisher information, equation 1.18, of this model with $K(\theta) > 0$ (see equation 1.8) is given by

$$\begin{aligned} J &= \sum_{i,j=1}^{N} (f_i' + \Delta m_i') C_{ij}^{-1} (f_j' + \Delta m_j') \\ &= \sum_{i,j=1}^{N} f_i' C_{ij}^{-1} f_j' + 2 \sum_{i,j=1}^{N} \Delta m_i' C_{ij}^{-1} f_j' + \sum_{i,j=1}^{N} \Delta m_i' C_{ij}^{-1} \Delta m_j'. \end{aligned} \tag{2.1}$$

The Fisher information of a specific system, that is, for a given realization of the $\{\Delta m_i\}$, is a random variable that fluctuates from one realization of the neuronal diversity to another with respect to the quenched distribution of the $\{\Delta m_i\}$. The statistics of the Fisher information can be characterized by its moments. We find (see appendix A) that to a leading order in N,

$$\ll J \gg = N\bar{K}d + J_{homog} \tag{2.2}$$

$$J_{homog} = \mathbf{f}'^T \mathbf{C}^{-1} \mathbf{f}' = \mathcal{O}(1) \tag{2.3}$$

$$\ll (\Delta J)^2 \gg = 2N\overline{K^2}d^2 + \mathcal{O}(1), \tag{2.4}$$

where d is the diagonal element of the inverse of the correlation matrix, $d = [\mathbf{C}^{-1}]_{ii}$; J_{homog} is the Fisher information of a homogeneous population ($\kappa = 0$), which is of $\mathcal{O}(1)$ in the presence of correlation; and

$$\bar{K} = \int \frac{d\theta}{2\pi} K(\theta) \tag{2.5}$$

$$\overline{K^2} = \int \frac{d\theta}{2\pi} K(\theta)^2. \tag{2.6}$$

From equations 2.2 and 2.4, one finds that for large systems, the sample-to-sample fluctuations of the Fisher information are small relative to its mean, $\sqrt{\ll (\Delta J)^2 \gg} / \ll J \gg = \mathcal{O}(1/\sqrt{N})$. Hence, for large populations, the Fisher information of a typical system will be equal to its mean across many samples. This property of the Fisher information is an example of self-averaging. Figure 3a shows the mean and Figure 3b the variance of the Fisher information as a function of the population size N. The mean and variance of J were computed by averaging over 400 realizations of the neuronal populations (open circles) in the amplitude diversity model with $\kappa = 0, 0.05, 0.1, 0.25$ from bottom to top. The analytical results, equations 2.2 and 2.4, are shown by the solid lines as a function of the population size. Note that $\kappa = 0$ is the case of an isotropic population of neurons. In this case, the Fisher information saturates to a finite value in the limit of large N. In contrast, for $\kappa > 0$, the Fisher information of the system increases linearly with the population size (top lines) with a slope that is linear in κ.

Interestingly, after averaging over the heterogeneity of the population, the statistics of the Fisher information are independent of the stimulus, θ. Hence, for all $\theta \in [-\pi, \ldots \pi)$ the Fisher information, $J(\theta)$, will be equal to its quench average up to $\mathcal{O}(\sqrt{N})$ corrections. Figure 4 shows the stimulus fluctuations of the Fisher information for a typical realization of the neuronal diversity in a system of $N = 1000$ neurons in the amplitude diversity model, equation 1.10, with $\kappa = 0.25$. From the figure, one can see that the Fisher information is a smooth function of θ and is equal to $\ll J \gg$ up to small fluctuations of $\mathcal{O}(\sqrt{N})$.

For local discrimination tasks, linear readout can extract all the information coded by the first-order statistics of the neural responses (see Shamir & Sompolinsky, 2004). Below we study the efficiency of the linear readout for global estimation tasks.

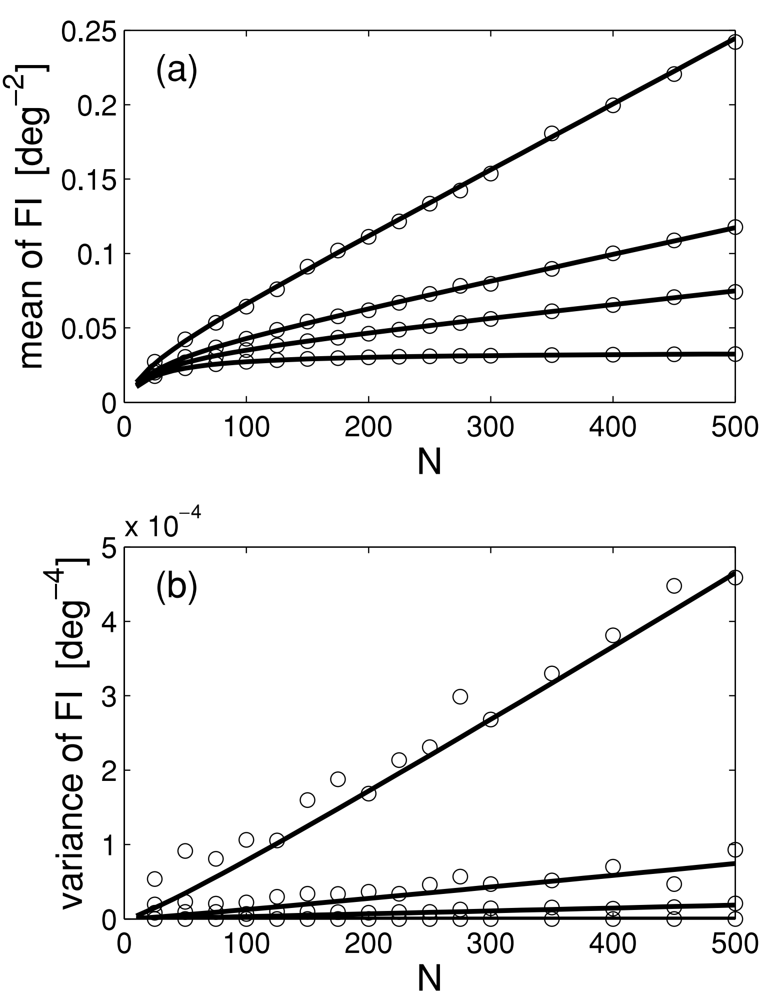

Figure 3: The mean (a) and variance (b) of the Fisher information with respect to the quenched statistics are shown as a function of the system size, N. The open circles show statistics of the Fisher information as calculated numerically by averaging the Fisher information of 400 different realizations of the neuronal diversity. The statistics were calculated in the amplitude diversity model with $\kappa = 0, 0.05, 0.1, 0.25$ from bottom to top. The solid lines in a show the analytical result for the average Fisher information, equations 2.2 and 2.3. The solid lines in b show the leading, $\mathcal{O}(N)$, term of equation (30) for the Fisher information variance, $2N\overline{K^2}d^2$.

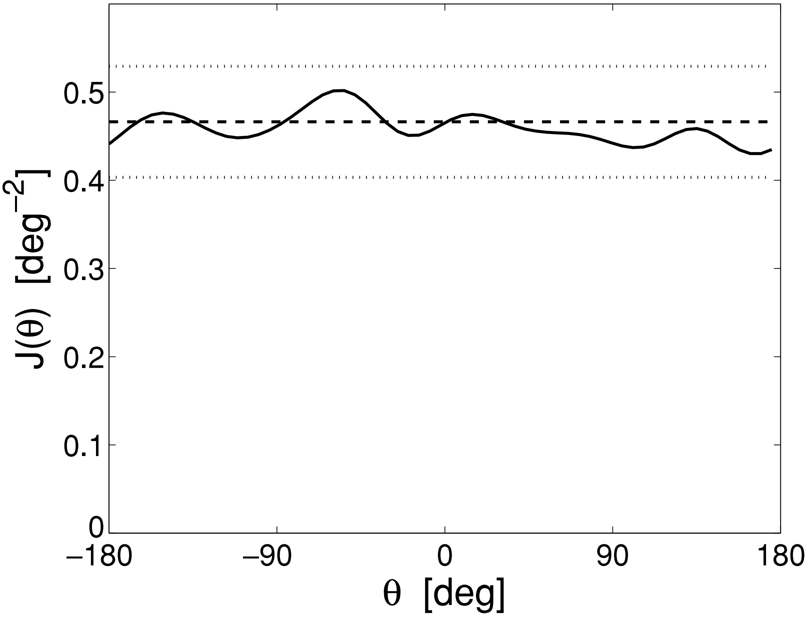

Figure 4: The Fisher information of a specific system. The Fisher information of a single system is plotted (solid line) as a function of the stimulus, θ, for a given typical realization of the neuronal diversity, the $\{\varepsilon_i\}$. The quenched average of the Fisher information, $\ll J \gg$, is shown in the dashed line for comparison. The dotted lines show $\ll J \gg \pm 2\sqrt{\ll \Delta J^2 \gg}$. In this figure, we used $N = 1000$.

3 Linear Readout

We now turn to the efficiency of linear readouts in a correlated heterogeneous population of neurons. For simplicity we restrict the discussion to the amplitude diversity model, equation 1.10. We first study the efficiency of the population vector readout. As mentioned above, in a homogeneous population of neurons, the population vector yields the optimal linear estimator. However, although the Fisher information in a diverse population grows linearly with the size of the system, the population vector efficiency saturates to a finite limit, as in the homogeneous case. In contrast, we show that the optimal linear estimator efficiency scales linearly with the number of cells in the population, N.

