Unconscious biases in neural populations coding multiple stimuli

Sander W. Keemink[1,2], Dharmesh V. Tailor[1] and Mark C.W. van Rossum[3*]

July 2, 2018

[1]Institute for Adaptive and Neural Computation, School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK

[2]Bernstein Center Freiburg, Faculty of Biology, University of Freiburg, Hansastr. 9a, 79104 Freiburg, Germany

[3]School of Psychology and School of Mathematical Sciences, University of Nottingham, Nottingham NH7 2RD, UK

* Corresponding author

swkeemink@scimail.eu, dharmesh.tailor@live.co.uk,

mark.vanrossum@nottingham.ac.uk

Abstract

Throughout the nervous system information is commonly coded in activity distributed over populations of neurons. In idealized situations where a single, continuous stimulus is encoded in a homogeneous population code, the value of the encoded stimulus can be read out without bias. However in many situations multiple stimuli are simultaneously present, for example, multiple motion patterns might overlap. Here we find that when

multiple stimuli that overlap in their neural representation are simultaneously encoded in the population, biases in the read-out emerge. Although the bias disappears in the absence of noise, the bias is remarkably persistent at low noise levels. The bias can be reduced by competitive encoding schemes or by employing complex decoders. To study the origin of the bias, we develop a novel general framework based on Gaussian Processes, that allows for an accurate calculation of the estimate distributions of maximum likelihood decoders, and reveals that the distribution of estimates is bimodal for overlapping stimuli. The results have implications for neural coding and behavioral experiments on, for instance, overlapping motion patterns.

Introduction

In many brain areas information is distributed across neurons using population codes in which many neurons respond collectively to a single stimulus. By pooling across neurons, population codes allow for accurate estimation of a stimulus from the population response even when neural noise is present. Given its ubiquity, understanding population coding is believed to be crucial to understand coding of information in the brain. Numerous studies have quantified, among other issues, the role of the tuning curves (Zhang and Sejnowski, 1999), noise-correlations (Sompolinsky et al., 2002; Moreno-Bote et al., 2014), heterogeneity (Shamir and Sompolinsky, 2006; Ecker et al., 2011; Shamir, 2014), and stimulus multiplicity (Orhan and Ma, 2015) on the coding accuracy.

However, coding accuracy as measured by the variance in the estimates is not the only performance metric. When the same stimulus is repeatedly estimated from a population response and these estimates are averaged over many trials, a systematic difference between the mean estimated value and its true value might remain; this is called bias.

In many idealized cases biases are absent from population coding estimation schemes. First, in the limit of low noise, estimators such as the maximum likelihood decoder can be shown to be unbiased under quite general conditions (Kay, 1993). Secondly, the coding problem might have an intrinsic symmetry that abolishes bias, that is, over- and underestimation of the stimulus are equally likely, e.g. the estimation of the orientation of a visual grating from a homogeneous population using a homogeneous decoder. Either condition by itself is sufficient to warrant unbiased estimation. For instance, while for one dimensional direction estimates the maximum likelihood decoder is sub-optimal at high noise, it remains unbiased (Xie, 2002).

Yet, in perception biases are common. To explain these, theoretical studies typically rely on mechanisms that modulate the neural response to break the homogeneity of the population without adjusting the decoder, such as occurs with adaptation (e.g. Stocker and Simoncelli, 2006; Seriès, Stocker, and Simoncelli, 2009; Cortes et al., 2012; Wei and Stocker, 2015) or with contextual changes in the neural tuning (e.g. Schwartz, Hsu, and Dayan, 2007; Keemink and van Rossum, 2016).

In contrast to those studies we show that biases can occur even in homogeneous population codes. We consider the case where multiple variables are simultaneously coded in a population, such as occurs in visual cortical area MT when two overlapping transparent random dot motion patterns are presented. We find that in these situations biases in estimation emerge, even though the decoder has full knowledge of the encoding process. Furthermore, when multiple overlapping stimuli are presented, the number of perceived stimuli can be fewer than the number presented, resembling psycho-physical findings (Treue, Hol, and Rauber, 2000; Edwards and Greenwood, 2005).

To explain these findings we develop a mathematical framework based on Gaussian Processes - an extension of multivariate Gaussian distributions - which is generally

applicable to maximum likelihood decoders for systems with Gaussian noise. We use this framework to calculate and understand the implications of the bias for neural computation and perceptual biases.

Results

To examine the emergence of biases we consider a population of neurons described by their firing rates. The average response of each neuron is given by its tuning curve $f(\mathbf{s})$, where $\mathbf{s}$ is a vector of stimulus parameters encoded by the neuron. Gaussian white noise ν_i with mean zero and variance σ^2 is added to the response, so that on a given trial the firing rate r_i of neuron i is

$$r_i = f_i(\mathbf{s}) + \nu_i. \tag{1}$$

Commonly one studies the case where $\boldsymbol{s}$ is one-dimensional. Here we consider the coding of two stimuli $\mathbf{s} = (s_{1,}s_2)$ simultaneously. For concreteness we consider the coding of two overlapping random dot motion patterns in area MT; in this case s_1 and s_2 represent the two motion directions, Fig. 1A. The response of MT neurons to such a stimulus has been modeled by the linear average or the sum of the tuning curves to the individual stimuli (van Wezel et al., 1996; Treue, Hol, and Rauber, 2000). Under this assumption, the mean firing rate of neuron i is

$$f_i(\mathbf{s}) = g_i(s_1) + g_i(s_2) \tag{2}$$

where $g_i(s)$ is the bell-shaped tuning of neuron i to a single stimulus (Methods). Competitive interactions between the responses are considered below.

Decoding of the neural response

We draw stochastic responses from the above model (see Methods for details) and then decode the stimulus parameters from the noisy population response using the

