Inferring decoding strategy from choice probabilities in the presence of noise correlations

Ralf M. Haefner*[123], Sebastian Gerwinn[23], Jakob H. Macke[234], and Matthias Bethge[235]

[1] Volen National Center for Complex Systems, Brandeis University, Waltham, MA, USA
[2] Centre for Integrative Neuroscience, Tübingen, Germany
[3] Max-Planck-Institute for Biological Cybernetics, Tübingen, Germany
[4] Gatsby Computational Neuroscience Unit, UCL, London, United Kingdom
[5] Bernstein Centre for Computational Neuroscience, Tübingen, Germany

Abstract

The activity of cortical neurons in sensory areas covaries with perceptual decisions, a relationship often quantified by choice probabilities. While choice probabilities have been measured extensively, their interpretation has remained fraught with difficulty. Here, we derive the mathematical relationship between choice probabilities, read-out weights and noise correlations within the standard neural decision making model. Our solution allows us to prove and generalize earlier observations based on numerical simulations, and to derive novel predictions. Importantly, we show how the read-out weight profile, or decoding strategy, can be inferred from experimentally measurable quantities. Furthermore, we present a test to decide whether the decoding weights of individual neurons are optimal, even without knowing the underlying noise correlations. We confirm the practical feasibility of our approach using simulated data from a realistic population model. Our work thus provides the theoretical foundation for a growing body of experimental results on choice probabilities and correlations.

1 Introduction

Understanding how external stimuli give rise to sensory percepts and how individual sensory neurons support this process remain central questions of systems neuroscience. One of the crucial requirements for the claim that a particular group of neurons plays a critical role in the generation of a perceptual event is that "Fluctuations in the firing of some set of the candidate neurons to the repeated presentation of identical external stimuli should be predictive of the observer's judgement on individual stimulus presentations" ([29]). Such correlations between the noise fluctuations in a single neuron's firing rate and the subject's perceptual decision have been found in many areas (e.g. V1: [15], V2: [25], IT: [35], MT: [3,11,12,15,20,21,28,30,34], MST: [5], VIP: [11]). They are usually quantified as choice probabilities ([3]). The quantitative interpretation of choice probabilities has been problematic, however, since their connection to the read-out weight of a neuron is confounded by correlations among the sensory neurons ([31]). For instance, a neuron that itself is not

*Corresponding author (ralf.haefner@gmail.com)

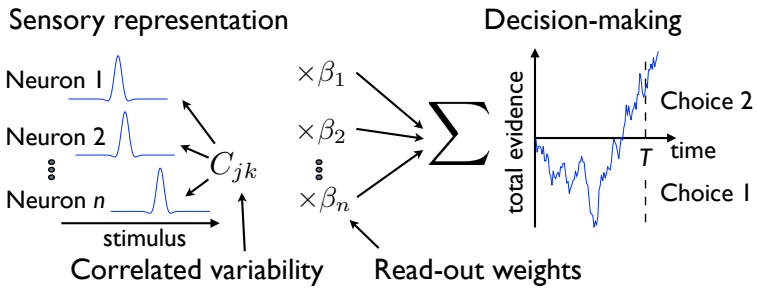

Figure 1: **Model setup:** The weighted activity of a pool of sensory neurons with stimulus-tuned responses and correlated noise is summed linearly by the decision area. At the time of the decision, a choice is triggered depending on the accumulated evidence. Right side adapted from [14].

directly involved in a decision, i.e. has a read-out weight of zero, might display a significant choice probability purely due to the fact that it is correlated with another sensory neuron that does directly contribute to the decision, i.e. that has a non-zero read-out weight ([10]). This indicates the importance of correlated noise and in particular its structure, something that had been recognized early on ([31]) and was highlighted again more recently in a review by [24]. A central challenge for all studies of choice probability to date is that this relationship between correlation structure and choice probabilities has not been characterized analytically. Therefore all previous studies are based on numerical simulations in which the key parameter, the correlation matrix, is very high-dimensional – quadratic in the number of neurons in the considered population. This makes it infeasible to exhaustively explore the behavior of the system, fit it to empirical data, draw conclusions about the incompatibility of a particular model with a set of data, or indeed acquire a deep understanding of the relationship between choice probabilities, sensory encoding and decision-making.

Here, we mathematically derive the relationship between noise correlations, choice probabilities and read-out weights. We find surprisingly simple relationships that make explicit which aspects of the correlation matrix affect choice probabilities and in what way. This allows us to analytically prove earlier numerical results as well as a recent conjecture by [24] about how choice probabilities depend on correlations in a special case. Further, we apply our framework to two correlation structures – a simplified one that modelers have used extensively in the past and one that is based on empirical observations – and show how choice probabilities and decoding weights are linked. Notably, we show how our analytical framework allows us to invert this relationship in order to compute the decoding weight profile from empirically observed neuronal correlations and choice probabilities. Finally, we derive a test for whether the read-out mechanism is optimal – even in the absence of any knowledge about noise correlations.

2 Results

2.1 Analytical link between choice probabilities, noise correlations and read-out weights

We start by describing the mathematical framework for computing choice probabilities from read-out weights and noise correlations. We follow previous studies in modeling the decision-making circuit by assuming that a decision-making area linearly reads out and accumulates the activity of