3.1 The Population Vector Readout. The efficiency of the population vector readout depends on the specific realization of the $\{\varepsilon_i\}$. We therefore characterize the population vector efficiency by its statistics with respect to the quenched disorder. Calculation of the quenched statistics of the population vector readout (see appendix B) reveals that the population vector bias is typically of order $1/\sqrt{N}$ with $\ll b_{pv} \gg = 0$ and variance,

$$\ll b_{pv}^2 \gg = \frac{\kappa}{2N} \frac{\widetilde{f^2}_{(0)} - \widetilde{f^2}_{(2)}}{|\tilde{f}^{(1)}|^2} = \mathcal{O}\left(\frac{\kappa}{N}\right), \tag{3.1}$$

where we have used the following notation for the Fourier transform of $f^2(\theta)$:

$$\widetilde{f^2}_{(n)} = \int \frac{d\varphi}{2\pi} f^2(\varphi) e^{in\varphi}. \tag{3.2}$$

Note that after the quenched averaging, the statistics of the bias and variance of the population vector readout are independent of the stimulus. Analysis of the population vector trial-to-trial fluctuations (see appendix B) shows that the variance of the population vector estimator, $\langle (\delta\hat{\theta}_{pv})^2 \rangle$, is a self-averaging quantity on the order of 1 with

$$\ll \langle (\delta\hat{\theta}_{pv})^2 \rangle \gg = \frac{\tilde{c}_1}{2|\tilde{f}^{(1)}|^2} = \mathcal{O}(1), \tag{3.3}$$

which is the same as the population vector efficiency in the homogeneous case, equation 1.26. Figure 5 shows the average efficiency of the population vector in a diverse population (circles) in terms of one over the average square angular estimation error, $\ll \langle (\hat{\theta} - \theta)^2 \rangle \gg^{-1}$. The population vector efficiency was calculated by averaging over 400 realizations of the neuronal diversity. The estimation error for a given realization of the neuronal diversity was computed by averaging the angular estimation error of the population vector for 200 trials of estimating $\theta = 0$. The analytical results of substituting equations 3.1 and 3.3 into equation 1.16 are shown by the overlapping solid line. Comparing Figures 1 and 5 and equations 1.26 and 3.3, it is easy to see that the efficiency of the population vector readout is almost unaffected by the neuronal diversity in the limit of large N. Thus, although the Fisher information of a diverse population scales linearly with the size of the system, the efficiency of the population vector saturates to a finite limit as N grows, in the presence of correlations.

Similarly, we find that to a leading order in N, the neuronal diversity has no effect on the population vector performance in terms of the Euclidean distance measure, equation 1.19. Hence, the result of equation 1.25 still holds up to a correction of $O(1/\sqrt{N})$ (see appendix B), yielding $\mathcal{O}(1)$ error even in the large N limit. These results raise the question of whether it is possible to obtain a linear estimator that will be able to use the neuronal diversity in order to overcome the correlated noise and obtain an efficiency that scales linearly with the size of the system. As discussed above, in

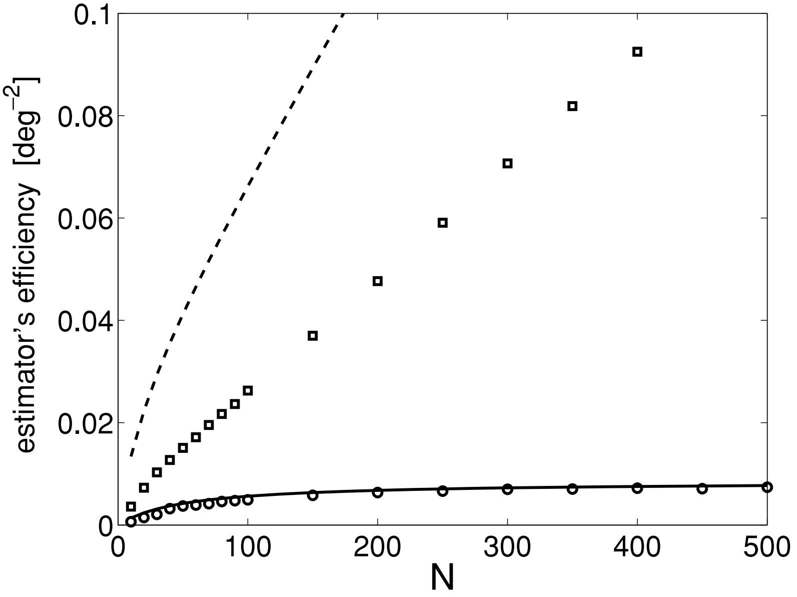

Figure 5: Linear readout average efficiency. The average efficiency of the population vector (open circles) and of the optimal linear estimator (open boxes) is plotted as a function of the size of the system. The efficiency of these readouts was calculated numerically by averaging over 400 different realizations of the neuronal diversity for each point. The estimation error for a given realization of the neuronal diversity was computed by averaging the angular estimation error of the readout over 200 trials of estimating $\theta = 0$. The Fisher information is shown in the upper dashed line for comparison. The bottom solid line shows the analytical result for the population vector error, equations 1.16, 3.3, and 3.1.

an isotropic population of neurons, the population vector is the optimal linear estimator. However, in a diverse population, the population vector is no longer optimal. Below we study the efficiency of the optimal linear estimator in a heterogeneous population of neurons.

3.2 The Optimal Linear Estimator. The optimal linear estimator weights are given by (see equations 1.19–1.21)

$$\mathbf{w}_{ole} = \mathbf{Q}^{-1}\mathbf{U} \tag{3.4}$$

$$Q_{ij} = C_{ij} + (1+\varepsilon_i)(1+\varepsilon_j)\int \frac{d\theta}{2\pi} f(\phi_i - \theta) f(\phi_j - \theta) \tag{3.5}$$

$$U_j = (1+\varepsilon_j)\tilde{f}^{(1)} e^{i\phi_j}. \tag{3.6}$$

The efficiency of the optimal linear estimator, in terms of one over the average quadratic estimation error, $\ll \langle(\hat{\theta} - \theta)^2\rangle \gg^{-1}$, is shown in Figure 5 (boxes). The optimal linear estimator efficiency was calculated by averaging over 400 realizations of the neuronal diversity. The estimation error for a given realization of the neuronal diversity was computed by averaging the angular estimation error of the optimal linear estimator over 200 trials of estimating $\theta = 0$. From the figure, one can see that while the population vector efficiency (solid line and open circles) saturates to a finite limit, the optimal linear estimator efficiency scales linearly with the population size, N. Obtaining a complete analytical expression for the optimal linear estimator and its efficiency is not an easy task, mainly due to the difficulty of inverting the random matrix $\mathbf{Q}$. However, for large populations, one can expand the optimal linear estimator to a leading order in $1/N$, as presented in the following section.

3.2.1 The Zeroth Approximation of the Optimal Linear Estimator. To a leading order in $1/N$, the optimal linear estimator weights are given by (see appendix C)

$$w_{ole,j} = w_j^{(0)} + \mathcal{O}(N^{-3/2}) \tag{3.7}$$

$$w_j^{(0)} = \frac{1}{N\kappa\tilde{f}^{(1)}}\epsilon_j e^{i\phi_j}. \tag{3.8}$$

Figure 6 shows the average overlap between the optimal linear estimator and the zeroth approximation readout weights: $Real\{\ll \frac{\mathbf{w}_{ole}^{\dagger}\mathbf{w}^{(0)}}{\sqrt{\|\mathbf{w}_{ole}\|\|\mathbf{w}^{(0)}\|}} \gg\}$. As can be seen from the figure, the overlap approaches the value of 1 as N grows. Hence, for large N, the zeroth approximation converges to the optimal linear estimator, $\mathbf{w}^{(0)} \overset{N\to\infty}{\longrightarrow} \mathbf{w}_{ole}$.