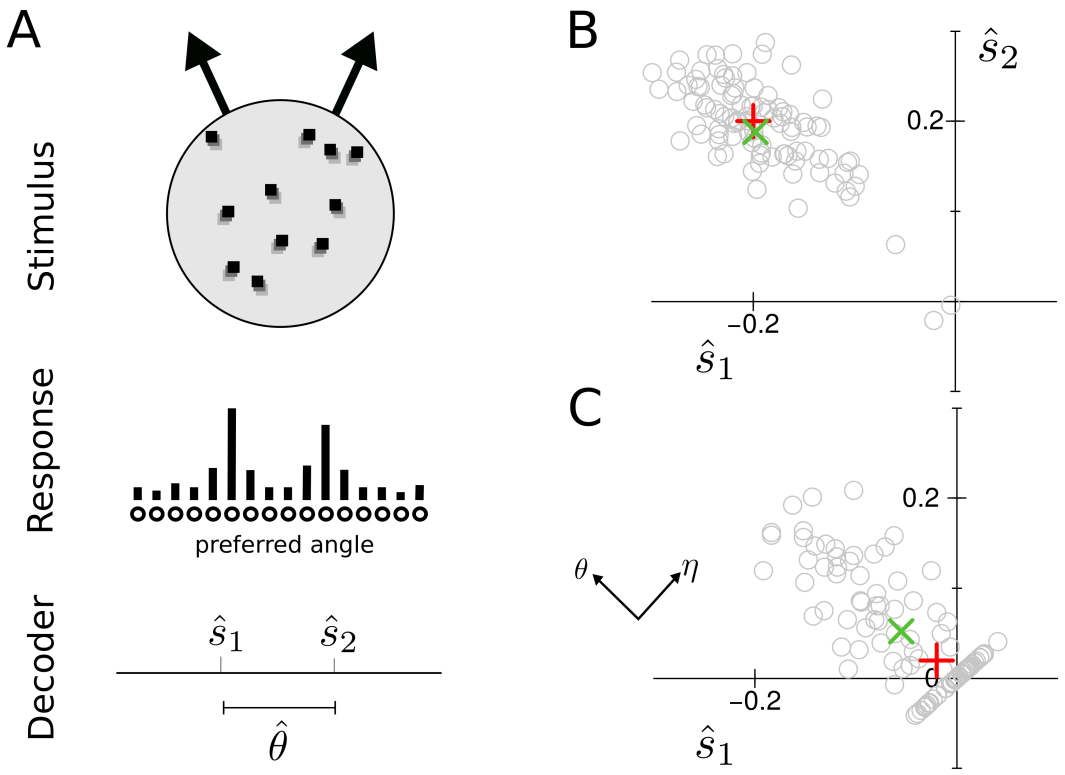

Figure 1: A) Basic encoding-decoding setup. The stimulus consists of two overlapping moving random dot patterns. A population of neurons codes for the two simultaneous stimuli. The task is to estimate the stimulus parameters, here the motion directions $\hat{s}_1$ and $\hat{s}_2$, from the noisy population response. B) Maximum likelihood estimates across a number of trials. For a wide opening angle $\mathbf{s} = (-0.2, 0.2)$, the distribution of estimates follows approximately a 2D Gaussian distribution. True stimulus (red plus) and average estimate (green X) overlap. C) For narrow opening angles, $\mathbf{s} = (-0.02, 0.02)$, the distribution of estimates falls into two roughly equal parts, a Gaussian-shaped distribution and a distribution along the line $\hat{s}_1 = \hat{s}_2$. True stimulus and average estimate now diverge, i.e. the estimate is biased. The sum and difference angles are indicated by η and θ, respectively. (all angles in radians).

maximum likelihood (ML) decoder. That is, estimates of the stimulus $\hat{\boldsymbol{s}}$ are obtained by finding the stimulus vector that was most likely given the noisy neural response vector $\boldsymbol{r}$,

$$\hat{\boldsymbol{s}} = \text{argmax}_{\boldsymbol{s}} \log P(\mathbf{r}|\boldsymbol{s}).$$

The hat indicates estimates throughout. Because the encoder loses the identity of the two stimuli, we additionally impose that $s_2 \geq s_1$.

We first consider the case when the opening angle is large, so that the two peaks in the tuning curve are far apart ($|s_1 - s_2| \gg w$, where w is the tuning width). In this case the stimulus estimates are centered around the true stimulus value, approximately according to a two-dimensional Gaussian distribution, Fig. 1B. The true stimulus value (cross) and the mean estimate (marked by the X) coincide.

However, when the motion directions are instead almost the same so that the peaks in the population response partly overlap, the distribution radically changes shape, Fig. 1C. Now the estimates fall essentially in two categories: Either the estimates are strongly positively correlated, and cluster on the diagonal where $\hat{s}_1 = \hat{s}_2$; in this case the most likely explanation for the neural response is that the two motion directions were the same. Alternatively, on other trials the estimates are negatively correlated, and the angular difference in the motion direction is over-estimated. The mean of neither component of the distribution coincides individually with the true stimulus vector, nor does the mean of the full distribution; the estimate is biased.

To more easily understand these results we transform the coordinates and describe the system in the sum and difference of the angles. The sum of the angles, $\eta = s_1 + s_2$ follows a Gaussian distribution and is unbiased as dictated by the rotational invariance of the setup. More interesting, however, is the opening angle $\Theta = s_2 - s_1$. Estimator bias b is defined as the difference between mean estimate and true stimulus value, $b(\Theta) = \langle\hat{\theta}\rangle - \Theta$, where the angular brackets denote the average over trials of the

estimates $\hat{\theta}$. The estimator bias is shown as a function of true stimulus value Θ in Fig. 2A. When the opening angle Θ is small, the bias is repulsive (the apparent angle is larger than the true value). As the opening angle increases, the bias changes sign and becomes attractive, before reducing to zero for even larger angles, Fig. 2A.

One can wonder whether the repulsive bias is simply caused by imposing $s_2 \geq s_1$. But this would not explain the biphasic nature of the bias nor the bimodal decoding distribution. Furthermore, if the ordering of s_1 and s_2 were randomly assigned, the estimate distribution would become tri-modal with some estimates lying on the diagonal, and others clustering in clouds on the anti-diagonal on either side of the origin. On average one would find $\langle s_1 \rangle = \langle s_2 \rangle$, i.e. $\left\langle \hat{\theta} \right\rangle = 0$, irrespective of the true angle, so it would be a strongly biased estimator.

In summary, in this relatively simple coding problem biphasic biases emerge. Next, we attempt to understand why this occurs.

Emergence of bias

To understand the emergence of the bias we analyze the Maximum Likelihood estimator in detail. For independent Gaussian noise the maximum likelihood estimate is equivalent to minimizing the Mean Squared Error (MSE) E between observed and expected response

$$\hat{\boldsymbol{s}} \;=\; \text{argmin}_{\boldsymbol{s}} E(\boldsymbol{s})$$

where $E(\boldsymbol{s}) = \sum_{i=1}^{N} [r_i - f_i(\boldsymbol{s})]^2$. The emergence of the bias and the underlying distribution of estimates can be understood from the mean square error that the estimator seeks to minimize. The MSE is a smooth curve that varies from trial to trial, Fig. 2B. This collection of curves constitutes a Gaussian Process (a generalization from a Gaussian distribution to a distribution of functions).

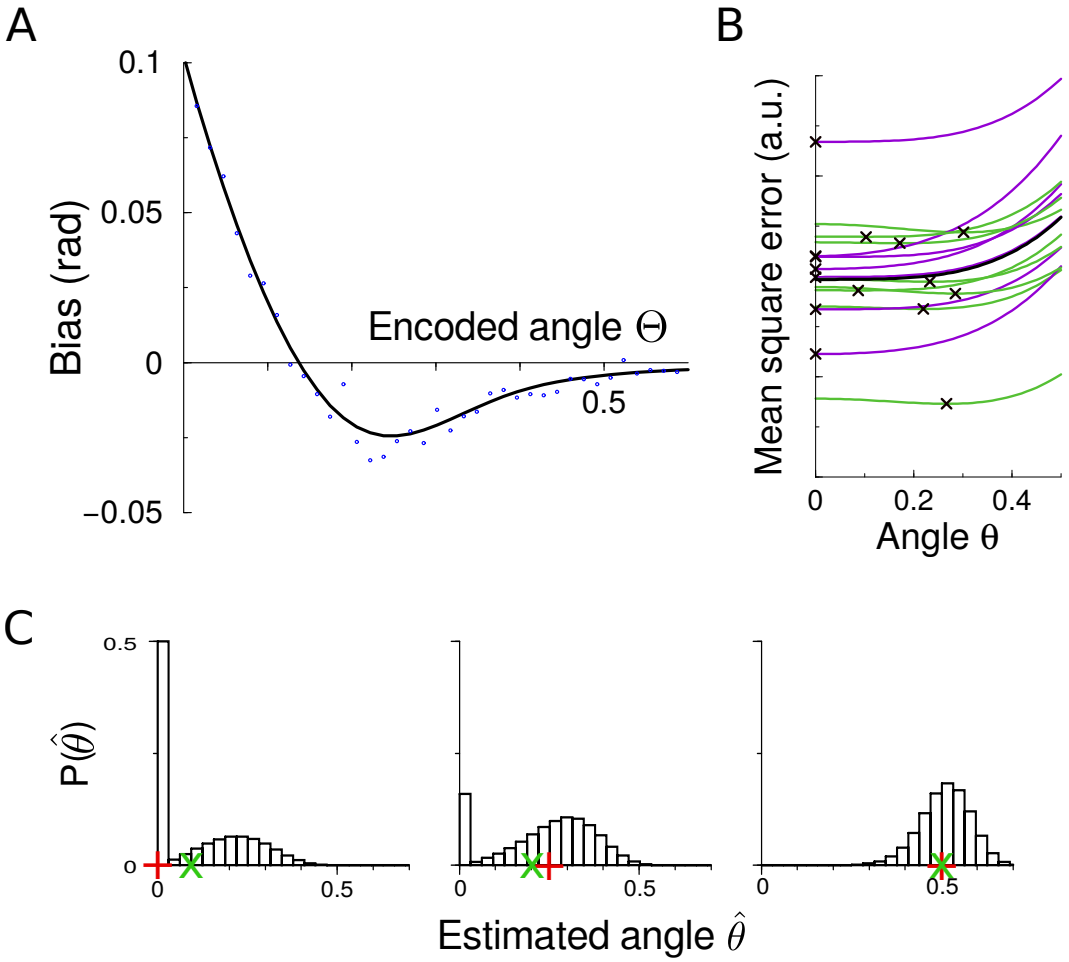

Figure 2: Decoding biases of the opening angle and the underlying decoding distribution.