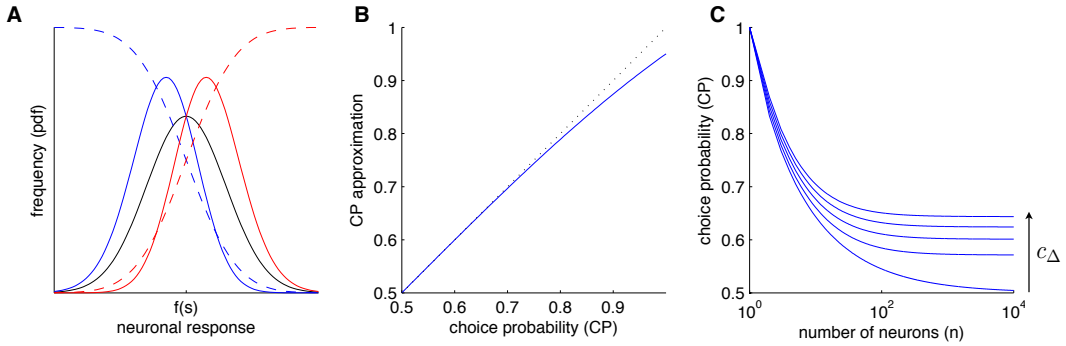

Figure 2: **A:** Choice-conditioned response distribution. Black: Distribution of sensory response distribution across all trials. Red & Blue: Sensory response distributions only for those trials in which the subject selected choice 1 and 2, respectively. They are obtained by a point-wise multiplication of the total distribution (black) with cumulative Gaussians (shown in dashed). **B:** Comparison of choice probability (CP) computed from eq (1) with its first-order approximation eq (2). The dotted line is the identity line. **C:** CP is shown as a function of the number of neurons n, for different levels $c_\Delta = (0, 0.05, 0.1, 0.15, 0.2)$, the difference between the correlations within and across decision pools, respectively. $\mathrm{CP}(n \to \infty) \approx 0.5 + \sqrt{c_\Delta}/\pi$

a population of neurons in a sensory area ([14, 24, 31]). Our setup is illustrated in Figure 1. On the left we illustrate the representation of a stimulus s by a population of sensory neurons $k = 1..n$ with tuning functions $f_k(s)$. Their activity is weighted by read-out weights β_k and accumulated over time. A decision is made at time T, depending on the final value of the accumulated activity (formulas provided in Online Methods). We assume the sensory responses to be variable and that this variability is correlated, as observed in cortex. We model this correlated variability by a multivariate Gaussian distribution with covariance matrix $\mathbf{C}$, where C_{jk} is the covariance between the responses of neuron j and neuron k, and C_{kk} is the variance of neuron k. $\mathbf{C}$ is also called the noise covariance matrix, and, when normalized by the variance, the noise correlation matrix.

Consider now the case of an ambiguous stimulus, one that contains equal evidence for either decision. Examples may be a stimulus consisting of moving dots, with equal numbers in the two directions that are to be distinguished ([3, 23]); or a stereo image with as much evidence in favor of the stimulus being behind as being in front of the fixation point ([25, 26]). Each sensory neuron will fire with a mean firing rate determined by its tuning function and a certain variability which we model as Gaussian distributed. If we split the trials into two groups – those that led to decision 1 and those that led to decision 2 – and determine the distribution of firing rates, we may find that they are slightly different for each group. One is shifted to lower, and the other shifted to higher firing rates (see Figure 2A). Such an effect has been observed in several sensory areas in cortex (for a review see [24]) and its strength is usually quantified as "choice probability" ([3]). More precisely, choice probability is defined as the probability that a random sample from the red distribution is indeed larger than a random sample from the blue distribution. It is 0.5 if both distributions are identical and increases up to 1 as they are more and more separated. Since the stimulus contains only negligible evidence, or equal evidence for either choice, the stimuli eliciting the responses in either group are essentially identical, i.e. they are not responsible for the differential firing distributions. It is tempting to link this separation, i.e. choice probability, to the read-out weight assigned to this

neuron. In fact, in the absence of any noise correlations, that link would be straightforward. We could infer the read-out weight for a particular neuron from its choice probability alone: the larger the choice probability, the larger the neuron's read-out weight. However, in the presence of noise correlations, this is no longer true: a neuron might be assigned a zero read-out weight and show a large choice probability only because it is correlated with a neuron with non-zero read-out weight.

We have derived an analytical relationship linking choice probabilities and read-out weights in the presence of arbitrary correlations, and find that it has a surprisingly simple form (see Methods):

$$\mathrm{CP}_k = \frac{1}{2} + \frac{2}{\pi} \arctan \frac{(\mathbf{C}\boldsymbol{\beta})_k}{\sqrt{2C_{kk}\,\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta} - (\mathbf{C}\boldsymbol{\beta})_k^2}}. \tag{1}$$

Here, CP_k is the choice probability of neuron k, which depends on the high-dimensional noise covariance matrix only through three numbers:

C_{kk}: the response variance of neuron k

$(\mathbf{C}\boldsymbol{\beta})_k = \sum_{j=1}^{n} C_{jk}\beta_j$: the sum over all covariances of neuron k with all other neurons, weighted by their read-out weights

$\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta} = \sum_{k=1}^{n}\sum_{j=1}^{n} \beta_k C_{jk} \beta_j$: the total variance summed across all neurons, weighted with the respective read-out weights.

A simplification to eq. (1), allowing for an easier intuition, is given by its first order approximation[1]:

$$\mathrm{CP}_k \approx \frac{1}{2} + \frac{\sqrt{2}}{\pi} \frac{(\mathbf{C}\boldsymbol{\beta})_k}{\sqrt{C_{kk}\,\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta}}}. \tag{2}$$

The error incurred by this approximation is very small – ranging from zero at $\mathrm{CP} = 1/2$ to 5% for $\mathrm{CP} = 1$. For clarity of exposition we will use this approximation to present our results for the CP in various scenarios below. The actual computations for the figures will be performed by using the exact result in eq. (1). Figure 2B compares the choice probability as computed according to equations (1) and (2) and shows the negligible error due to the approximation.

Eq (2) exposes the intrinsic linearity of the relationship. $\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta}$ is simply a normalization constant that affects the overall magnitude of the CPs in a population, but does not affect individual neurons differentially. This means that the CPs, after subtracting 1/2 and adjusting them for the intrinsic response variability C_{kk} of each neuron, is given by the covariance matrix $\mathbf{C}$ times the read-out weights vector $\boldsymbol{\beta}$. This means that in addition to its own weight and response variance, the CP of any neuron depends on its covariance with all the other neurons in the population and their weights. It also follows immediately from the fact that only 1 of n terms in the sum $(\mathbf{C}\boldsymbol{\beta})_k = \sum_{j=1}^{n} C_{jk}\beta_j$ contains β_k, that as the number n of neurons in a positively correlated population increases, its own read-out weight has less and less influence on the CP of a neuron.