We find (see appendix D) that the average Euclidean error, equation 1.19, of the zeroth approximation fluctuates from sample to sample with mean and standard deviation of the order of $1/N$. Assuming small angular estimation errors, we can study the angular estimation efficiency of the zeroth approximation. Our calculations (see appendix D) show that the bias of the zeroth approximation is zero, on average, with fluctuations that are of order $1/\sqrt{N}$:

$$\ll b \gg = 0 \tag{3.9}$$

$$\ll (\Delta b)^2 \gg = \frac{1 + \frac{1}{2\kappa}}{N} \frac{\widetilde{f^2}_{(0)} - \widetilde{f^2}_{(2)}}{|\tilde{f}^{(1)}|^2}. \tag{3.10}$$

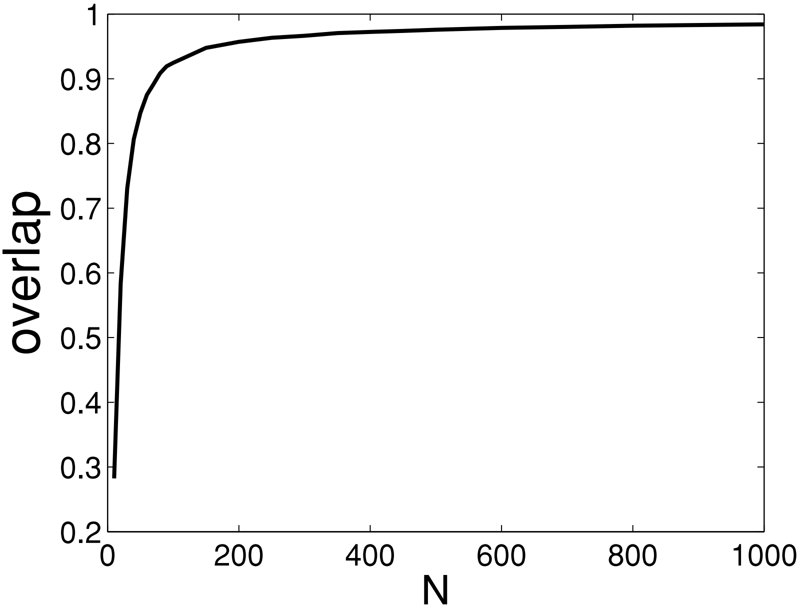

Figure 6: The average overlap between the optimal linear estimator and the zeroth approximation of the optimal linear estimator linear weights, *Real* $\{\ll \frac{\mathbf{w}_{ole}^{\dagger}\mathbf{w}^{(0)}}{\sqrt{\|\mathbf{w}_{ole}\|\|\mathbf{w}^{(0)}\|}} \gg\}$, is shown as a function of the system size N. Note that we find the imaginary part of the average overlap to be on the order of 10^{-3} and decreases rapidly with N (results not shown). Every point on the graph shows the overlap averaged over 400 realizations of the neuronal diversity.

The trial-to-trial fluctuations of the zeroth approximation obey the following statistics (see appendix D),

$$\ll \langle (\delta\hat{\theta})^2 \rangle \gg = \frac{a}{2N\kappa|\tilde{f}^{(1)}|^2} \tag{3.11}$$

$$\ll (\Delta \langle (\delta\hat{\theta})^2 \rangle)^2 \gg = \frac{\tilde{B}^{(0)} + \frac{1}{2}\tilde{B}^{(2)}}{2(N\kappa)^2|\tilde{f}^{(1)}|^4}, \tag{3.12}$$

where a is the single-cell response variance (see equation 1.14) and $B(\theta) \equiv C^2(\theta)$. Thus, the efficiency of the zeroth-order approximation scales linearly with the size of the system, yielding an estimation error on the order of $1/\sqrt{N}$. Figure 7 shows the efficiency of this readout as a function of the population size N; the open circles show the numerical evaluation of the efficiency, and the solid line shows the analytical results of substituting

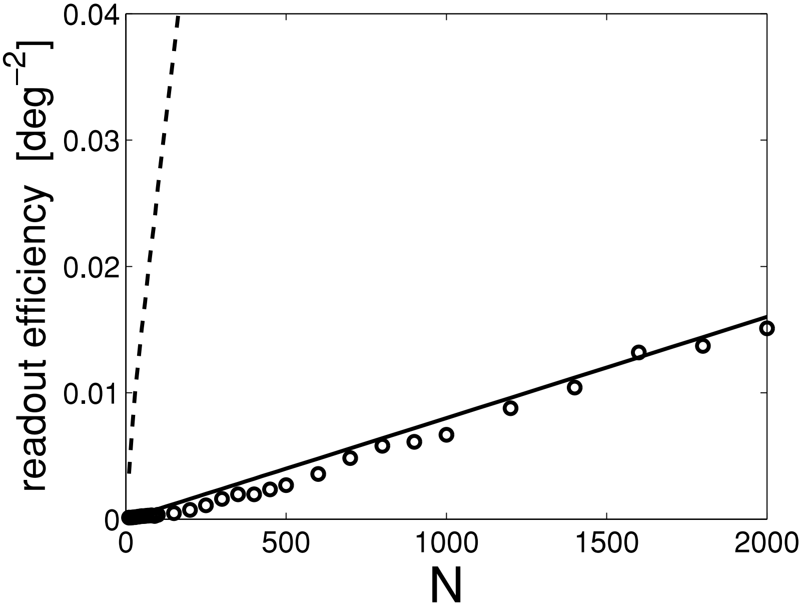

Figure 7: The efficiency of the zeroth-order approximation of the optimal linear estimator as a function of the population size, N. The efficiency is shown in terms of one over the average quadratic angular estimation error, $\ll \langle \hat{\theta} - \theta \rangle^2 \gg^{-1}$. The analytical results of equations 1.16, 3.10, and 3.11 are shown by the solid line. The open circles show the numerical estimation of the efficiency. For comparison, we present the optimal linear estimator efficiency (dashed line), as calculated numerically. The numerical calculation of the readouts efficiencies was done by averaging over 400 different realizations of the neuronal diversity and over 200 trials of simulating the neuronal stochastic responses to stimulus $\theta = 0$ for each realization.

equations 3.11 and 3.10 into equation 1.16. For comparison, the optimal linear estimator efficiency is shown by the dashed line. As can be seen from the figure, the efficiency of the zeroth approximation scales linearly with the size of the system, even in the presence of correlated noise. Nevertheless, its performance is considerably inferior to that of the optimal linear estimator. This result contrasts with the high degree of similarity between $\mathbf{w}^{(0)}$ and $\mathbf{w}_{ole}$ (see Figure 6), emphasizing the importance of the higher-order corrections to the optimal linear estimator readout. We find that by also incorporating the first-order corrections, we can retrieve most of the efficiency of the optimal linear estimator. However, the first-order corrections are nonlocal in space and involve global averages of the neuronal diversity across the entire population (results not shown). On the other hand, the analysis of the zeroth-order approximation efficiency is sufficient to prove the linear scaling of the optimal linear estimator efficiency with the population size. The problem of fine tuning of the optimal linear estimator weights is addressed in section 4.

4 Summary and Discussion

The efficiency of population codes has been the subject of considerable theoretical effort. Early studies investigated the efficiency of the code using the theoretical concept of the Fisher information and quantifying the accuracy of simple readout mechanisms (Paradiso, 1988; Seung & Sompolinsky, 1993). Assuming the trial-to-trial fluctuations in the responses of different cells have zero correlation, these studies have shown that the coding efficiency of the population grows linearly with the number of cells. Zohary et al. (1994) have shown the possible detrimental effect of nonzero cross-correlations on the coding efficiency of the population. On the other hand, differing results were presented by Abbott and Dayan (1999) claiming that correlations do not have a detrimental effect on the coding efficiency. Abbott and Dayan considered several models for the cross-correlations. One model incorporated a short-range correlation structure. In terms of the current study, this corresponds to scaling the correlation length, ρ, inversely with the number of cells in the population, $\rho \sim 1/N$ (see equation 1.14); hence, in the large N limit, this model is very similar to a model without correlations. Another interesting model studied by Abbott and Dayan is one with uniform correlations;[1] it corresponds to the other extreme of taking the limit of very large correlation length, $\rho \longrightarrow \infty$ (see equation 1.14). Uniform correlations generate large collective fluctuations in the neural responses. However, these collective fluctuations are limited to the uniform direction, $(1, 1, \ldots, 1)$, that contains no information about the stimulus identity. Thus, signal and noise in this model are completely segregated into orthogonal subspaces of the phase space of neural responses. Due to this segregation, one can treat this model as one of an independent population of neurons and obtain qualitatively similar results.

Qualitatively different results are obtained in the intermediate regime of $\rho = \mathcal{O}(1)$. In this regime, a considerable fraction of neuronal pairs shows significant correlations. Moreover, correlations are stronger for pairs of neurons with closer preferred directions. These properties of the $\rho = \mathcal{O}(1)$ regime are in agreement with experimental findings (see, e.g., Zohary et al., 1994; Lee et al., 1998). In this case, signal and noise are not segregated, and the collective fluctuations in the population response to the stimulus cause the

[1] Abbott and Dayan also considered a model in which information is coded by the correlated neuronal responses. This coding strategy is not the topic of our article and was addressed elsewhere (Shamir & Sompolinsky, 2004; see also Wu, Amari, & Nakahara, 2004).

saturation of the information capacity of the system. Typical values for the neural response properties and pairwise correlations yield an accuracy bound that is inconsistent with the known psychophysical accuracy (for a discussion in greater details, see Sompolinsky et al., 2001; Wu, Amari, & Nakahara, 2002). This raises the theoretical question of how a neural population with the experimentally observed correlation structure can overcome these collective fluctuations and obtain a more accurate code that will be able to account for psychophysical performance. One possible mechanism for overcoming the strong collective noise fluctuations is by utilizing the heterogeneity inherent in any neural population.

In this work, we studied the effect of neuronal heterogeneity on the coding capabilities of large populations of neurons with correlated firing rate fluctuations. The heterogeneity of the system enables information to be coded in all of the spatial modes of the network. The population vector readout extracts information only from the slowly varying collective modes of the system; hence, the optimal linear estimator yields superior performance to that of the population vector (see Figure 2). Moreover, the diversity of the population adds a bias to the population vector estimate of θ, thus decreasing its efficiency with respect to the homogeneous case (see Figure 2). However, in an uncorrelated population of neurons, the efficiency of both readouts scales linearly with the number of cells in the population (see Figure 2).