A) Bias in estimation of the opening angle as a function of its true value, showing both a repulsive bias at small angles and a attractive one at larger angles. The curve was calculated using the Gaussian Process approach given in the Methods. Also shown for comparison are simulations (dots) averaged over 1000 trials per point.

B) Samples of the Mean Square Error in case the true opening angle is zero. The minima of the MSE correspond to the estimates of the maximum likelihood estimator. While the average MSE has a minimum at the true value (black curve), on a given noisy trial the estimate can either be exactly $\theta = 0$ (shown in purple), or repulsed away from it (shown in green). The black crosses indicate the estimates, i.e. the angle that minimizes the error, on the individual trials.

C) Distributions of estimates that underly the bias. The true stimulus value is indicated with the red plus on the x-axis, the mean estimate is denoted with the green

To write the MSE as a Gaussian Process (Williams and Rasmussen, 2006) we first split the MSE up as

$$E(\theta) \quad = \quad E_{\text{mean}}(\theta) + E_{\text{noise}}(\theta) + C, \tag{3}$$

where C is a stimulus independent term, and θ denotes the candidate stimulus[1]. The stimulus dependent part consists of two terms: the first term is the mean error $E_{\text{mean}}(\theta) = \sum_{i=1}^{N}[f_i(\theta) - f_i(\Theta)]^2$ that is identical across trials and attains its minimal value of zero at the true stimulus value, Θ. The second term is the noise term that varies from trial to trial $E_{\text{noise}}(\theta) = -2\sum_{i=1}^{N}\nu_i f_i(\theta)$, with covariance $\Sigma_{\theta,\theta'} = 4\sigma^2\sum_{i=1}^{N} f_i(\theta)f_i(\theta')$.

Of particular interest is the limiting case of $\Theta = 0$. While somewhat contrived as the presented motion directions are identical in that case, exact results can be obtained in this limit that approximately hold for any small Θ. In this limit $E_{\text{mean}}(\theta)$ is lowest at $\theta = 0$, as expected, Fig. 2B, thick black curve. Because of symmetry in the combined tuning curves, Eq. 7, not only all odd derivatives, but also the second derivative of E_{mean} is zero. Thus the dependence is quartic, $E_{\text{mean}}(\theta) \sim \theta^4$, i.e. very flat, and for small opening angles, this error term hardly changes as the estimate is altered.

As the noise term E_{noise} combines the signals from overlapping tuning curves it is smooth. It is also symmetric in θ, however its second derivative is non-zero. Depending on the noise it is in leading order either an upward or downward curved parabola centered around the origin. For small θ this parabola will dominate over E_{mean}. Therefore, if the E_{noise} parabola is U-shaped and thus with a minimum at $\theta = 0$, the total MSE also has a global minimum there, Fig. 2B, purple curves. If, on the other hand, the noise term has a maximum at $\theta = 0$, the global minimum will be repulsed

[1]As a reminder to the reader: Θ denotes the true stimulus value, θ the possible candidate estimates, and $\hat{\theta}$ the estimate (the θ which is most likely).

away from the true solution, Fig. 2B, green curves. As a result the distribution shows a sharp peak at 0, and a smeared peak further away, Fig. 2C (left). Furthermore, when the encoded angle $\Theta = 0$, exactly half of the estimates will be at $\theta = 0$ (i.e. fall on the diagonal in Fig. 1C) and the other half will not. As Θ increases, the probability to find estimates $\hat{\theta} = 0$ will decrease and the second peak will gain more mass, Fig. 2C (middle and right). The net effect is that this will first decrease the repulsive bias, then turn into an attractive bias, and finally the bias disappears.

The Gaussian Process approach can be used to calculate the probability of estimates $P(\hat{\theta}|\Theta)$ in a numerically exact way without relying on simulations. Briefly, for a given true stimulus Θ, we run over all candidate stimulus estimates θ and find the probability that it minimizes E (see Methods). This gives an accurate and noise-free estimation of the decoding distribution, and thus of the decoding bias. This method was used to create Fig. 2A+C, and compares well to explicit stochastic simulations over many trials (dots in Fig. 2A).

In an elegant, but little known, paper Amari and Burnashev (2003) calculated the bias analytically in the case of Gaussian white noise (see Methods). While our approach does not yield the analytical form of the distribution, it has as advantage that it allows for more general encoding and noise models as we examine below.

The bias depends on the neural noise level and other system parameters. In the limit of small angles the bias can be found by estimating the expected location of the minima of the Mean Square Error (crosses in Fig. 2B). As shown in the Methods this gives for the tuning curves used,

$$b(0) = c\sqrt{\frac{\sigma}{A}}.\sqrt[4]{\frac{w^3}{N}}. \tag{4}$$

where σ is the std.dev. of the neural noise, w is the tuning width, A is the maximum neural response amplitude, N is the number of neurons and $c \approx 1.2$ is a numerical constant. Therefore to, say, half the bias, one needs 4 times smaller noise, or 16

times as many neurons. The second effect of the noise level is a shift in the angle at which repulsion becomes attraction, i.e. where the curve in Fig. 2A crosses the x-axis. The location of this transition point is approximated by the bias at zero, as the derivative of the bias at the origin equals $b'(0) = -1$, Fig. 2A. The reason for this is that the estimator $\langle\hat{\theta}\rangle$ is a smooth, symmetric function in Θ, so for small Θ, $\langle\hat{\theta}\rangle \approx const + O(\Theta^2)$, and so $b'(0) = -1$.

Interestingly, as the noise is reduced, the distribution of estimates remains bimodal. While in the limit of zero noise the bias disappears (as the theory of maximum likelihood estimation states), the transition in the limit of small angles is not due to a collapse of $P(\hat{\theta})$ into a single Gaussian distribution, rather it is due to the two peaks in the distribution of estimates moving closer and closer together.

Intuitively, the bias emerges due to the interaction two effects: 1) for small opening angles the two stimuli are interpreted as being just a single stimulus leading to attractive bias, and 2) when the stimulus is correctly perceived as being two directions, the angle estimate is broad and tends to overestimate, leading to a repulsive bias.