2.2 Choice probabilities for special cases

We first apply our result to the most widely discussed model in the literature, that of [31]. It assumes that the population of sensory neurons consists of two groups of neurons with opposing stimulus

[1] in $(\mathbf{C}\boldsymbol{\beta})_k/\sqrt{C_{kk}\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta}}$ around $1/2$.

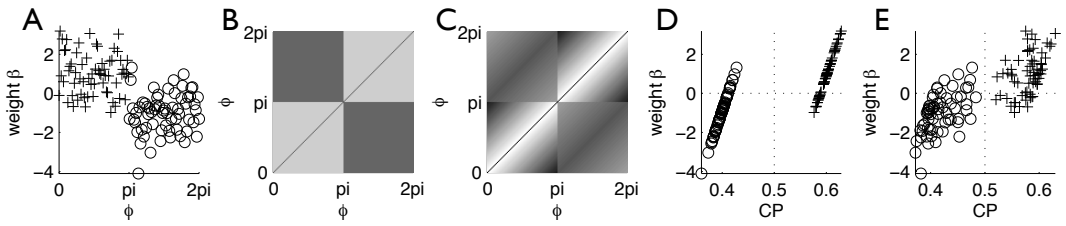

Figure 3: **Relationship between CPs and read-out weights β_k for two example correlation structures: A:** Weights sampled randomly around 1 and -1, respectively, depending on the pool (symbols '+' and 'o'). **B:** Correlation matrix with constant within and across pool correlations of 0.1 and 0, respectively. **C:** Limited range correlations, separate within and across pools. Based on empirical findings of [9]. For the exact dependency of correlations on $\Delta\phi$ see Figure 4A. **D:** Relationship between CPs and weights for the correlation matrix in panel B. Symbols '+' and 'o' represent different pools as in panel A. Linear relationship within each pool. **E:** Relationship between CPs and weights for the correlation matrix in panel C.

preferences. The activity within each group is averaged before the two means are subtracted from each other. In our terms this corresponds to constant weights of 1 for $n/2$ of the neurons, and -1 for the other $n/2$ neurons. The noise correlation between neurons within each pool is assumed to be $c_{||}$ and between neurons of opposing pools $c_{\perp}$. Evaluating eqs (1 and 2) now simply involves summing over entries in the noise covariance matrix to yield (see Methods):

$$\mathrm{CP}_k \approx \frac{1}{2} + \frac{\sqrt{2}}{\pi}\sqrt{\frac{1}{n} + \frac{1}{2}c_{\Delta}} \quad \text{with} \quad c_{\Delta} = \frac{n-2}{n}c_{||} - c_{\perp} \tag{3}$$

We find that the CP of a neuron only depends on the number of neurons involved in the decision and the difference between the mean correlations within a pool, $c_{||}$, and across pools, $c_{\perp}$. Figure 2C shows the CP as a function of the number of neurons and c_{Δ}. Numerically, this relationship was first reported by [31] for the special case of zero correlations between neurons in different pools, however its parametric form had not been known. Our result also shows that a recent conjecture by [24] – again based on numerical simulations – that CPs depend on the correlation structure only through $c_{||} - c_{\perp}$ is true for large neuronal populations. For small numbers of neurons the correction shown in eq. (3) needs to be applied. Based on our analytical relationship, it is easy to generalize these findings to arbitrary correlation structures simply replacing $c_{||}$ and $c_{\perp}$ by their averages $\langle c_{||} \rangle$ and $\langle c_{\perp} \rangle$.

For the rest of this paper we present our results in the context of the classic motion discrimination task ([23]) in which a subject has to decide whether the net motion in a random dot kinematogram is in one or the opposite direction. Our neuronal population consists of n neurons with preferred directions ϕ between 0 and 2π. Our convention will be that the two to-be-discriminated directions are $\pi/2$ and $3\pi/2$. This implies that neurons with preferred direction $0 < \phi < \pi$ are in one pool, and neurons with $\pi < \phi < 2\pi$ are in the other pool.

In Figure 3 we illustrate the relationship between read-out weights and CPs for two different noise correlation structures. In both cases we assume random weights within each pool (shown in panel A). First, we posit constant correlations within and across pools (Fig. 3B). This implies a piecewise linear relationship between weights and CPs as shown in panel D. The fact that this relationship is linear within each pool makes it possible potentially to deduce the read-out weight of a neuron

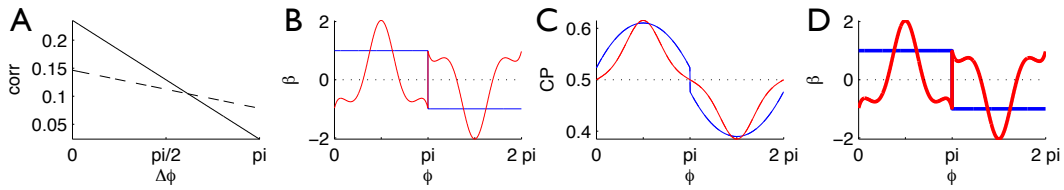

Figure 4: **Reconstruction of weights from CPs and correlations. A:** Correlation structure: limited range correlations within pools (solid) and across pools (dashed) similar to [9]. **B:** Weights β chosen uniformly (blue) and linearly-optimally (red) based on the correlation structure in A. **C:** CPs implied by correlation structure in A and weights in B. **D:** Reconstructed weights from knowledge of correlation structure and weights. Colors as defined in panel B.

from its CP. However, the slope of the lines in Figure 3D increases with the number of neurons in the population. This means that for very large pool sizes, all neurons within a pool have identical CPs, regardless of their actual weight, i.e. even if zero, just as found by [10] (based on numerical simulations.) This means that for large neuronal populations and constant correlations within and across pools, it is *in principle* impossible to reconstruct the read-out weights. However, if we assume a more realistic correlations structure (Fig. 3C), based on the one reported by [9], this convergence of CPs regardless of read-out weight does not happen (see Suppl Materials, Fig. 8). Even so, knowledge of the correlation structure is crucial for inferring the weights from the CPs: in Figure 3E we show the relationship between CPs and weights based on the correlation structure shown in Figure 3C. Without accounting for noise correlations, the relationship between CPs and weights is very weak, and becomes even more so as the number of neurons increases ($n = 256$ for the case shown here).