In a correlated population of neurons, in the biologically interesting regime of $\rho = \mathcal{O}(1)$, correlations generate large fluctuations in the slowly varying collective modes of the system. If the information coded by the neuronal responses resides only in this low-dimensional subspace of collective fluctuations, as in the homogeneous case, then the information capacity of the system will remain finite even in the limit of infinitely large networks (see Figure 3, bottom line). The neuronal diversity, applied here to the first-order statistics of the neuronal responses, allows information to be coded in other, more rapidly varying spatial modes of the system. In this case, signal and noise are not segregated; however, they are also not entirely overlapping, as in the homogeneous case. Consequently, the Fisher information of this system does not saturate to a finite limit; rather, it scales linearly with the population size (see equations 2.2–2.4 and Figure 3).

We further investigated the efficiency of the population vector and the optimal linear estimator in the case of correlated heterogeneous populations of neurons. It was shown that the population vector readout fails to extract the information coded by the neuronal diversity and that its efficiency is bound due to the correlated noise in the slowly varying collective modes of the system (see Figure 5). On the other hand, the optimal linear estimator can extract most of the information coded by the quenched fluctuations of the neuronal responses. A numerical study of the optimal linear estimator performance revealed that its efficiency scales linearly with the size of the system (see Figure 5). Note that for both correlated and uncorrelated cases, the optimal linear estimator bias has a negligible contribution to the error. Our study shows that for large N, the optimal linear estimator converges to the simple form of the zeroth approximation, $w_i^{(0)} \propto \varepsilon_i e^{i\phi_i}$ (see Figure 6). We have shown analytically that the efficiency of the zeroth approximation scales linearly with N, equations 3.9 to 3.12. However, its performance is considerably inferior to that of the optimal linear estimator (see Figure 7). These two findings highlight the sensitivity of the optimal linear estimator performance to the higher-order corrections and raise the question of whether the central nervous system can perform such an accurate readout.

This last question is addressed in the context of supervised learning of the optimal linear estimator weights. We assume the linear readout weights of the system are learned by an online learning process (e.g., Radons, 1993; Hansen, Pathria, & Salamon, 1993) and ask whether the performance of this readout will converge to that of the optimal linear estimator in a reasonable time. In every learning step, an example stimulus is chosen randomly from a uniform distribution on the circle, $\{\theta^k\}_{k=1}^{p} \sim$ i.i.d. $U([-\pi, \pi))$. The kth example stimulus, θ^k, generates a response of the neural population, $\mathbf{r}^{(k)}$, that is distributed according to $\mathbf{r}^{(k)} \sim P(\mathbf{r}^{(k)}|\theta^k)$, as defined in equation 1.1. Given the neural responses, the system calculates its estimation of the kth example, $\hat{z}^k = \mathbf{r}^{(k)\dagger}\mathbf{w}(k)$, and updates the readout weights according to the estimation error in the kth example. Formally, we define a momentary cost function,

$$E_k(\mathbf{w}) = \frac{1}{2}|\hat{z}^k - z^k|^2 = \frac{1}{2}|\mathbf{r}^{(k)\dagger}\mathbf{w} - e^{i\theta_k}|^2, \tag{4.1}$$

and the learning dynamics are defined by

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \eta(k)\frac{\partial E_k(\mathbf{w}(k))}{\partial \mathbf{w}} = \mathbf{w}(k) - \eta(k)(\hat{z}^k - z^k)\mathbf{r}^{(k)}, \tag{4.2}$$

where $\eta(k)$ is the momentary learning rate and $\frac{\partial}{\partial \mathbf{w}}$ is the gradient with respect to $\mathbf{w}$. Note that in Hebbian learning, the synaptic weight is modified in proportion to the product of its input and output. Here, in the case of supervised learning of a linear system, the update takes the form of the output error, $(\hat{z}^k - z^k)$, times the input, $\mathbf{r}^{(k)}$. Figure 8 shows the online learning curve of the optimal linear estimator for a population of $N = 400$ neurons (solid line). The Euclidean error, equation 1.19, of the zeroth-order approximation to the optimal linear estimator and of the population vector readout are plotted for comparison by the horizontal solid and dashed lines, respectively. As can be seen from the figure, the learning converges rather fast and presents a superior performance to that of the zeroth-order approximation after $p \sim N$ learning steps. Hence, despite the sensitivity of the optimal linear estimator to the higher-order, $\mathcal{O}(N^{-3/2})$, corrections, a

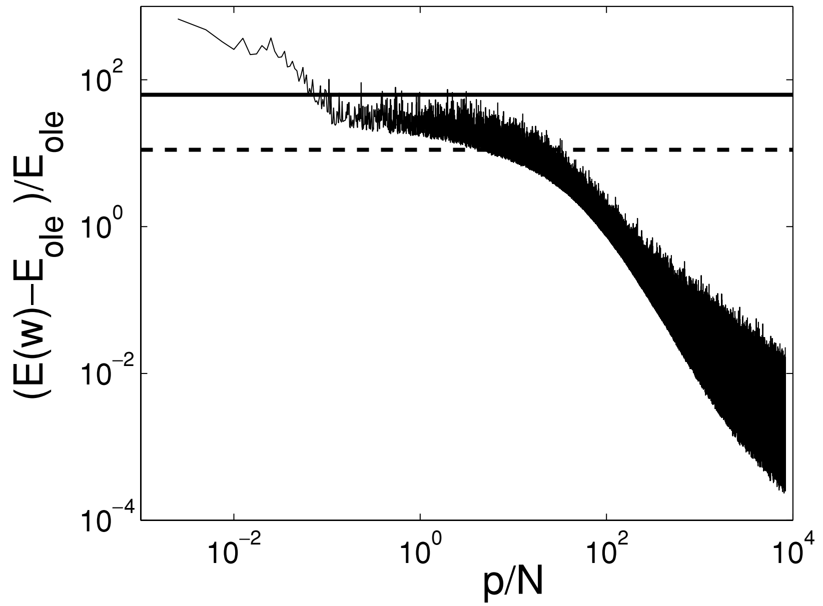

Figure 8: Online learning curve for the optimal linear estimator. Generalization error, equation 1.19, is plotted as a function of the number of examples shown to the system, p, per weight that needs to be learned, N. For comparison, we show the zeroth-order error and the population vector readout error by the horizontal solid and dashed lines, respectively. In this graph, $N = 400$ was used. In our learning dynamics, we have scaled the learning rate according to $\eta(k) = \frac{\eta_0}{1+k/k_0}$ with $\eta_0 = 5 \cdot 10^{-6}$ and $k_0 = 20{,}000$.

simple learning rule converges to the optimal linear estimator performance rather fast.

Recently, an interest in the learning of the optimal linear estimator has appeared in the context of constructing a brain machine interface (e.g., Wessberg et al., 2000; Carmena et al., 2003; Schwartz, Taylor, & Tillery, 2001). For this application, learning can be done using a batch learning algorithm (Seung, Tishby, & Sompolinsky, 1992). We found that similar to online learning, batch learning converges to the optimal linear estimator performance on the scale of several N examples, where N is the size of the system (results not shown). Wessberg et al. (2000) and later Carmena et al. (2003) applied a batch algorithm to learning of optimal linear estimator weights for predicting motor commands from cortical neuron activity (mainly from motor cortex). Wessberg et al. also investigated the scaling of the learned readout accuracy by the number of its input neurons. Their results suggest that the optimal linear estimator accuracy scales linearly with the number of pooled neurons. Their results are consistent with our finding on the scaling of the optimal linear estimator performance with the population size in a correlated heterogeneous network. Our results indicate that in order to obtain an accuracy of 5 degrees, similar to the psychophysical accuracy in a simple reaching task, it is sufficient to implement the optimal linear estimator readout on a population of several hundred neurons (see Figure 5).

The results of this work are not limited to the amplitude diversity model, equation 1.10, and can be generalized to other kinds of neuronal heterogeneities. Assuming a small measure of heterogeneity in the system, $\kappa \ll 1$, one can approximate the tuning curves by equation 1.9 and generalize the calculation of appendixes C and D, yielding a zeroth approximation to the optimal linear estimator of the form $w_j^{(0)} = \frac{1}{N\kappa\tilde{g}^{(1)}}\epsilon_j e^{i\phi_j}$ (see equation 1.9 for the definition of g). For example, one may consider the case of a population with diverse tuning curves widths. The tuning curve width of a specific cell, σ_i, can be characterized as a sum of the average width, $\ll \sigma \gg \equiv \bar{\sigma}$, and a fluctuation, $\sigma_i = \bar{\sigma} + \varepsilon_i$. Interestingly, in this case, one obtains that $\tilde{g}^{(1)} < 0$; hence, the zeroth-order approximation predicts that the optimal estimator will give a higher weight to neurons with a sharper tuning curve ($\varepsilon < 0$) than for neurons with broader-than-average tuning curves ($\varepsilon > 0$). Similarly, the calculation can be expanded to a model in which the tuning curve of every cell is characterized by a vector of heterogeneity parameters. Qualitatively, results are the same. This argument excludes diversity in the preferred directions of the cells. The deviations from homogeneity in the distribution of preferred directions in the population result from finite sampling size. This corresponds to scaling the measure of heterogeneity, κ, inversely with the number of cells, $\kappa \sim 1/N$. Hence, in the limit of large populations, this source of heterogeneity is not expected to have a significant contribution to the coding efficiency of the system. It is interesting to note, however, that one of the problems that motivated the work of Salinas and Abbott (1994), which introduced the concept of the optimal linear estimator, was the finite-size heterogeneity in the distribution of preferred directions. Further extensions of this theory for studying the coding of a more complex stimulus can be obtained once a biologically plausible model for the tuning and pairwise correlation of the neural responses is established.