Effect of noise correlation and heterogeneity

Next we examine how the bias depends on the neural noise, tuning curve heterogeneity, and encoding model; all these effects can be included in the Gaussian Process approach without additional computational cost or complexity. First, we consider the effect of correlations in the neural noise. In studying the coding of single stimuli it was found that correlations in the neurons' noise, so called noise correlations, limit the ability to average across neurons. This effect was particularly important when the spatial scale of the correlations, ω, matched the tuning curve width (Sompolinsky et al., 2002; Wu, Amari, and Nakahara, 2002). A similar effect is found here, Fig. 3A: for uncorrelated noise (small ω) the bias can be reduced by using larger neural populations as in Eq. 4,

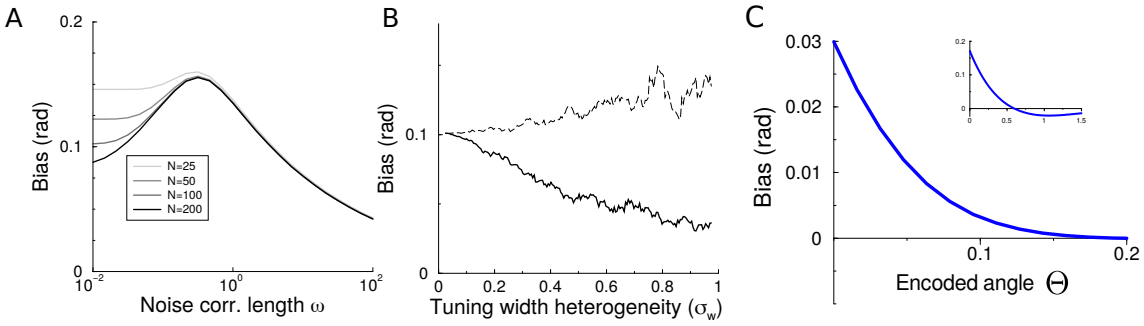

Figure 3: The dependence of bias on the encoding model: noise, heterogeneity, and competitive coding.

A) The bias at zero angle as a function of the noise correlation length across different neural population sizes. Increasing the number of neurons reduces the bias only for small correlation lengths (independent noise).

B) The bias at zero angle as a function of tuning curve heterogeneity when tuning curves widths were drawn from a log-normal distribution so that the distribution of widths has a standard deviation σ_w, and a mean of $1/2$. Heterogeneity decreases the bias (solid curve). Dashed curve: control case when the bias is averaged across a set of homogeneous populations (see text).

C) Bias in the estimates in a competitive coding model where the response of any neuron to two stimuli equals the maximum response to the individual stimuli for the same noise level as used in Fig. 2A (note the difference in scales). Only at high noise levels ($\sigma = 1$), an attractive bias manifests itself (inset).

but for correlated noise it can not. The bias is maximal for intermediate correlation length. The bias diminishes (but remains finite) for large correlation lengths in which case all neurons co-vary across trials.

Next, we consider heterogeneity among the tuning curves, which again from univariate coding studies is known to improve population coding accuracy (Shamir and Sompolinsky, 2006; Ecker et al., 2011). Similarly, heterogeneity resolves some of the degeneracy at $\Theta = 0$ that underlie the bias. To examine this, we drew the widths of the individual tuning curves from a log-normal distribution. Indeed, increasing the heterogeneity among the tuning curves decreased the bias, Fig. 3B. The bias reduction could be simply the result of the inclusion of a few neurons with very narrow tuning curves in the population that allow a precise estimate. To check against this explanation we calculated the average bias from a set of homogeneous populations as $\int b(\theta)P(w)dw$, with $P(w)$ the distribution of widths and $b(\theta)$ from Eq. 4. This has only a weak effect on the bias, Fig. 3B (dashed line), instead it is the heterogeneity that underlies the bias reduction.

Bias reduction strategies

Bias reduction by the encoder

The estimation bias depends on how the stimuli are encoded in the neural response. Above it was assumed that the neural response to two simultaneous stimuli was the linear sum of the responses to the individual stimuli. While there is some experimental evidence for such a linear interaction, it is known that this type of interaction limits coding accuracy (Orhan and Ma, 2015). Furthermore, in other studies evidence for more competitive interaction has been found in area MT (Britten and Heuer, 1999), as well as other visual cortices (Gawne and Martin, 2002; Lampl et al., 2004; Oleksiak et al., 2011). Such interactions have been modeled using a maximum-like interaction,

so that instead of Eq. 2, the response of a single neuron to two simultaneous stimuli is

$$f_i(s_1, s_2) = \max[g_i(s_1), g_i(s_2)]. \tag{5}$$

Since under this encoding model the mean term in the MSE is not quartic but quadratic in θ, one would expect a lower bias. Indeed, when the simulations are repeated for this encoding model, the bias is still present, but it is substantially smaller, Fig. 3C. The repulsive bias is now approximately linear in the noise and the attractive component of the bias is smaller and becomes only apparent at high noise levels (inset). Thus we find that the encoding model is an important determinant of the size of the bias, and these findings suggest a functional role for the competitive interaction observed experimentally.

Bias reduction by the decoder

We wondered whether the bias is unavoidable or is perhaps particular to the ML decoder. First, we use a Bayesian decoder, which calculates the full distribution of possible stimulus estimates given the response and the noise model, $P_B(\theta|\boldsymbol{r})$. For a flat prior for Θ, this distribution is proportional to $P(\boldsymbol{r}|\theta)$. Whereas the maximum likelihood decoder takes the maximum of this distribution, using a square loss function the Bayesian estimate equals the expected value of this distribution, $\hat{\theta}_B = \int \theta P_B(\theta|\boldsymbol{r})\, d\theta$ (Kay, 1993; Salinas and Abbott, 1994). This estimator minimizes the mean square error in the estimate. We find that with a Bayesian decoder the bias is slightly more pronounced, Fig.4A.

Can one design a bias-free decoder? If there is a smooth, monotonic, potentially noisy, relation between an estimate and the true stimulus, one can hope to compensate the bias. While here the bimodal distribution of estimates makes this more challeng-

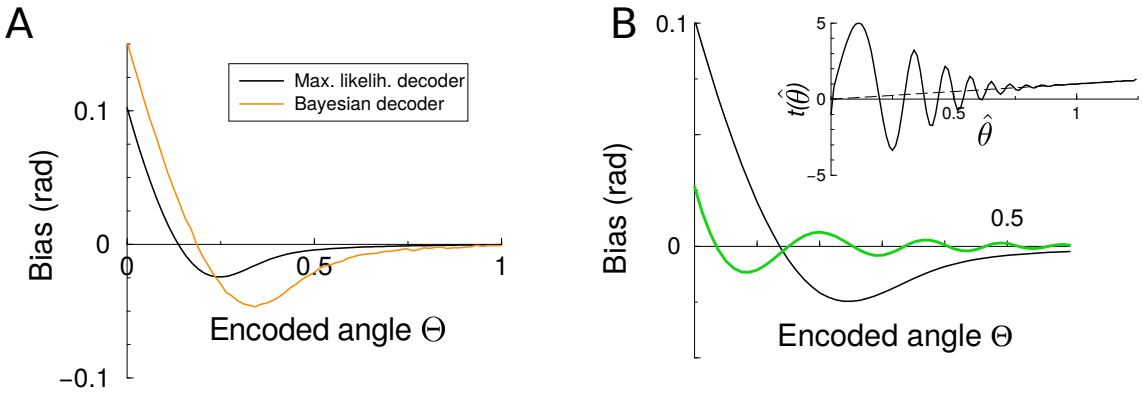

Figure 4: Effect of decoder on bias. A) Bias of a Bayesian decoder (orange). Shown for comparison, the ML decoder used throughout (black). The biases are of comparable magnitude and share the biphasic character. B) The bias after applying an optimized non-linearity on the individual estimates (green); the bias of the ML-decoder is shown for comparison (black). Inset: the non-linearity found. Dashed line shows the identity function. (Both Θ and $\hat{\theta}$ were discretized into 100 bins, regularization parameter $\lambda = 10^{-6}$).

ing, we wondered if nevertheless one can compensate for the ML-decoder bias. Is there a non-linear mapping replacing each estimate of the ML-decoder $\hat{\theta}$ with a transformed estimate $t(\hat{\theta})$, that reduces the bias across the range of possible encoded angles?