To summarize: inferring the read-out weights from the CPs is generally impossible in large neuronal populations if the correlations are constant within and across pools. However, for a realistic correlations structure, reconstructing the weights *is* possible, provided the correlations are known. In the next section we will describe how.

2.3 Reconstruction of read-out weights

If the correlation structure is known in addition to the choice probabilities, we can invert eqs (2) and deduce the read-out weights associated with individual neurons directly. We find (see methods):

$$\beta_k \approx \frac{\pi}{\sqrt{2}} \left[\sum_{l=1}^{n} \{\mathbf{C}^{-1}\}_{kl} \sqrt{C_{ll}} (\mathrm{CP}_l - \frac{1}{2}) \right] \tag{4}$$

where $\mathbf{C}$ is the noise covariance matrix of the population. Essentially, the vector of read-out weights is the vector of choice probabilities multiplied with the inverse of the correlation matrix. From this follows immediately that if the noise correlations between the neurons are zero, the weights are directly proportional to the CPs. (For more details and the precise version of eq (4) based on eq (1) see eq (14) in the Methods.)

Figure 4 illustrates this framework. Panel A shows the noise correlations depending on the difference $\Delta\phi$ in preferred orientation of any two neurons separated by whether they belong to the same pool (solid line) or different pools (dashed lines). The noise correlations we use are linear fits to the

data in [9]. For this example we investigate two different read-out weight profiles: constant weights within each pool (blue lines in panel B) and the optimal weights implied by the correlation structure (red lines). For the case of optimal decoding, within each pool, neurons far away from the decision direction are subtracted from those whose preferred direction is close to the decision direction. This improves decision-making by subtracting (positively correlated) noise while leaving the signal largely unchanged ([6]). Panel C shows the implied CPs for all the neurons in our model. For both cases – constant and optimal weights – the magnitude of the CPs of those neurons close to a decision direction are largest, while those close to the decision boundaries at $\phi = 0$ and $\phi = \pi$ are smallest. (Since both pools and their results are symmetric with respect to each other, we will only discuss the left pool with $0 < \phi < \pi$.) More importantly, the CPs are different for the two cases, allowing us to differentiate between the two scenarios based on empirical data. The difference in CP between the two decoding strategies is largest for neurons close to a decision boundary: while constant weights imply choice probabilities larger than 0.5, the optimal read-out scheme forces them to be exactly 0.5 at the decision-boundary. This implies that empirical data about the neurons closest to a decision boundary is going to be the most informative for distinguishing between constant and optimal read-out weights. From this figure we again see that the relationship between CPs and weights is far from trivial: CPs greater than 0.5 do not necessarily indicate a positive contribution to the decision, but the neuron may in fact have negative weights (optimal weight profile). At the same time, neurons differing in CPs may have identical weights (constant weight profile). But most importantly, and in contrast to the case of constant correlations within and across pools, CPs are not independent of the read-out weights, so that the read-out weights can be reconstructed from the CPs given knowledge about the underlying noise correlations. Figure 4D shows how the weights are correctly reconstructed from the CPs in panel C using eq 4.

After having gained a better conceptual understanding, we now put our framework to test in the more scenario of a non-smooth covariance matrix and heterogeneous tuning curves, i.e. neurons with differing response properties. As before, we use a correlation structure similar to the one reported by [9]. But instead of assuming that any variability in correlation coefficients at a given difference in preferred orientation $\Delta\phi$ is due to measurement noise, we allow for an actual underlying variability among the correlation coefficients – illustrated in Figure 5A. Its entries are sampled around a mean that corresponds to Figure 4A and 3C. In order to increase visibility, only 128 neurons in the simulated population are displayed in panels A-C of this figure. Furthermore, we assumed a high degree of tuning curve variability in terms of response amplitude and – through a Poisson-like mean-variance relationship – response variability (details in Methods). Panel B shows the weights for the two models that we consider here. One set of weights (in blue) has been sampled from a Gaussian distribution around a constant mean for each pool (mean shown in blue in Fig. 3B). The other set of weights (in red) are the optimal weights implied by the particular correlation matrix in Figure 5A and the particular tuning curves. The large variability among the optimal weights is due to the variability in correlation coefficients and particularly in tuning curves for nearby neurons. Those two sets of weights imply CPs as shown in Figure 5C. They vary around the CPs implied by the homogenous case discussed above, however with less variability around their mean than found among the weights themselves.