In this work, we considered only variability in the first-order statistics of the neural responses. However, the second-order statistics of neural responses, namely, the firing rates covariance, is also very diverse (see, e.g., Zohary et al., 1994; Lee et al., 1998; Maynard et al., 1999). The way in which diversity in the second-order statistics affects our conclusions depends on the amount of overlap and segregation between eigenvectors of $\mathbf{C}$ corresponding to its largest eigenvalues and $\mathbf{f}'$. Unfortunately, the relation between the diversity of the first- and second-order statistics of the neural responses is not yet clear. One possibility is to consider independent

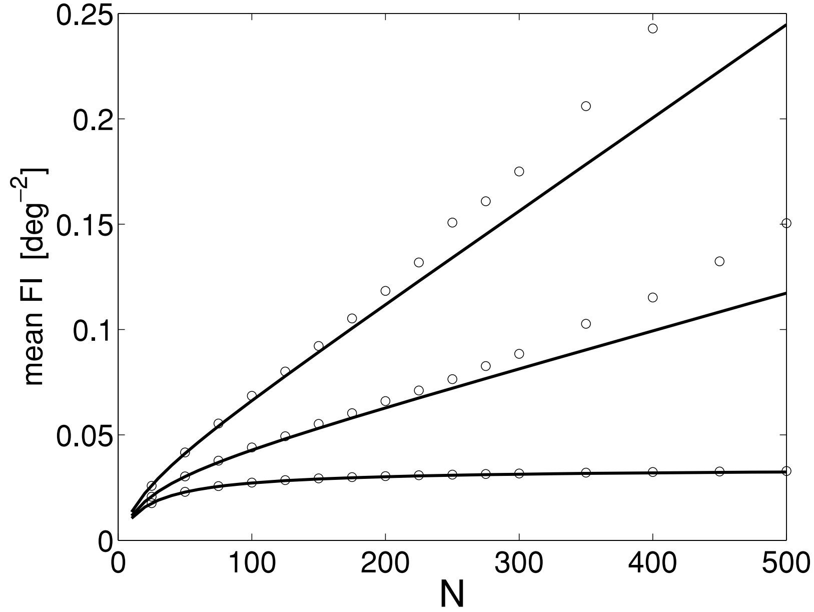

Figure 9: The mean Fisher information of a model with heterogeneity in the first- and second-order statistics. The quenched mean of the Fisher information, $\ll J \gg$, was calculated by averaging over the 100 realizations of the neuronal diversity (open circles) with $\Xi = 0.25$ in the amplitude diversity model with $\kappa = 0, 0.1, 0.25$ from bottom to top. The Fisher information was calculated for the stimulus value $\theta = 0$. For comparison, the average Fisher information of the system with $\Xi = 0$ is shown by the solid lines, as computed by equation 2.2.

additive variability to the covariance matrix. Assuming the covariance matrix obeys

$$C_{ij} = \bar{C}_{ij} + T_{ij} \tag{4.3}$$

$$\bar{C}_{ij} = \bar{C}(\phi_i - \phi_j) \tag{4.4}$$

$$T_{ij} = \xi_{ij} + \xi_{ji}, \tag{4.5}$$

where ξ is a random gaussian matrix, that is, the $\{\xi_{ij}\}$ are i.i.d. gaussian random variables with means 0 and variance Ξ that are also independent of the heterogeneity of the first-order statistics. Hence, $\mathbf{T}$ is a real symmetric matrix with maximum eigenvalue $2\sqrt{2N}\Xi$ (see, e.g., Mehta, 1991). The matrix $\bar{\mathbf{C}}$ obeys equation 1.14. Figure 9 shows the mean Fisher information of such a system with $\Xi = 0.25$ for different values of κ in the amplitude diversity model (open circles); for comparison, the results with $\Xi = 0$ are shown in solid lines. As can be seen from the figure, adding heterogeneity to the second-order statistics does not alter our previous results qualitatively. In particular, the Fisher information of a system with $\kappa = 0$ saturates to a finite limit, whereas the Fisher information of systems with $\kappa > 0$ grows linearly with the size of the system.

In this work, we emphasized the possible role of the "quenched" fluctuations (neuronal heterogeneity) in coding information in correlated populations of neurons. In another study, the stimulus-dependent "thermal" (trial-to-trial) fluctuations were suggested as a primary source of information (Shamir & Sompolinsky, 2001, 2004). Neural populations exhibit both stimulus-dependent thermal and quenched fluctuations. Studying the optimal linear estimator efficiency, equations 1.19 to 1.21, we find that only the stimulus average of the correlation matrix appears in $E(\mathbf{w})$. Hence, the performance of the optimal linear estimator is not affected by the tuning of the higher-order statistics. Similarly the contribution of the stimulus-dependent correlations to the Fisher information, $J_{cov} = Tr\{\mathbf{C}'\mathbf{C}^{-1}\mathbf{C}'\mathbf{C}^{-1}\}$ (where $\mathbf{C}' = \frac{d\mathbf{C}'}{d\theta}$; see Shamir & Sompolinsky 2004), is not affected by the heterogeneity of the first-order statistics of the neural responses. Thus, although thermal and quenched fluctuations are usually considered sources of noise in the system, our theory suggests that these fluctuations may have a central role in coding information in populations of neurons with correlated firing rate statistics. However, since the information capacity of both thermal and quenched stimulus-dependent fluctuations scales linearly with the population size, several decoding strategies are possible. Our current theory is unable to make a definitive claim about which strategy is most plausible. Nevertheless, in order to utilize neuronal diversity, fine tuning of the readout mechanism is required. On the other hand, our theory suggests that if neuronal diversity is not utilized by the readout, a nonlinear readout is essential to obtain a high degree of accuracy. Further theoretical as well as experimental effort is needed to provide additional tests for the alternative decoding schemes used in the brain.

Appendix A: Calculation of the Fisher Information Statistics

The Fisher information of a diverse population, equation 2.1, can be written as the sum of three terms: $J = I_1 + 2I_2 + I_3$, where

$$I_1 = \sum_{i,j=1}^{N} f_i' C_{ij}^{-1} f_j' \tag{A.1}$$

$$I_2 = \sum_{i,j=1}^{N} \Delta m_i' C_{ij}^{-1} f_j' \tag{A.2}$$

$$I_3 = \sum_{i,j=1}^{N} \Delta m_i' C_{ij}^{-1} \Delta m_j'. \tag{A.3}$$

The first term, I_1, is the Fisher information of an isotropic system. As discussed in section 1, in the biologically relevant parameter regime for the correlations, $\rho = \mathcal{O}(1)$ and $c = \mathcal{O}(1)$, $c > 0$, this term saturates to a finite limit even when the population size grows to infinity, yielding a contribution to J that is $\mathcal{O}(1)$. The second term, I_2, has zero mean, $\ll I_2 \gg = 0$, and variance

$$\ll (I_2)^2 \gg = \sum_{i=1}^{N} K(\phi_i - \theta) v_i^2 \tag{A.4}$$

$$v_i = \sum_{j=1}^{N} C_{ij}^{-1} f_j'. \tag{A.5}$$

Now, $\sum K^2(\phi_i - \theta) = \mathcal{O}(N)$. Since $\|\mathbf{v}\|^2 = \mathbf{f}'^T \mathbf{C}^{-2} \mathbf{f}' = \mathcal{O}(1/N)$ (eigenvalues of $\mathbf{C}^{-2}$ scales like N^{-2}), then $v_i = \mathcal{O}(1/\sqrt{N})$ $\forall i$, yielding $\sum_i v_i^4 = \mathcal{O}(1/N)$. Hence, applying the Cauchy-Schwartz inequality to equation 4.4, one obtains

$$\ll (I_2)^2 \gg \leq \sqrt{\sum_{i=1}^{N} K^2(\phi_i - \theta) \sum_{i=1}^{N} v^4} = \mathcal{O}(1). \tag{A.6}$$

The last term, I_3, has a mean that scales linearly with the size of the system, N,

$$\ll I_3 \gg = \sum_{i=1}^{N} K(\phi_i - \theta) C_{ii}^{-1} = N\bar{K}d, \tag{A.7}$$

where d is the diagonal element of $\mathbf{C}^{-1}$ and $\bar{K} = \int \frac{d\varphi}{2\pi} K(\varphi)$. Note that one can approximate d using the fact that the eigenvalue spectrum of the correlation matrix, $\mathbf{C}$, contains a small number, $p(N) = o(N)$, of low Fourier modes with eigenvalues scaling like N, while the rest of the spectrum rapidly decays to $a(1-c)$. Hence, for large N, one can approximate $C_{ii}^{-1} = \frac{1}{a(1-c)} + O(p(N)/N)$, $\forall i \in \{1, \ldots, N\}$. In this model, the eigenvalue spectrum of $\mathbf{C}$ decays algebraically with the Fourier mode, $n, C_n \propto n^{-2}$; hence, $p(N) \propto \sqrt{N}$, and in the limit of large N, one can neglect the contribution of $\mathcal{O}(p(N)/N)$ to the diagonal of $\mathbf{C}^{-1}$.