After discretizing both Θ and $\hat{\theta}$, the bias after correction ($b = \langle t(\hat{\theta}) \rangle - \Theta$) becomes $b(\Theta_j) = \sum_i P(\hat{\theta}_i|\Theta_j)t_i - \Theta_j$, where $t_i = t(\hat{\theta}_i)$. The goal now is to find the vector $\mathbf{t} = (t_1, t_2, \ldots)$ so that $b(\Theta_j) = 0$ for all j. This is a linear algebra problem of the classic form $A\mathbf{x} = \mathbf{y}$, where $\mathbf{x}$ (here $\mathbf{t}$) has to be found for known A (here P) and $\mathbf{y}$ (here Θ). Because the problem is ill-conditioned, we use singular value decomposition to find $\mathbf{t}$. Furthermore, as the entries of $\mathbf{t}$ diverge, we apply Tikhonov regularization. As a result of the regularization the bias won't be exactly zero, but the norm of $\mathbf{t}$ is limited (Methods).

With this procedure the bias is substantially reduced compared to the uncorrected estimator, Fig.4B (green vs black curve). The non-linearity found to achieve this fluctuates smoothly around zero (inset). The intuition is that the oscillations make the non-linearity highly sensitive to small changes in $P(\hat{\theta}|\Theta)$, while maintaining the correct mean. Reassuringly, for larger estimates, where bias was small anyway, no non-linearity is needed and $t(\hat{\theta}) = \hat{\theta}$ (inset, dashed line). However, the bias reduction comes at a cost: because of the oscillations in $\mathbf{t}$, the estimates vary wildly from trial to trial. For the case illustrated the variance in the estimate is some 300 times larger than for the ML-decoder for a fourfold bias reduction. As the regularization is relaxed, the bias can be made smaller, but the amplitude (and frequency) of $\mathbf{t}$ increase, and hence the variance grows.

Bias and estimator efficiency

The quality of population code readout is not quantified by the bias alone, but also by the amount of trial-to-trial variations in the estimates, i.e. the variance in the

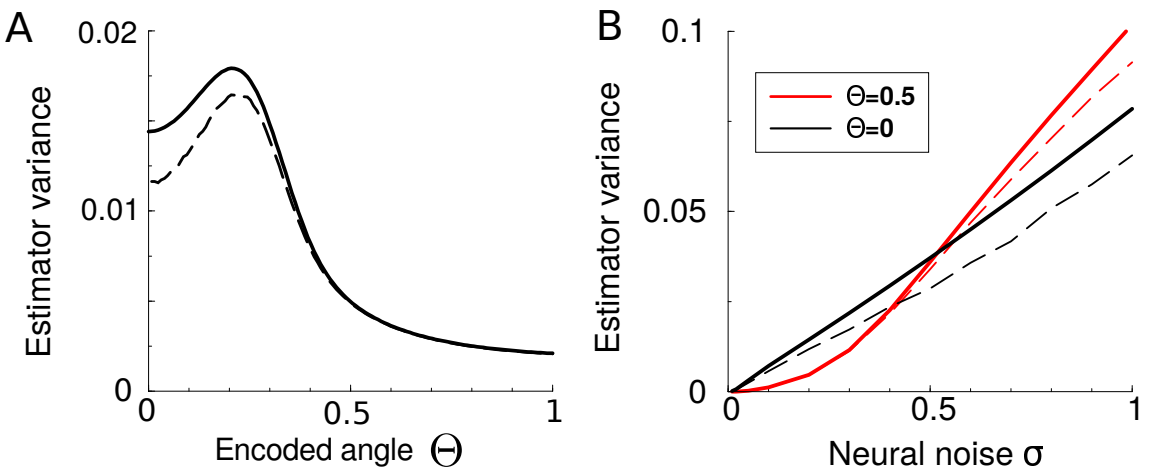

Figure 5: Variance and efficiency of the maximum likelihood decoder and its dependence on encoded angle and neural noise.

A. Variance in the estimates of the ML decoder (solid curve) depends non-monotonically on the encoded angle. Dashed curve:the Cramèr-Rao bound with the bias taken into account - no estimator can achieve a lower variance.

B. Variance in the estimates of the ML decoder (solid curve) as a function of the neural noise comparing large and small encoded angles. At small angles the strong bias alters the expected square dependence on noise into linear behavior. Dashed curves correspond to the Cramèr-Rao bound with the bias taken into account.

distributions in Fig. 1B+C. The variance of the estimates from the ML decoder follows directly from the distribution of estimates $P(\hat{\theta}|\Theta)$ that our approach yields. The variance in the estimator is plotted in Fig. 5A.

For large opening angles Θ, the estimate distributions are Gaussian with a width proportional to the neural noise, Fig. 2C, right; and the bias plays a minor role. As expected from the Fisher Information, the variance of the estimator is proportional to the square of the neural noise, Fig. 5B, red curve. However, for small opening angles the bias has a profound effect on the estimator, causing a linear dependence on the noise, Fig. 5B, black curve. This can be understood as follows: The estimator variance at $\Theta = 0$ can be approximately found by describing the estimate distribution $P(\hat{\theta}|\Theta)$ as a peak at zero and a Gaussian, Fig. 2C. The variance is similar to the bias squared, and thus its parameter dependence follows from squaring Eq. 4. Therefore in contrast to the behavior at large angles, the variance in the estimates is only linear in the neural noise.

The minimal variance any estimator can achieve is limited by the Fisher Information through the Cramèr-Rao bound which states that the variance of any estimator obeys (Methods)

$$\mathrm{var}(\hat{\theta}) \geq \frac{[1 + b'(\Theta)]^2}{\mathcal{I}(\Theta)}, \tag{6}$$

where b' is the derivative of the bias, and the Fisher Information $\mathcal{I}(\Theta)$ is given by Eq. 9, Methods. The efficiency of an estimator expresses how close it comes to this bound. The resulting Cramèr-Rao bound is indicated by the dashed curve in Fig. 5A.

While the Fisher Information is proportional to the neural noise squared σ^2, see Eq. 9, the Cramèr-Rao bound at small opening angles is only linear in the neural noise, Fig. 5B, dashed curves. The reason is that for small angles the Fisher Information goes to zero (Eq.9), but the bias' derivative at the origin equals $b'(0) = -1$. Hence at

small angles, both numerator and denominator in the Cramèr-Rao bound (Eq. 6) go to zero. The bound does therefore not diverge and a linear dependence on the neural noise remains. For the parameters used, the ML always achieves an efficiency $\geq 80\%$.

As an aside, here another advantage of the Gaussian Process approach shows. With simulations the bias and in particular its derivative are hard to calculate accurately, even using a large number of realizations, Fig. 2A dots. However, the numerically exact method to calculate the bias allows for a precise calculation of the bias and its derivative.

Discussion

Traditionally, theoretical studies of the accuracy limits of population codes have focused on estimator variance. Whenever biases have been studied theoretically, they have typically been explained from inhomogeneities in the neural encoding. Here we find that when multiple stimuli are encoded simultaneously in a relatively simple coding problem, substantial biases arise with standard decoders. That biases occur is in itself should not be surprising. Apart from cases where symmetry rules out biases, the absence of biases can be proven in the limit of low noise, but in general an ML decoder will not be unbiased, nor efficient (Kay, 1993; Seriès, Stocker, and Simoncelli, 2009; Pilarski and Pokora, 2015). Yet, the rich structure of the bias in these simple models, including its biphasic character and its relative persistence at low noise, is surprising. The reason for the bias is the bimodal distribution of decoding estimates. The bias will disappear in the limit of zero noise, but it diminishes only slowly as noise is reduced (proportional to the square root of the std. dev. of the neural noise, Eq. 4). The persistence of the bias at low noise stands in contrast to other studies of bias and efficiency where below a critical noise level the decoders abruptly become optimal (Kay, 1993; Xie, 2002). Simulations show that the shape of the bias curve and

the distribution of estimates is very similar when instead of Gaussian noise, Poisson or multiplicative Gaussian noise is considered (not shown), but the Gaussian process approach can not be used in the Poisson case.