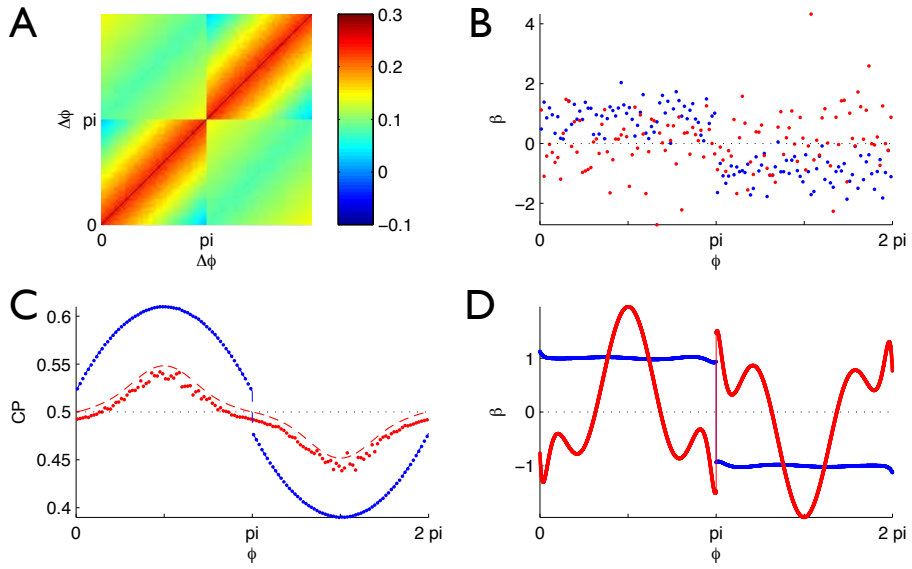

Figure 5: **Reconstruction of weights from limited data in the case of heterogenous weights and noisy correlation structure from CPs and correlations with heterogenous weights and limited data. A:** Noisy correlation structure around means as in Figure 4. **B:** Weights β chosen randomly around a uniform mean (blue) and linearly-optimally (red) based on the noisy correlation structure in A. **C:** Implied CPs associated with correlation structure in A and weights in B. Dashed red line is a parametric fit of the noisy CPs. **D:** Reconstructed weights assuming the mean of the correlation structure in A and the fits to the noisy CPs in C. Colors as defined in panel B.

While a precise reconstruction of every individual weight is possible in cases where every neuron in the population can be recorded from, this is not realistic in cortex given current technology. Instead, we demonstrate in the following how measurements from a small subset of observed neurons can be used to reconstruct an "average model". The central idea is to reconstruct an "averaged" weight profile which explains the observed mean correlations and the observed mean CPs. First, we fit a smooth correlation function to the measured data. For this demonstration, we use 64 "recorded" pairs of neurons (out of 1024 for the entire population) assuming that the correlation coefficients only depend on the difference in preferred orientation ([9]). Next, we fit a smooth function to the observed CPs in order to extrapolate from our sparse measurements to the entire population (for our example we fit polynomials to the CPs in each pool, assumed to be symmetric round $\phi = \pi$ – shown as dashed lines in Figure 5C). We can now use the smooth approximations to correlation matrix and CPs to reconstruct the weights profile – results shown in Figure 5D. The profile reconstructed from the noisy constant weights (in thick blue) is itself largely constant across almost the entire range of directions, with noteworthy deviations only very close to the decision boundaries. These deviations are an illustration of the fact that the reconstructed weights are particularly sensitive to approximation errors close to the decision-boundaries (or generally discontinuities in the system) and therefore need to be interpreted with care. The weight profile reconstructed for the system with optimal weights (Fig. 5D in thick red) is very similar to the optimal profile based on the observed means of the noisy correlation matrix in A with its characteristic negative weights close to the decision-boundary (compare with 'ground truth' in Fig. 4B&D). This means that even for a very

large heterogenous population of neurons, measurements of only a small fraction of the neurons can be used to reconstruct the decoding weight profile.

Figure 5 also illustrates another point: the CPs observed for the optimal weights are much smaller in the heterogenous case than the homogenous situation (see Fig. 4). Based on the homogenous case, the peak CPs is over 0.6 while in the heterogenous case, it is about 0.55 for the optimal weights. The solid lines in Figure 6A replot the smooth fits to the actual CPs for the heterogenous case from Figure 5C. For comparison, we have now computed the CPs implied by the smooth reconstructed weights shown in Figure 5D – shown as dashed lines in Figure 6A. We find that the CPs implied by an averaged, or smooth, model are much larger than those that would actually be observed. The reason for the discrepancy is the $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}$ term in the denominator of the CP formula (2). Averaging away variability in the $\boldsymbol{\beta}$ typically decreases $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}$. For the amount of heterogeneity assumed in our simulation, $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}$ decreases by a factor of roughly 5 when calculated from the averaged weights (for details see Methods). This means that the CPs in our averaged system, based on the averaged correlations and averaged weights, would be larger than actually observed by a factor of about $\sqrt{5} \approx 2.2$ (eq. (4)). Note that the CP profile is unchanged, only the magnitude is greater. It also does not mean that the reconstruction is wrong, it simply implies that we need to take the underlying neuronal heterogeneity into account to explain the (simulated) data. Empirical data from [10] is overlaid in black in Figure 6A. This data was measured concurrent to the correlation data underlying our simulation ([9]). Without taking the underlying neuronal variability into account, a model based only on the data means would not be able to bring the measured correlations into agreement with the measured CPs – at least not without invoking an additional noise term at the decision stage that would have to be more influential than all the sensory evidence together. To summarize: we have just shown that CPs are not only determined by the absolute magnitude of the correlations, but also the variability in the underlying read-out weights, which in turn may depend on the variability in response properties (as for an optimal read-out). Furthermore, based on the data from [9, 10] we conclude that this variability is likely to be crucial for explaining existing empirical data.

2.4 Optimality test

We now show how to use our framework to answer the question whether the weights are optimal, even if the correlation structure is *not* known. This is possible because the mathematical formula that relates optimal weights to the correlation structure is very similar to the one relating CPs to the correlation structure. The optimal weights for a linear discriminator are given by the inverse of the covariance matrix multiplied by the difference between the responses to the two stimuli, s_1 and s_2, that are to be distinguished ([2]):

$$\boldsymbol{\beta}^{\mathrm{optimal}} \propto \mathbf{C^{-1}}[\mathbf{r}(s_1) - \mathbf{r}(s_2)].$$

At the same time, from eqs. (2 & 4) it follows that the actual read-out weights are essentially the inverse of the covariance matrix multiplied by the CPs. Hence, optimal read-out weights predict that the CPs are proportional to the difference between the population responses to the two stimuli that are to be distinguished:

$$\mathrm{CP}_k - \frac{1}{2} \propto \frac{r_k(s_1) - r_k(s_2)}{\sqrt{C_{kk}}}. \tag{5}$$