The variance of I_3 is given by

$$\ll (\Delta I_3)^2 \gg = 2\sum_{ij} K(\phi_i - \theta)K(\phi_j - \theta)(C_{ij}^{-1})^2$$
$$= 2\sum_{i=1}^{N} K^2(\phi_i - \theta)(C_{ii}^{-1})^2 + 2\sum_{i\neq j} K(\phi_i - \theta)K(\phi_j - \theta)(C_{ij}^{-1})^2. \tag{A.8}$$

The last term in the right-hand side of equation A.8 can be bound by $\max\{K\}^2 \sum_{i\neq j}(C_{ij}^{-1})^2$. For a long-range correlation matrix with a strongly decaying eigenvalue spectrum, we can approximate the off-diagonal terms of $\mathbf{C}^{-1}$ by

$$C_{kj}^{-1} \approx -\frac{1}{Na(1-c)} \sum_{|n|<p(N)} e^{in(\phi_k - \phi_j)} \quad (k \neq j). \tag{A.9}$$

Using this approximation, one obtains $\sum_{i\neq j}(C_{ij}^{-1})^2 \propto p(N) = o(N)$. Hence, for large N, to a leading order in N, the quenched fluctuations in I_3 are given by

$$\ll (\Delta I_3)^2 \gg = 2N\overline{K^2}d^2, \tag{A.10}$$

where $\overline{K^2} = \int \frac{d\varphi}{2\pi} K^2(\varphi)$. In corollary of the above, the Fisher information of such a system is a self-averaging quantity with a quenched mean and variance that scale linearly with the population size.

Appendix B: Statistics of the Population Vector Efficiency

B.1 Average Euclidean Error. The population vector linear readout weights are given by $w_j = \frac{\gamma}{N} e^{i\phi_j}$ with[2] $\gamma = \frac{\tilde{f}^{(1)}}{|\tilde{f}^{(1)}|^2 + \tilde{c}_1}$. In the amplitude variability model, equation 1.10, one obtains

$$U_j = (1 + \varepsilon_j)\tilde{f}^{(1)} e^{i\phi_j} \tag{B.1}$$

$$Q_{ij} = C_{ij} + F_{ij} \tag{B.2}$$

$$F_{ij} = (1 + \varepsilon_i)(1 + \varepsilon_j) \int \frac{d\theta}{2\pi} f(\phi_i - \theta) f(\phi_j - \theta). \tag{B.3}$$

[2] Note that $\tilde{f}^{(1)}$ and hence also γ are real.

For the calculation of $E(PV)$, equation 1.19, we need to compute the following terms:

$$\mathbf{U}^{\dagger}\mathbf{w} = \sum_{j=1}^{N} \tilde{f}^{(1)}(1+\varepsilon_j)e^{-i\phi_j}\frac{\gamma}{N}e^{i\phi_j}$$

$$= \tilde{f}^{(1)}\gamma\frac{1}{N}\sum_{j=1}^{N}(1+\varepsilon_j) = |\gamma\tilde{f}^{(1)}|\,(1+I_1) \tag{B.4}$$

$$\mathbf{w}^{\dagger}\mathbf{C}\mathbf{w} = \gamma^2\tilde{c}_1 \tag{B.5}$$

$$\mathbf{w}^{\dagger}\mathbf{F}\mathbf{w} = \frac{\gamma^2}{N^2}\sum_{j,k=1}^{N}(1+\varepsilon_j)(1+\varepsilon_k)e^{i(\phi_j-\phi_k)}\int\frac{d\theta}{2\pi}f(\phi_j-\theta)f(\phi_k-\theta)$$

$$= \gamma^2\int\frac{d\theta}{2\pi}\left|\frac{1}{N}\sum_{j=1}^{N}f(\phi_j-\theta)(1+\varepsilon_j)e^{i\phi_j}\right|^2$$

$$= \gamma^2\int\frac{d\theta}{2\pi}\left|\tilde{f}^{(1)}+\frac{1}{N}\sum_{j=1}^{N}f(\phi_j-\theta)\varepsilon_j e^{i\phi_j}\right|^2$$

$$= |\gamma\tilde{f}^{(1)}|^2 + 2|\gamma\tilde{f}^{(1)}|^2 I_1 + \gamma^2 I_2, \tag{B.6}$$

where we have used in equation B.5 the fact that **w** is an eigenvector of the correlation matric **C** with an eigenvalue of $N\tilde{c}_1$. The terms I_1 and I_2 are defined by

$$I_1 = \frac{1}{N}\sum_{i=1}^{N}\epsilon_i \tag{B.7}$$

$$I_2 = \frac{1}{N^2}\sum_{kj}\epsilon_k\epsilon_j\int\frac{d\theta}{2\pi}f(\phi_k-\theta)f(\phi_j-\theta)e^{i(\phi_k-\phi_j)}. \tag{B.8}$$

Substituting the above in equation 1.19 and expressing γ in terms of the Fourier transforms of the average tuning curve and of the correlation matrix, we obtain

$$E(PV) = 1 - \frac{|\tilde{f}^{(1)}|^2}{|\tilde{f}^{(1)}|^2+\tilde{c}_1} - 2\frac{|\tilde{f}^{(1)}|^2\tilde{c}_1}{\left(|\tilde{f}^{(1)}|^2+\tilde{c}_1\right)^2}I_1 + \frac{|\tilde{f}^{(1)}|^2}{\left(|\tilde{f}^{(1)}|^2+\tilde{c}_1\right)^2}I_2. \tag{B.9}$$

Now I_1 is a gaussian random variable that fluctuates from one realization of the neural population to another with $\ll I_1 \gg = 0$ and $\ll (I_1)^2 \gg = \frac{\kappa}{N}$.

The random variable I_2 obeys the following statistics,

$$\ll I_2 \gg = \frac{\kappa}{N}\widetilde{f^2}_{(0)} = \mathcal{O}(\kappa/N) \tag{B.10}$$

$$\ll (\Delta I_2)^2 \gg = \frac{\kappa^2}{N^2}\left(\widetilde{G^2}_{(0)} + \widetilde{G^2}_{(2)}\right) = \mathcal{O}(\kappa^2/N^2), \tag{B.11}$$

with $G(\phi) = \int \frac{d\theta}{2\pi} f(\theta) f(\theta - \phi)$, $\widetilde{G^2}_{(n)} = \int \frac{d\varphi}{2\pi} G^2(\varphi) e^{in\varphi}$, and $\widetilde{f^2}_{(n)} = \int \frac{d\varphi}{2\pi} f^2(\varphi) e^{in\varphi}$. Hence, $E(PV)$ is a self-averaging quantity of $\mathcal{O}(1)$ with fluctuations that are $\mathcal{O}(1/\sqrt{N})$:

$$E(PV) = \frac{1}{1 + |\tilde{f}^{(1)}|^2/\tilde{c}_1} + \mathcal{O}(1/\sqrt{N}) \tag{B.12}$$

B.2 Angle Estimation Error. Let $\hat{x}$ and $\hat{y}$ be the real and imaginary parts of the population vector, respectively:

$$\hat{z} = \hat{x} + i\hat{y} = \mathbf{w}^T \mathbf{r}. \tag{B.13}$$

The population vector angular estimator is given by

$$\hat{\theta}(\hat{x}, \hat{y}) = \arctan\left(\frac{\hat{y}}{\hat{x}}\right). \tag{B.14}$$

For small, angular estimation errors, we can expand $\hat{\theta}$ in powers of $\delta\hat{x}$ and $\delta\hat{y}$ around their means and approximate

$$\hat{\theta} = \hat{\theta}(\langle\hat{x}\rangle, \langle\hat{y}\rangle) + \frac{\partial\hat{\theta}(\langle\hat{x}\rangle, \langle\hat{y}\rangle)}{\partial\hat{x}}\delta\hat{x} + \frac{\partial\hat{\theta}(\langle\hat{x}\rangle, \langle\hat{y}\rangle)}{\partial\hat{y}}\delta\hat{y}. \tag{B.15}$$

The first term in the right-hand side of equation B.15 yields the bias; the two other terms provide the trial-to-trial fluctuations of the estimator. Thus, to the lowest order in fluctuations at $\theta = 0$, the bias and variance of $\hat{\theta}$ are given by

$$\langle\hat{\theta}\rangle = \arctan\left(\frac{\langle\hat{y}\rangle}{\langle\hat{x}\rangle}\right) \approx \frac{\langle\hat{y}\rangle}{\langle\hat{x}\rangle} \tag{B.16}$$

$$\langle(\delta\hat{\theta})^2\rangle = \frac{\langle(\delta\hat{y})^2\rangle}{\langle\hat{x}\rangle^2}. \tag{B.17}$$

Computing the statistics of $\hat{x}$ and $\hat{y}$, one obtains

$$\ll \langle \hat{x} \rangle \gg = \gamma \tilde{f}^{(1)} \tag{B.18}$$

$$\ll (\Delta \langle \hat{x} \rangle)^2 \gg = \frac{\gamma^2 \kappa}{2N} \left(\widetilde{f^2}_{(0)} + \widetilde{f^2}_{(2)} \right). \tag{B.19}$$