This particular coding problem has been studied twice before. Orhan and Ma (2015) calculated the Fisher information but did not consider decoder biases. Amari and Burnashev (2003) showed that for uncorrelated noise a singularity in the Fisher Information leads to a bound on $\langle(\hat{\theta} - \Theta)^2\rangle = var(\hat{\theta}) + b^2(\Theta)$ for any decoder. The Gaussian Process approach numerically matches their analytical results, but also allowed us to study correlated noise, non-linear encoding and heterogeneity.

While the ML decoder and the Bayesian decoder have a strong bias, it can be arbitrarily reduced by a non-linear mapping. However, this comes at the cost of a much increased variance (inline with Amari and Burnashev, 2003) and the compensating decoder is probably too complicated and fragile for biological implementation and addition it would need to be made dependent on noise level. The neural implementation of decoding mechanisms is currently not clear, although it has been argued that it is straightforward to implement ML decoders neurally (Deneve, Latham, and Pouget, 1999; Jazayeri and Movshon, 2006). One could wonder for which coding problems ML decoders are biased. While we don't currently have a general answer to this, one can employ the Gaussian Process approach to explore potential cases (under a Gaussian noise assumption).

The question which decoder the brain implements and whether bias is present is ultimately an empirical one and can be tested in psycho-physical experiments. For instance, two overlapping random dot motion patterns with different directions can be presented and subjects are asked to guess the angle between the two directions. In such experiments repulsive biases have commonly been observed (Marshak and Sekuler, 1979), but attractive effects have also been observed (Braddick, Wishart,

and Curran, 2002). Several effects have been hypothesized to underlie these biases, including adaptation (Rauber and Treue, 1999), cortical interactions (Carandini and Ringach, 1997) and repulsion from the cardinal directions (Rauber and Treue, 1998). The bias described here, is not at odds with those explanations, but presents a novel contribution to the total bias. It should be most prominent at small angles and when presentation times are short so that the signal-to-noise ratio is small.

The estimated decoding distribution reflects an ambiguity between the presence of one or two stimuli. Apart from predicting a bias, the theory predicts a bimodal distribution of direction difference estimates and for small angles about half the time the two motions should be perceived as one. In experiments the number of stimuli that can simultaneously be perceived using overlapping motions is limited (e.g. Edwards and Greenwood, 2005) and when three or five overlapping motions are presented, they can sometimes be perceived as two (so called metamers, Treue, Hol, and Rauber, 2000); an effect which previously has been explained using the probabilistic population code framework (Zemel, Dayan, and Pouget, 1998; Zemel and Dayan, 1999). The results here suggest that differences in the numerosity between presented and perceived stimuli already emerge with maximum likelihood decoders. Quantitative verification of the predictions of our study regarding bias and numerosity should be possible but attention, participants' expectations, and natural priors for perceiving a single motion direction instead of two directions should be taken into account.

More generally, these result might also be relevant for other brain areas such as higher visual areas. Here our findings pose limits on the number of objects that can be represented simultaneously in a neural population. The competitive coding studied here, or alternatively, complex temporal dynamics during simultaneous stimulus presentation (Gawne, 2008; Li et al., 2016), might help to alleviate such limitations (Amari and Nakahara, 2005).

Acknowledgments

This work was supported in part by grants EP/F500386/1 and BB/F529254/1 to the University of Edinburgh School of Informatics Doctoral Training Centre in Neuroinformatics and Computational Neuroscience from the UK Engineering and Physical Sciences Research Council (EPSRC), UK Biotechnology and Biological Sciences Research Council (BBSRC), and the UK Medical Research Council (MRC). SK was supported by the EuroSpin Erasmus Mundus program and the EPSRC NeuroInformatics DTC. The work has made use of resources provided by the Edinburgh Compute and Data Facility (ECDF; www.ecdf.ed.ac.uk), which has support from the eDIKT initiative (www.edikt.org.uk). The authors would like to thank Udo Ernst, Richard van Wezel, Lawrence York, Peggy Series, and Chris Williams for discussions.

Methods

Neural population response

We use a population of $N = 100$ neurons. The tuning of neuron i to a single stimulus is given by $g_i(s) = A \exp\left[-\frac{(s-\phi_i)^2}{2w^2}\right]$. Here A is the response amplitude (arbitrarily set to 1), w is the width of the tuning curve (set to 1/2). The preferred directions ϕ_i of the neurons are equally spaced between 0 and 2π. As is common, we assume that the angles involved are relatively small, so that we don't have to worry their circularity, which would add complication through the need for circular statistics but does not change the results qualitatively.

When multiple stimuli are present, the neural response is modeled as the sum of the responses to the individual stimuli. After transforming the variables to the sum angle η and the difference angle Θ (see Main text) we can set η to zero, so that the

tuning of neuron i becomes

$$f_i(\Theta) = g_i(\Theta/2) + g_i(-\Theta/2). \tag{7}$$

By replacing A by half its value, the joint tuning curve equals the average (instead of the sum) of the tuning curves. The default value of the std. dev. of the noise in Eq. 1 was $\sigma = 0.2$. Correlated noise (Fig. 3A) was parameterized as $Q_{ij} = \sigma^2[\delta_{ij} + c(1 - \delta_{ij}) \exp(-|\phi_i - \phi_j|/\omega)]$, where ω is the range of the correlation and the strength of the correlation c was set to 1. To include response heterogeneity (Fig. 3B), the widths of the tuning curves were drawn from a log-normal distribution.

Algorithm to calculate of maximum likelihood estimate

Here we demonstrate how to calculate the distribution of estimates $P(\hat{\theta}|\Theta)$ of the ML estimator in a numerically exact manner. In case of correlated noise the negative log-likelihood becomes

$$E(\theta) = \frac{1}{2} \sum_{i,j} [r_i - f_i(\theta)] Q_{ij}^{-1} [r_j - f_j(\theta)]$$

where $Q_{ij} = \langle \nu_i \nu_j \rangle$, noting that in case of uncorrelated noise $Q_{ij} = \sigma^2 \delta_{ij}$, we retrieve the MSE up to a factor.

Given a noisy response $\mathbf{r}$, we run over all candidate stimulus estimates θ and find the probability that it minimizes E. Because E is a smooth Gaussian process, and nearby E's are correlated, we finely discretize the possible estimates θ and define a set of M candidate estimates $(\theta_1, ..., \theta_M)$.

To calculate the probability that a certain estimate θ_m yields the lowest MSE, it is compared to the MSE that all other $M - 1$ estimates yield. We define the $M - 1$ dimensional set of MSE differences

$(E(\theta_m) - E(\theta_1), \ldots, E(\theta_m) - E(\theta_{m-1}), E(\theta_m) - E(\theta_{m+1}), \ldots, E(\theta_m) - E(\theta_M))$. We write this in short-hand as the vector $\mathbf{D}_m = E(\theta_m) - E(\mathbf{\Phi}_m)$, where $\mathbf{\Phi}_m = \{\theta_1, ..., \theta_M\} \backslash \theta_m$.

The distribution of differences $\mathbf{D}_m$ is a $(M-1)$-dimensional multivariate normal distribution

$$p(\mathbf{D}_m|\Theta) = \mathcal{N}(\boldsymbol{\mu}^m, \Sigma^m),$$

where $\boldsymbol{\mu}^m = E_{\text{mean}}(\theta_m) - E_{\text{mean}}(\boldsymbol{\Phi}_m)$ and the $(M-1) \times (M-1)$ covariance matrix has entries $\Sigma^m_{ab} = \sum_{i=1}^{N}[f_i(\theta_m) - f_i(\theta_a)]Q^{-1}[f_i(\theta_m) - f_i(\theta_b)]$. The probability that θ_m has a lower MSE than all other candidate estimates is

$$p(\mathbf{D}_m < \mathbf{0}|\Theta) \;\; = \;\; \int_{-\infty}^{0} ... \int_{-\infty}^{0} p(\mathbf{D}_m|\Theta)\mathrm{d}\boldsymbol{D}_m, \tag{8}$$

which is a multi-variate cumulative normal distribution. We evaluated the integral for all values of m, to yield $P(\hat{\theta}|\Theta)$.