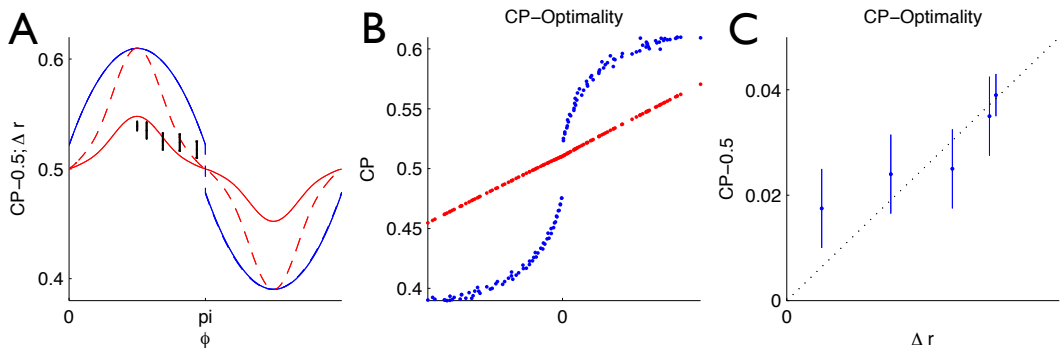

Figure 6: **Application to empirical data: A:** CP data from [10] in black (errorbars represent 1 standard error of the mean). CPs based on constant weights in blue, based on the optimal linear decoding weights in red. The solid red line corresponds to the actual CPs that would be measured in the heterogenous case considered above. The dashed red line is the CP prediction based on the averaged 'data' from the same simulation. **B:** Optimality test from eq. (5) applied to the simulated data in Figure 5. In blue the data based on a model with weights sampled randomly around a constant within each pool. In red the data for a model with optimal weights. **C:** Optimality test applied to average data from [10]. Dotted black line indicates proportionality.

In words: for a single neuron k, the CP minus 0.5 should be proportional to the difference between its neuronal responses to the two stimuli, divided by the square root of its response variance. This relationship is shown for our model case in panel F of Figure 5. The correlation is perfect for the optimal linear decoder weights (red) and has characteristic deviations for the constant weights case (blue). The beauty of this relationship is that it involves only quantities that can easily be measured by traditional techniques (e.g. individual extracellular recordings).

In Figure 6B we apply eq. (5) to the results from our realistic simulation in the previous section. We see that despite the significant neuron-to-neuron variability in terms of tuning curves, response variance, and correlation coefficients, the data based on a mode with optimal weights form a perfectly straight line (red). If the underlying weights are roughly constant, however, the relationship shows characteristic deviations from a straight line (blue). In panel C of the same figure we re-plot the averaged data underlying panel A in the same form as for panel B. The data is not significantly different from the proportionality line implying that optimality cannot be statistically excluded. Remember, however, that we found in the previous section, that the published average data cannot rule out constant weights either. We emphasize also that applying eq. (5) to averaged data ignores the main power of the test which applies to individual neurons, taking their individual tuning curves and response variability into consideration. These quantities have not always been published, probably in part because of the questions concerning their interpretation. However, most existing studies on CPs will have recorded these values for each of their neurons so that testing the hypothesis of an optimal read-out scheme with respect to amplitude and variability should be possible in many of them.

3 Discussion

The discovery of [3] that the response of individual sensory neurons is correlated with the animal's behavior even when there is no signal in the stimulus has since been replicated for different decision paradigms and in different sensory cortices ([3, 5, 11, 12, 12, 15, 20, 21, 25, 28, 30, 34, 35] with reviews in [14, 24]). Over the same time, the appreciation has grown for the crucial role of noise correlations in the brain for our understanding of sensory coding ([1, 8]). With this study we link those two concepts in an analytical framework. We show how choice probabilities, noise correlations and read-out weights are related and how fixing two of these quantities determines the third one. There are clear benefits of an analytical framework over numerical simulations (as used in previous studies): it allows us to fit parameters, to investigate much larger and realistic populations sizes by cutting computation time by several magnitudes, to mathematically extrapolate to arbitrarily large populations and to make rigorous statements about which types of models may or may not be compatible with the data.

We furthermore used our framework to derive a test for whether the read-out of sensory neurons is optimal with respect to tuning curves, response variability and noise correlations – *without* requiring knowledge of noise correlations that are difficult to measure. We find that on the basis of existing data that has only been published in an averaged form, it is not possible to distinguish between two major hypotheses about sensory read-out: constant weights within each pool, and linearly optimal weights. Applying this test to already existing datasets in their raw form should reveal whether decision neurons 'know' about tuning curves, response variability and noise correlations of their input neurons and account for them when pooling sensory information.

At this point there is little data available that combines a measurement of interneuronal correlations with choice probabilities in a decision-making task ([9, 10, 17]). However, with the increased use of population recording techniques ([4, 19, 33]), we expect more such data to become available very soon and our framework will be readily applicable to infer decoding weights and shed light on the decoding strategies of the brain in various contexts.

3.1 Model assumptions

Our analytical framework assumes a population of sensory neurons whose responses are variable, correlated, and linearly read out – assumptions that it shares with most previous studies on perceptual decision-making (reviews in [14, 24]). We furthermore assume that the noise is additive and Gaussian. Since in the scenario that we model, spike rates are approximately constant from trial to trial (although not from neuron to neuron), any observed mean-dependent response variance of cortical neurons (e.g. Poisson) is covered by our model. It only breaks down when spike counts over the timespan during which the decision is made are very low. Our framework implies few assumptions about the temporal integration process and our findings hold even if the integration is not perfect and weights effectively change over the time of the integration, as in integration-to-bound ([14]) or attractor-based read-out models ([36]). In both cases, only the overall magnitude of the weight profile might change over time, or from trial to trial, something to which choice probability is invariant to.