Thus, $\langle \hat{x} \rangle$ is a self-averaging quantity with respect to the quenched fluctuations. For $\hat{y}$, we obtain

$$\ll \langle \hat{y} \rangle \gg = 0 \tag{B.20}$$

$$\ll (\Delta \langle \hat{y} \rangle)^2 \gg = \frac{\gamma^2 \kappa}{2N} \left(\widetilde{f^2}_{(0)} - \widetilde{f^2}_{(2)} \right) \tag{B.21}$$

$$\langle (\delta \hat{y})^2 \rangle = \frac{1}{2} \gamma^2 \tilde{c}_1. \tag{B.22}$$

Note that the trial-to-trial fluctuations of the population vector readout, reflected here by $\langle (\delta \hat{y})^2 \rangle$, have zero variability with respect to the quenched statistics. Summarizing the above results yields, for the bias,

$$\ll b_{pv} \gg = \ll \frac{\langle \hat{y} \rangle}{\langle \hat{x} \rangle} \gg = \frac{\ll \langle \hat{y} \rangle \gg}{\ll \langle \hat{x} \rangle \gg} = 0 \tag{B.23}$$

$$\ll b_{pv}^2 \gg = \frac{\ll (\Delta \langle \hat{y} \rangle)^2 \gg}{\ll \langle \hat{x} \rangle \gg^2} = \frac{\kappa}{2N} \frac{\widetilde{f^2}_{(0)} - \widetilde{f^2}_{(2)}}{|\tilde{f}^{(1)}|^2}. \tag{B.24}$$

Note that in the above equations, we have used the self-averaging property of $\langle \hat{x} \rangle$, neglecting its quenched fluctuations. For the population vector variance, one obtains

$$\ll \langle (\delta \hat{\theta}_{pv})^2 \rangle \gg = \frac{\ll \langle (\delta \hat{y})^2 \rangle \gg}{\ll \langle \hat{x} \rangle^2 \gg} = \frac{\tilde{c}_1}{2|\tilde{f}^{(1)}|^2}. \tag{B.25}$$

Note that since $\langle (\delta \hat{y})^2 \rangle$ has zero sample-to-sample variability and $\langle \hat{x} \rangle$ is a self-averaging quantity, then $\langle (\delta \hat{\theta}_{pv})^2 \rangle$ is a self-averaging quantity. To summarize the above results, the population vector readout in a diverse population has a small bias that decays with the increase of the pool size and trial-to-trial fluctuations that remain $\mathcal{O}(1)$ even in the limit of infinitely large neuronal populations.

Appendix C: The Zeroth-Order Approximation to the Optimal Linear Estimator

The main difficulty in calculating the optimal linear estimator, equations 3.4 to 3.6, is the inversion of the random matrix $\mathbf{Q}$. It is useful to first study how $\mathbf{Q}$ acts as a linear transformation. Let us denote $v_j^n = \frac{1}{N} e^{in\phi_j}$ and $t_j^n = \frac{1}{N}\varepsilon_j e^{in\phi_j}$. Note that $\mathbf{U} = N\tilde{f}^{(1)}(\mathbf{v}^1 + \mathbf{t}^1)$,

$$\sum_{j=1}^{N} Q_{ij} v_j^n = N(\tilde{c}_n + |\tilde{f}^{(n)}|^2) v_i^n + N|\tilde{f}^{(n)}|^2 t_i^n$$
$$+ \sum_m S_1^{(n-m)} |\tilde{f}^{(m)}|^2 (v_i^m + t_i^m) \tag{C.1}$$

$$\sum_{j=1}^{N} Q_{ij} t_j^n = N\kappa |\tilde{f}^{(n)}|^2 (t_i^n + v_i^n) + \sum_m \tilde{c}_m S_1^{(n-m)} v_i^n$$
$$+ \sum_m |\tilde{f}^{(m)}|^2 \left(S_1^{(n-m)} + S_2^{(n-m)} \right) (v_i^m + t_i^m), \tag{C.2}$$

where $S_1^{(n)} = \sum_{j=1}^{N} \varepsilon_j e^{in\phi_j}$ and $S_2^{(n)} = \sum_{j=1}^{N} \Delta\varepsilon_j^2 e^{in\phi_j}$. Note that S_1^n and S_2^n are of order $\sqrt{N}$. Hence, to a leading order in N, we can treat the subspace spanned by $\{\mathbf{v}^{(1)}, \mathbf{t}^{(1)}\}$ as an invariant subspace of $\mathbf{Q}$. Let us denote by $\hat{\mathbf{x}} = \binom{\hat{x}_1}{\hat{x}_2}$ the reduction of an N-dimensional vector $\mathbf{x}$ to the subspace spanned by $\{\mathbf{v}^{(1)}, \mathbf{t}^{(1)}\}$, that is, $(\hat{x}_1 \mathbf{v}^1 + \hat{x}_2 \mathbf{t}^1)$ is equal to the projection of $\mathbf{x}$ on the two-dimensional subspace. With this notation, $\hat{\mathbf{U}} = N\tilde{f}^{(1)}\binom{1}{1}$. The reduction of $\mathbf{Q}$ is given by

$$\hat{\mathbf{Q}} = N \begin{bmatrix} \tilde{c}_1 + |\tilde{f}^{(1)}|^2 & \kappa|\tilde{f}^{(1)}|^2 \\ |\tilde{f}^{(1)}|^2 & \kappa|\tilde{f}^{(n)}|^2 \end{bmatrix} + \mathcal{O}(\sqrt{N}). \tag{C.3}$$

Thus, to a leading order in $1/N$, one obtains $\hat{\mathbf{Q}}^{-1}\hat{\mathbf{U}} = \frac{1}{\kappa \tilde{f}^{(1)}}\binom{0}{1}$, and hence to this order of approximation,

$$w_{ole,j} = \frac{1}{N\kappa \tilde{f}^{(1)}} \varepsilon_j e^{i\phi_j} + \mathcal{O}(N^{-3/2}). \tag{C.4}$$

Appendix D: Statistics of the Efficiency of the Zeroth Approximation to the Optimal Linear Estimator

D.1 The Euclidean Error. The zeroth approximation readout weights are given by $w_j^{(0)} = \frac{1}{N\kappa \tilde{f}^{(1)}} \varepsilon_i e^{i\phi_j}$. For the calculation of $E^{(0)} \equiv E(\mathbf{w}^{(0)})$ (see equation 1.19), we need to compute the following terms:

$$\mathbf{U}^{\dagger}\mathbf{w}^{(0)} = \mathbf{w}^{\dagger}\mathbf{U} = 1 + \frac{1}{N\kappa}\sum_{i=1}^{N}(\Delta\varepsilon_i^2 + \varepsilon_i) \tag{D.1}$$

$$\mathbf{w}^{(0)\dagger}\mathbf{C}\mathbf{w}^{(0)} = \frac{1}{\kappa^2|\tilde{f}^{(1)}|^2} I_1 \tag{D.2}$$

$$\mathbf{w}^{(0)\dagger}\mathbf{F}\mathbf{w}^{(0)} = 1 + \frac{2}{N\kappa}\sum_{i=1}^{N}(\Delta\varepsilon_i^2 + \varepsilon_i) + \frac{1}{\kappa^2|\tilde{f}^{(1)}|^2}\left(I_2 + 2Real\{I_3\} + I_4\right), \tag{D.3}$$

where

$$I_1 = \frac{1}{N^2}\sum_{jk}\varepsilon_j\varepsilon_k e^{i(\phi_j-\phi_k)}C_{jk} \tag{D.4}$$

$$I_2 = \frac{1}{N^2}\sum_{jk}\varepsilon_j\varepsilon_k\int\frac{d\theta}{2\pi}f(\phi_j-\theta)f(\phi_k-\theta)e^{i(\phi_j-\phi_k)} \tag{D.5}$$

$$I_3 = \frac{1}{N^2}\sum_{jk}\varepsilon_j\Delta\varepsilon_k^2\int\frac{d\theta}{2\pi}f(\phi_j-\theta)f(\phi_k-\theta)e^{i(\phi_j-\phi_k)} \tag{D.6}$$

$$I_4 = \frac{1}{N^2}\sum_{jk}\Delta\varepsilon_j^2\Delta\varepsilon_k^2\int\frac{d\theta}{2\pi}f(\phi_j-\theta)f(\phi_k-\theta)e^{i(\phi_j-\phi_k)}. \tag{D.7}$$

The statistics of I_1 are given by $\ll I_1 \gg = \frac{\kappa a}{N}$ and $\ll (\Delta I_1)^2 \gg = \frac{\kappa^2}{N^2}(\tilde{B}^{(0)} + \tilde{B}^{(2)})$, where $\tilde{B}^{(n)} = \int\frac{d\phi}{2\pi}C^2(\phi)e^{in\phi}$. For I_2, I_3, and I_4, one obtains: $\ll I_2 \gg = \frac{\kappa}{N}\widetilde{f^2}_{(0)}$, $\ll I_3 \gg = 0$, $\ll I_4 \gg = 2\frac{\kappa^2}{N}\widetilde{f^2}_{(0)}$ for the means, and for the variances:

$$\ll (\Delta I_2)^2 \gg = \frac{\kappa^2}{N^2}(\widetilde{G^2}_{(0)} + \widetilde{G^2}_{(2)}) \tag{D.8}$$

$$\ll (\Delta 2Real\,\{I_3\})^2 \gg = 8\frac{\kappa^3}{N^2}(\widetilde{G^2}_{(0)} + \widetilde{G^2}_{(2)}) + \mathcal{O}(N^{-3}) \tag{D.9}$$

$$\ll (\Delta I_4)^2 \gg = 4\frac{\kappa^4}{N^2}(\widetilde{G^2}_{(0)} + \widetilde{G^2}_{(2)}) + \mathcal{O}(N^{-3}), \tag{D.10}$$

where $G(\varphi) = f * f(\varphi) \equiv \int\frac{d\theta}{2\pi}f(\theta)f(\varphi-\theta)$. Substituting the above in equation 1.19 yields

$$E^{(0)} = \frac{1}{\kappa^2|\tilde{f}^{(1)}|^2}\left(I_1 + I_2 + 2Real\,\{I_3\} + I_4\right). \tag{D.11}$$

Hence, $E^{(0)}$ is a random variable with quenched fluctuations with mean and standard deviation that are $\mathcal{O}(1/N)$,

$$\ll E^{(0)} \gg = \frac{1}{N\kappa|\tilde{f}^{(1)}|^2}\left(a + \widetilde{f^2}_{(0)}[1+2\kappa]\right) \tag{D.12}$$

$$\ll (\Delta E^{(0)})^2 \gg = \mathcal{O}(1/N^2) \tag{D.13}$$

where we have used $\ll \Delta I_x \Delta I_y \gg \leq \sqrt{\ll (\Delta I_x)^2 \gg \ll (\Delta I_y)^2 \gg}$ in the last inequality.