While the orthant integral, Eq. 8 is not analytically tractable, efficient algorithms exist that calculate it to a high precision for values of M up to in the hundreds. We used the quasi-Monte Carlo integration function `mvnun` from Scipy (Genz, 1992, 1998) with $M = 100$ and $\theta = 0 \ldots \pi$ (using a larger M had negligible effects). The code is available at https://github.com/swkeemink/DeDist.

We note that the algorithm also extends to higher dimensional stimuli, but is in practice limited by the dimensionality of the integral (which is equal to the number of bins used for the stimulus space discretization). However, algorithms for even higher dimensions exist (e.g. Azzimonti and Ginsbourger, 2016).

Scaling of the bias

Here we calculate the bias for small angles analytically and estimate how the bias scales with the model parameters under the assumption of uncorrelated noise. We use that in case of small Θ and the limit of small candidate angles θ, the mean square

error, Eq. 3, can be Taylor expanded as (ignoring the scaling with $1/2\sigma^2$)

$$\begin{aligned} E_{\text{mean}}(\theta) &= \sum_i [f_i(\theta) - f_i(\Theta)]^2 \\ &\approx \rho \int_{-\infty}^{\infty} [g_a(-\theta/2) + g_a(\theta/2) - 2g_a(0)]^2 \mathrm{d}a \\ &= 2\sqrt{\pi}\rho w A^2 \left[3 + \exp(-\theta^2/4w^2) - 4\exp(-\theta^2/16w^2)\right] \\ &\approx \frac{3\sqrt{\pi}}{64}\frac{\rho A^2}{w^3}\theta^4 \equiv \alpha\theta^4 \end{aligned}$$

where we replaced the sum by an integral and where ρ is the coding density (the number of neurons per angle, $\rho = N/2\pi$) and a indexes the neurons. Similarly, the noise term on a given trial can expanded as

$$E_{\text{noise}}(\theta) = -2\sum_{i=1}^{N} \nu_i f_i(\theta) \approx -\theta^2[\sum_i \nu_i f_i''(0)]$$

The coefficient in the square brackets is a sum of Gaussian variables and so is itself a Gaussian random variable with zero mean and a variance $4\sigma^2 \sum_i [f_i''(0)]^2 \approx \frac{3\sqrt{\pi}}{4}\sigma^2\rho A^2/w^3$. We are interested in the cases where the coefficient will be negative as these are the repulsive trials, which happens in half of the trials. The mean value of a Gaussian truncated below zero is $-\sqrt{2/\pi}$ times the standard deviation, so that for these cases $\langle E_{\text{noise}}(\theta)\rangle \approx -\beta\theta^2$, with $\beta^2 = \frac{3}{4\sqrt{\pi}}\sigma^2\rho A^2/w^3$.

The approximate location of the repulsed minimum is given by $\frac{dE(\theta)}{d\theta}|_{\theta=\hat{\theta}} = 0$, or $\frac{d}{d\theta}(\alpha\theta^4 - \beta\theta^2)|_{\theta=\hat{\theta}} = 0$, and thus $\hat{\theta}^2 = \frac{\beta}{2\alpha}$. The bias in the other half of the trials is zero (purple traces in Fig. 2B), hence the average bias is $b(0) = \frac{1}{2}\sqrt{\frac{\beta}{2\alpha}}$. This yields the relation in the main text, Eq.4. The dependency of Eq.4 on its parameters was confirmed numerically.

Calculation of the Cramèr-Rao bound

Here we show how the Fisher Information is calculated which we use to compare to the variance in the estimator. The Fisher Information matrix for additive, uncorrelated

Gaussian noise is given by $\mathcal{I}_{kl} = \frac{1}{\sigma^2}\sum_{i=1}^{N} \partial_{s_k} f_i(\boldsymbol{s})\partial_{s_l} f_i(\boldsymbol{s})$. While in the original $\mathbf{s}$-coordinates the Information matrix has off-diagonal elements (Orhan and Ma, 2015), in the coordinates (Θ, η) it becomes diagonal. To calculate the Fisher Information we use that in the limit of dense tuning curves, the sum becomes an integral. For example,

$$\mathcal{I}_{11}(\Theta) = \rho \int_{-\infty}^{\infty} [g_a'(\Theta/2) + g_a'(-\Theta/2)]^2 \mathrm{d}a$$

where as above a replaces ϕ_i. We find that

$$\mathcal{I}(\Theta) = \frac{A^2\rho\sqrt{\pi}}{8w^3\sigma^2} \begin{pmatrix} 2w^2 + (\Theta^2 - 2w^2)\mathrm{e}^{-\Theta^2/4w^2} & 0 \\ 0 & 2w^2 + (2w^2 - \Theta^2)\mathrm{e}^{-\Theta^2/4w^2} \end{pmatrix}, \quad (9)$$

The diagonal nature confirms the intuition that the opening and the sum angles can be estimated independently. Further note that both information components depend on the opening angle Θ, but neither depends on the sum angle η. This is due to the rotation invariance of the problem w.r.t. η. Finally, the Fisher Information for Θ, that is $\mathcal{I}_{11}$, goes to zero for small Θ (Amari and Nakahara, 2005).

An estimator is called *efficient* if its decoding covariance satisfies the Cramèr-Rao bound (CRB) (Rao, 1945, 2008; Cramér, 1946). In the often studied case of un-biased, one-dimensional estimators, the CRB is $\mathrm{var}(\hat{\theta}) \geq 1/\mathcal{I}(\Theta)$. In the case of biased vector parameters the CRB states that the matrix

$$C - B\mathcal{I}^{-1}B^T,$$

should be a positive definite matrix (Cover and Thomas, 1991; Kay, 1993). Here C is the covariance matrix of the stimuli that the estimator yields, B is the sum of the Jacobian matrix of $\mathbf{b}$ and the identity matrix $B_{ij} = \delta_{ij} + \partial_j b_i$. In our case this reduces to the bound in the main text.

Fisher Information in competitive coding model

For the competitive coding the Fisher Information is identical for both sum and difference angles, and again only depends on Θ, $\mathcal{I}(\Theta) = \frac{A^2\rho}{8w^2}\{\sqrt{\pi}w[1 + \mathrm{erf}(\Theta/2w)] - \Theta e^{-\Theta^2/4w^2}\}I$, where I is the 2×2 identity matrix. This is a monotonically increasing function in Θ. When there are two separate peaks in the population response ($\Theta \gg w$), the information is twice that when $\Theta = 0$, where there is a single peak in the tuning.

De-biasing non-linearity

To calculate transformations that reduce the bias, we solve

$$\sum_j P(\hat{\theta}_i|\Theta_j)t_j = \Theta_i \tag{10}$$

for t_j, where Θ_i is the discretized encoded angle written as a vector $\mathbf{\Theta} = (0,\, \Delta\Theta,\, 2\Delta\Theta,\, 3\Delta\Theta, \ldots)$, and similar for $\hat{\theta}$.

We use the analytical expression for the conditional probability distribution $P(\hat{\theta}|\Theta)$ derived in Amari and Burnashev (2003) for not too large Θ. After transformation of variables of Eqs. 34 and 47 there and ignoring the scaling with noise, one has

$$P(\hat{\theta}|\Theta) = \sqrt{\frac{2}{\pi}}\theta \exp[-\frac{1}{2}(\theta^2 - \Theta^2)^2] + p_0\delta(\theta)$$

with $p_0 = \frac{1}{2} - \frac{1}{2}\mathrm{erf}\left(\frac{\Theta^2}{\sqrt{2}}\right)$. As an aside, with some effort this expression can be integrated to give the following curious, analytical expression for the bias in terms of Bessel functions

$$B(\Theta) = \sqrt{\frac{\pi}{2}}\frac{e^{-z}}{\Theta}\left[z\left(I_{-\frac{1}{4}}(z) + I_{\frac{1}{4}}(z) + I_{\frac{3}{4}}(z) + I_{\frac{5}{4}}(z)\right) + \frac{1}{2}I_{\frac{1}{4}}(z)\right] - \Theta$$

where $z = \Theta^4/4$.