3.2 Role of correlations

A topic of recent debate has been the origin of choice probabilities ([24]). We emphasize that while our model appears explicitly bottom-up or feedforward, it is in fact agnostic about the source of its main ingredient, noise correlations. While some noise correlation structures have traditionally been explained in a bottom-up way ([32]), correlations that depend on whether neurons belong to the same or to different pools, might more plausibly be explained by top-down influences like fluctuations in attention-like processes ([9]). On the other hand, collective processes like attractor dynamics within a sensory area might also contribute to the correlations within a sensory area.

Arguably, the most remarkable result of our study is that choice probabilities are useful at all in inferring read-out weights in very large populations of neurons. This appeared impossible after [10] demonstrated that choice probabilities for a group of neurons with zero weights, and those for a group of neurons with positive weights, converged to the same value as the overall number of neurons became very large. This seemed to imply that it would be practically impossible to reconstruct which of the neurons contributed to a decision and which did not. The difference between our case and the one previously studied numerically, is that [10] for their simulation assumed a quadrant-wise constant correlation structure, while we work with one that is not. Our analytical equations apply to all correlations structures, of course. However, quadrant-wise constant correlations define a mapping from weights to choice probabilities that is not invertible in the limit of infinitely many neurons. For our reconstruction we averaged over all neuronal dimensions other than preferred direction and only reconstruct average weights along those "averaged" dimensions. This means that we, too, cannot distinguish between neurons of different weights but same preferred direction. However, since our correlation matrix along the dimension of interest – preferred direction – is not constant, but has a profile that implies an invertible mapping for arbitrarily many neurons, it is possible to reconstruct the decoding strategy by inverting eq (1). Such a non-constant relationship between sensory dimension and noise correlation has been found by many empirical studies (usually reported as a relationship between signal correlations and noise correlations – reviewed in [8]) in addition to [9] on whose data our examples were based. Generally, our work implies that in any dimension in which the empirical correlation matrix is invertible for infinitely many neurons, one can recover the weight profile from the choice probabilities. While originally noise correlations were primarily seen as a complicating factor in inferring read-out weights from choice probabilities, our work implies that it is only because of their presence and their observed structure, that we can infer read-out weights at all.

3.3 Neuronal heterogeneity

Physiologists have long known that response properties of sensory neurons are very heterogenous. Our work shows that accounting for this heterogeneity in theoretical models is important. The magnitude of the observed choice probabilities depends strongly on the amount of heterogeneity in the underlying weights. Simply averaging over variability in the data will lead to models that are not self-consistent. The main drivers for the discrepancy in our model are the number of neurons in the population and the variability in their tuning curve amplitudes divided by the square root of their response variability. Since the latter quantities are easily accessible empirically (and are likely already available in existing studies), quantitatively fitting a model to data even holds the promise

of determining the size of the pool of sensory neurons contributing to the decision. Our study adds to a growing body of literature that emphasizes the importance of reporting and modeling that heterogeneity – for understanding low-level function ([22]), sensory ([13, 27]) and motor ([7]) circuits. This heterogeneity is usually discarded in favor of summary statistics most relevant to the focus of the particular study. Our findings – both with respect to weight reconstruction as well as with respect to the optimality test – emphasize that valuable information and understanding can be gained by reporting on and modeling that underlying variability.

3.4 Optimality test

The right hand side of eq (5) is sometimes called neuronal sensitivity and a positive correlation between choice probabilities and neuronal sensitivity has been observed in several studies ([3,5,16, 28,30]). It was alternatively attributed to increased correlations between the most sensitive neurons ([31]) or to a preferential read-out of the most informative neurons based on intuition and numerical simulations (e.g. [18,20]). Here we derive this relationship mathematically for the case of an optimal read-out code. Comparing predictions for a model with constant weights and those for a model with optimal weights we find that the correlation between neuronal sensitivity and choice probability is positive in both cases. If the correlation coefficient is 1, the read-out weights are optimal, and vice versa: optimal weights imply a correlation of 1. Surprisingly, a positive correlation of less than 1 cannot be taken as evidence that more sensitive neurons are weighted preferentially as has previously been assumed ([18]) since even an indiscriminate pooling model implies a positive correlation (Fig. 6B). However, it is likely that an increased correlation between sensitivity and choice probabilities, given unchanged noise correlations, (e.g. over time as in [20]) is indicative of a more selective weighting based on sensitivities, but whether this is always true needs to be investigated in more detail. By providing the exact relationship expected from an optimal read-out code, we provide the basis not only for rigorous statistical tests, but also for addressing the question of whether decision neurons take into account response amplitude, response variability, and/or correlations when reading out sensory neurons.

Acknowledgments

We thank A. Ecker and P. Berens for stimulating discussions and detailed comments on the manuscript. RMH is supported by a Swartz Fellowship and JHM is supported by an EC Marie Curie Fellowship.

References

[1] BB Averbeck, PE Latham, and A Pouget. Neural correlations, population coding and computation. *Nat Rev Neurosci*, 7(5):358–366, 2006.

[2] CM Bishop. Pattern recognition and machine learning. 2006.

[3] KH Britten, WT Newsome, MN Shadlen, S Celebrini, and JA Movshon. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci*, 13(1):87–100, 1996.

[4] G Buzsaki. Large-scale recording of neuronal ensembles. *Nat Neurosci*, 7(5):446–451, 2004.

[5] S Celebrini and WT Newsome. Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. *J Neurosci*, 14(7):4109–4124, 1994.

[6] Y Chen, WS Geisler, and E Seidemann. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nat Neurosci*, 9(11):1412–1420, 2006.

[7] MM Churchland and KV Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *J Neurophysiol*, 97(6):4235–4257, 2007.

[8] MR Cohen and A Kohn. Measuring and interpreting neuronal correlations. *Nat Neurosci*, 14(7):811–819, 2011.

[9] MR Cohen and WT Newsome. Context-dependent changes in functional circuitry in visual area MT. *Neuron*, 60(1):162–173, 2008.