D.2 Angle Estimation Error. Let $\hat{x}$ and $\hat{y}$ denote the real and imaginary parts, respectively, of the zeroth approximation estimator: $\hat{z} = \mathbf{r}^T\mathbf{w}^{(0)} = \hat{x} + i\,\hat{y}$. Assuming small angular estimation errors, we use equations B.14 to B.17 to calculate the statistics of the zeroth approximation, at $\theta = 0$. As above, after averaging over the statistics of the neuronal diversity, the results are independent of the specific choice of $\theta = 0$. The statistics of $\hat{x}$ and $\hat{y}$ are given by

$$\ll \langle \hat{x} \rangle \gg = 1 \tag{D.14}$$

$$\ll (\Delta \langle \hat{x} \rangle)^2 \gg = \frac{1+\frac{1}{2\kappa}}{N}\,\frac{\widetilde{f^2}_{(0)} + \widetilde{f^2}_{(2)}}{|\tilde{f}^{(1)}|^2}. \tag{D.15}$$

Hence, $\langle \hat{x} \rangle$ is a self-averaging quantity with respect to the quenched disorder. The statistics of $\hat{y}$ are given by

$$\ll \langle \hat{y} \rangle \gg = 0 \tag{D.16}$$

$$\ll (\Delta \langle \hat{y} \rangle)^2 \gg = \frac{1+\frac{1}{2\kappa}}{N}\,\frac{\widetilde{f^2}_{(0)} - \widetilde{f^2}_{(2)}}{|\tilde{f}^{(1)}|^2} \tag{D.17}$$

$$\ll \langle (\delta \hat{y})^2 \rangle \gg = \frac{a}{2N\kappa|\tilde{f}^{(1)}|^2} \tag{D.18}$$

$$\ll (\Delta \langle (\delta \hat{y})^2 \rangle)^2 \gg = \frac{\tilde{B}^{(0)} + \frac{1}{2}\tilde{B}^{(2)}}{2N^2\kappa^2|\tilde{f}^{(1)}|^4}. \tag{D.19}$$

Summarizing the above results yields for the bias

$$\ll \langle \hat{\theta} \rangle \gg = 0 \tag{D.20}$$

$$\ll (\Delta \langle \hat{\theta} \rangle)^2 \gg = \frac{1+\frac{1}{2\kappa}}{N}\,\frac{\widetilde{f^2}_{(0)} - \widetilde{f^2}_{(2)}}{|\tilde{f}^{(1)}|^2}, \tag{D.21}$$

where we have used the self-averaging property of $\langle \hat{x} \rangle$, neglecting its quenched fluctuations. For the trial-to-trial variability of the angle estimation, one obtains

$$\ll \langle(\delta\hat{\theta})^2\rangle \gg = \frac{a}{2N\kappa|\tilde{f}^{(1)}|^2} \tag{D.22}$$

$$\ll (\Delta\langle(\delta\hat{\theta})^2\rangle)^2 \gg = \frac{\tilde{B}^{(0)} + \frac{1}{2}\tilde{B}^{(2)}}{2N^2\kappa^2|\tilde{f}^{(1)}|^4}. \tag{D.23}$$

Hence, the zeroth approximation of the optimal linear estimator has bias and trial-to-trial fluctuations that are of order $\mathcal{O}(1/\sqrt{N})$.

Acknowledgments

This research is partially supported by the Israel Science Foundation, Center of Excellence grant 8006/00 and by the USA-Israel Binational Science Foundation. M.S. was supported in part by a scholarship from the Clore Foundation and by the Burroughs Wellcome Fund.

References

Abbott, L. F., & Dayan P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Comput., 11*(1), 91–101.

Carmena, J. M., Lebedev, M. A., Crist R. E., O'Doherty J. E., Santucci, D. M., Dimitrov, D., Patil, P. G., Henriquez, C. S., & Nicolelis, M. A. (2003). Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biol., 1*(2), 193–207.

Coltz, J. D., Johnson, M. T., & Ebner, T. J. (2000). Population code for tracking velocity based on cerebellar Purkinje cell simple spike firing in monkeys. *Neurosci. Lett., 296*(1), 1–4.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci., 2*(11), 1527–1537.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416–1419.

Hansen, L. K., Pathria R., & Salamon, P. (1993). Stochastic dynamics of supervised learning. *J. Phys. A: Math. Gen., 26*(1), 63–71.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol., 160*, 106–154.

Kay, S. M. (1993). *Fundamentals of statistical signal processing.* Upper Saddle River, NJ: Prentice Hall.

Lee, D., Port, N. L., Kruse, W., & Georgopoulos, A. P. (1998). Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. *J. Neurosci., 18*(3), 1161–1170.

Mastronarde, D. N. (1983). Correlated firing of cat retinal ganglion cells. II. Responses of X- and Y-cells to single quantal events. *J. Neurophysiol., 49*(2), 325–349.

Maynard, E. M., Hatsopoulos, N. G., Ojakangas, C. L., Acuna, B. D., Sanes, J. N., Normann R. A., & Donoghue, J. P. (1999). Neuronal interactions improve cortical population coding of movement direction. *J. Neurosci., 19*(18), 8083–8093.

Mehta, L. M. (1991). *Random matrices* (2nd ed.). San Diego, CA: Academic Press.

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

Radons, R. (1993). On stochastic dynamics of supervised learning. *J. Phys. A: Math. Gen., 26*(14), 3455–3461.

Razak, K. A., & Fuzessery, Z. M. (2002). Functional organization of the pallid bat auditory cortex: Emphasis on binaural organization. *J. Neurophysiol., 87*(1), 72–86.

Ringach, D. L., Shapley, R. M., & Hawken, M. J. (2002). Orientation selectivity in macaque V1: Diversity and laminar dependence. *J. Neurosci., 22*(13), 5639–5651.

Salinas, E., & Abbott, L. F. (1994). Vector reconstruction from firing rates. *J. Comp. Neurosci., 1* (1–2), 89–107.

Schwartz, A. B., Taylor, D. M., & Tillery, S. I. (2001). Extraction algorithms for cortical control of arm prosthetics. *Curr. Opin. Neurobiol., 11*(6), 701–707.

Seung H. S., & Sompolinsky H. (1993). Simple models for reading neuronal population codes. *Proc. Natl. Acad. Sci. USA, 90*(22), 10749–10753.

Seung H. S., Tishby N., & Sompolinsky H. (1992). Statistical mechanics of learning from examples. *Phys. Rev. A, 45*(8), 6056–6091.

Shamir, M., & Sompolinsky, H. (2001). Correlation codes in neuronal networks. In D. G. Thomas, B. Suzanna, & G. Zoubin (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

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

Sompolinsky, H., Yoon, H., Kang, K., & Shamir, M. (2001). Population coding in neuronal systems with correlated noise. *Phys. Rev. E, 64*(5 Pt. 1), 051904.

Thomas, J. A., & Cover, T. M. (1991). *Elements of information theory*. New York: Wiley.

van Kan, P. L., Scobey, R. P., & Gabor A. J. (1985). Response covariance in cat visual cortex. *Exp. Brain. Res., 60*(3), 559–563.

Wessberg, J., Stambaugh, C. R., Kralik, J. D., Beck, P. D., Laubach, M., Chapin, J. K., Kim, J., Biggs, S. J., Srinivasan, M. A., & Nicolelis, M. A. (2000). Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature, 408*(6810), 361–365.

Wu, S., Amari, S., & Nakahara, H. (2002). Population coding and decoding in a neural field: A computational study. *Neural Comput., 14*(5), 999–1026.

Wu, S., Amari, S., & Nakahara, H. (2004). Information processing in a neuron ensemble with the multiplicative correlation structure. *Neural Netw., 17*(2), 205–214.

Yoon, H., & Sompolinsky, H. (1999). The effect of correlations on the Fisher information of population codes. In M. J. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems, 11*. Cambridge, MA: MIT Press.

Zohary, E., Shadlen, M. N., & Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature, 370*(6485), 140–143.

Received May 26, 2005; accepted January 10, 2006.