To solve Eq.10 for $\mathbf{t}$, we use the standard SVD decomposition $P = U.S.V^T$, where U and V are orthogonal matrices, and S is a diagonal matrix. Now $\mathbf{t} = V.S^{-1}.U^T.\mathbf{\Theta}$.

To regularize this solution we replace the elements of the diagonal matrix S^{-1}, s_i^{-1}, with $s_i/(s_i^2 + \lambda)$.

References

Amari, S.-i. and M. V. Burnashev (2003). On some singularities in parameter estimation problems. *Problems of Information Transmission* 39(4): 352–372. Originally in Russian as Problemy Peredachi Informatsii, Russian Academy of Sciences, Branch of Informatics, Computer Equipment and Automatization, 2003, 39, 41-62.

Amari, S.-i. and H. Nakahara (2005). Difficulty of singularity in population coding. *Neural computation* 17(4): 839–858.

Azzimonti, D. and D. Ginsbourger (2016). Estimating orthant probabilities of high dimensional Gaussian vectors with an application to set estimation. *arXiv preprint* arXiv:1603.05031.

Braddick, O. J., K. A. Wishart, and W. Curran (2002). Directional performance in motion transparency. *Vision Res* 42(10): 1237–1248.

Britten, K. H. and H. W. Heuer (1999). Spatial summation in the receptive fields of MT neurons. *Journal of Neuroscience* 19(12): 5074–5084.

Carandini, M. and D. L. Ringach (1997). Predictions of a recurrent model of orientation selectivity. *Vision research* 37(21): 3061–3071.

Cortes, J. M., D. Marinazzo, P. Series, M. W. Oram, T. J. Sejnowski, and M. C. W. van Rossum (2012). The effect of neural adaptation on population coding accuracy. *J Comput Neurosci* 32(3): 387–402.

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

Cramér, H. (1946). *Mathematical Methods of Statistics.* NJ: Princeton Univ. Press.

Deneve, S., P. E. Latham, and A. Pouget (1999). Reading population codes: a neural implementation of ideal observers. *Nat. Neuro.* 2: 740–745.

Ecker, A. S., P. Berens, A. S. Tolias, and M. Bethge (2011). The effect of noise correlations in populations of diversely tuned neurons. *J Neurosci* 31(40): 14272–14283.

Edwards, M. and J. A. Greenwood (2005). The perception of motion transparency: A signal-to-noise limit. *Vision Research* 45(14): 1877–1884.

Gawne, T. J. (2008). Stimulus selection via differential response latencies in visual cortical area V4. *Neurosci Lett* 435(3): 198–203.

Gawne, T. J. and J. M. Martin (2002). Responses of Primate Visual Cortical V4 Neurons to Simultaneously Presented Stimuli. *J Neurophysiol* 88: 1128–1135.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comput. Graph. Statist.* 1(2): 141–149.

Genz, A. (1998). MVNDST: Software for the numerical computation of multivariate normal probabilities, available from web page at http://www. sci. wsu. edu/math/faculty/genz/homepage.

Jazayeri, M. and J. A. Movshon (2006). Optimal representation of sensory information by neural populations. *Nat Neurosci* 9(5): 690–696.

Kay, S. (1993). *Fundamentals of statistical signal processing: Estimation theory.* Prentice-Hall, NJ.

Keemink, S. W. and M. C. W. van Rossum (2016). A unified account of tilt illusions, association fields, and contour detection based on Elastica. *Vision Res* 126: 164–173.

Lampl, I., D. Ferster, T. Poggio, and M. Riesenhuber (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. *J Neurophysiol* 92(5): 2704–2713.

Li, K., V. Kozyrev, S. Kyllingsbæk, S. Treue, S. Ditlevsen, and C. Bundesen (2016). Neurons in primate visual cortex alternate between responses to multiple stimuli in their receptive field. *Frontiers in Computational Neuroscience* 10: 141.

Marshak, W. and R. Sekuler (1979). Mutual repulsion between moving visual targets. *Science* 205(4413): 1399–1401.

Moreno-Bote, R., J. Beck, I. Kanitscheider, X. Pitkow, P. Latham, and A. Pouget (2014). Information-limiting correlations. *Nat Neurosci* 17(10): 1410–1417.

Oleksiak, A., P. C. Klink, A. Postma, I. J. M. van der Ham, M. J. Lankheet, and R. J. A. van Wezel (2011). Spatial summation in macaque parietal area 7a follows a winner-take-all rule. *J Neurophysiol* 105(3): 1150–1158.

Orhan, A. E. and W. J. Ma (2015). Neural population coding of multiple stimuli. *The Journal of Neuroscience* 35(9): 3825–3841.

Pilarski, S. and O. Pokora (2015). On the Cramér–Rao bound applicability and the role of Fisher information in computational neuroscience. *Biosystems* 136: 11–22.

Rao, C. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society* 37(81-89).

Rao, C. (2008). Cramèr-Rao bound. *Scholarpedia* 3(8): 6533. doi:10.4249/scholarpedia.6533.

Rauber, H.-J. and S. Treue (1998). Reference repulsion when judging the direction of visual motion. *Perception* 27(4): 393–402.

Rauber, H.-J. and S. Treue (1999). Revisiting motion repulsion: evidence for a general phenomenon? *Vision research* 39(19): 3187–3196.

Salinas, E. and L. F. Abbott (1994). Vector reconstruction from firing rates. *J. of Comput. Neurosc.* 1: 89–107.

Schwartz, O., A. Hsu, and P. Dayan (2007). Space and time in visual context. *Nat Rev Neurosci* 8(7): 522–535.

Seriès, P., A. Stocker, and E. Simoncelli (2009). Is the homunculus "aware" of sensory adaptation? *Neural Comput* 21: 3271–3304.

Shamir, M. (2014). Emerging principles of population coding: in search for the neural code. *Curr Opin Neurobiol* 25: 140–148.

Shamir, M. and H. Sompolinsky (2006). Implications of neuronal diversity on population coding. *Neural Comput* 18(8): 1951–1986.

Sompolinsky, H., H. Yoon, K. Kang, and M. Shamir (2002). Population coding in neuronal systems with correlated noise. *Phys. Rev E* 64: 51904.

Stocker, A. A. and E. P. Simoncelli (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci* 9(4): 578–585.

Treue, S., K. Hol, and H. J. Rauber (2000). Seeing multiple directions of motion-physiology and psychophysics. *Nat Neurosci* 3(3): 270–276.

van Wezel, R. J., M. J. Lankheet, F. A. Verstraten, A. F. Marée, and W. A. Grind (1996). Responses of complex cells in area 17 of the cat to bi-vectorial transparent motion. *Vision research* 36(18): 2805–2813.

Wei, X.-X. and A. A. Stocker (2015). A Bayesian observer model constrained by efficient coding can explain'anti-Bayesian'percepts. *Nature Neuroscience* 18(10): 1509–1517.

Williams, C. K. and C. E. Rasmussen (2006). Gaussian processes for machine learning. *the MIT Press* 2(3): 4.

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

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

Zemel, R. S. and P. Dayan (1999). Distributional population codes and multiple motion models. *Advances in neural information processing systems* pp. 174–182.

Zemel, R. S., P. Dayan, and A. Pouget (1998). Probabilistic interpretation of population codes. *Neural Comput* 10(2): 403–430.

Zhang, K. and T. J. Sejnowski (1999). Neuronal Tuning: to sharpen or to broaden? *Neural Comp.* 11: 75–84.