[10] MR Cohen and WT Newsome. Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *J Neurosci*, 29(20):6635–6648, 2009.

[11] EP Cook and JH Maunsell. Dynamics of neuronal responses in macaque MT and VIP during motion detection. *Nat Neurosci*, 5(10):985–994, 2002.

[12] JV Dodd, K Krug, BG Cumming, and AJ Parker. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *J Neurosci*, 21(13):4809–4821, 2001.

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

[14] JI Gold and MN Shadlen. The neural basis of decision making. *Annu Rev Neurosci*, 30:535–574, 2007.

[15] A Grunewald, DC Bradley, and RA Andersen. Neural correlates of structure-from-motion perception in macaque v1 and MT. *J Neurosci*, 22(14):6195–6207, 2002.

[16] Y Gu, DE Angelaki, and GC Deangelis. Neural correlates of multisensory cue integration in macaque MSTd. *Nat Neurosci*, 11(10):1201–1210, 2008.

[17] Y Gu, S Liu, CR Fetsch, Y Yang, S Fok, A Sunkara, GC DeAngelis, and DE Angelaki. Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron*, 71(4):750–761, 2011.

[18] M Jazayeri. Probabilistic sensory recoding. *Curr Opin Neurobiol*, 18(4):431–437, 2008.

[19] JN Kerr and W Denk. Imaging in vivo: watching the brain in action. *Nat Rev Neurosci*, 9(3):195–205, 2008.

[20] CT Law and JI Gold. Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nat Neurosci*, 11(4):505–513, 2008.

[21] J Liu and WT Newsome. Correlation between speed perception and neural activity in the middle temporal visual area. *J Neurosci*, 25(3):711–722, 2005.

[22] E Marder. Variability, compensation, and modulation in neurons and circuits. *Proc Natl Acad Sci U S A*, 108 Suppl 3:15542–15548, 2011.

[23] WT Newsome, KH Britten, and JA Movshon. Neuronal correlates of a perceptual decision. *Nature*, 341(6237):52–54, 1989.

[24] H Nienborg and B Cumming. Correlations between the activity of sensory neurons and behavior: how much do they tell us about a neuron's causality? *Curr Opin Neurobiol*, 20(3):376–381, 2010.

[25] H Nienborg and BG Cumming. Macaque v2 neurons, but not v1 neurons, show choice-related activity. *J Neurosci*, 26(37):9567–9578, 2006.

[26] H Nienborg and BG Cumming. Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, 459(7243):89–92, 2009.

[27] K Padmanabhan and NN Urban. Intrinsic biophysical diversity decorrelates neuronal firing while increasing information content. *Nat Neurosci*, 13(10):1276–1282, 2010.

[28] AJ Parker, K Krug, and BG Cumming. Neuronal activity and its links with the perception of multi-stable figures. *Philos Trans R Soc Lond B Biol Sci*, 357(1424):1053–1062, 2002.

[29] AJ Parker and WT Newsome. Sense and the single neuron: probing the physiology of perception. *Annu Rev Neurosci*, 21:227–277, 1998.

[30] G Purushothaman and DC Bradley. Neural population code for fine perceptual decisions in area MT. *Nat Neurosci*, 8(1):99–106, 2005.

[31] MN Shadlen, KH Britten, WT Newsome, and JA Movshon. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci*, 16(4):1486–1510, 1996.

[32] MN Shadlen and WT Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neurosci*, 18(10):3870–3896, 1998.

[33] IH Stevenson and KP Kording. How advances in neural recording affect data analysis. *Nat Neurosci*, 14(2):139–142, 2011.

[34] T Uka and GC DeAngelis. Contribution of area MT to stereoscopic depth perception: choice-related response modulations reflect task strategy. *Neuron*, 42(2):297–310, 2004.

[35] T Uka, S Tanabe, M Watanabe, and I Fujita. Neural correlates of fine depth discrimination in monkey inferior temporal cortex. *J Neurosci*, 25(46):10796–10802, 2005.

[36] XJ Wang. Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234, 2008.

4 Methods

4.1 Notation, assumptions and derivation

We assume that the decision is based on a linear combination of the responses $\mathbf{r} = (r_1, .., r_n)$ of a population of n sensory neurons:

$$D = \sum_{k=1}^{n} \beta_k r_k \equiv \boldsymbol{\beta}^\top \mathbf{r} \tag{6}$$

where $\boldsymbol{\beta} = (\beta_1, ..\beta_n)$ are the weights with which neurons $1..n$ contribute to the decision. We assume that the decision is unbiased and that a stimulus that does not contain any evidence, or equal amounts of evidence for both choices, implies $\langle D \rangle = 0$. Our convention is that if $D < 0$, choice 1 is elicited, if $D > 0$, choice 2 is initiated.

We further assume that the neuronal responses can be modeled as a multivariate Normal distribution whose means are given by the neuron's tuning functions $\langle r_k \rangle = f_k(s)$. We assume that the neuronal responses can be correlated and denote the noise covariance matrix with $\mathbf{C}$. C_{kk} is the response variance of neuron k and C_{jk} is the covariance between the responses of neuron j and neuron k.

Based on these assumption, we can now derive the choice-conditioned stimulus distribution $P(r_k|D < 0)$ for choice 1, and $P(r_k|D > 0)$ for choice 2. Since

$$P(r_k|D < 0)P(D < 0) = P(r_k, D < 0) = P(D < 0|r_k)P(r_k) \tag{7}$$

and $P(D < 0) = P(D > 0) = 1/2$ (assuming unbiased decision-making) it follows:

$$P(r_k|D < 0) = 2P(r_k)P(D < 0|r_k) \tag{8}$$

and similarly for $P(r_k|D > 0)$. Since we assumed the r_k to be normally distributed, D also is, and therefore $D < 0|r_k$. Applying the formula for the mean and variance of conditional Gaussians, we find:

$$P(D|r_k) = \phi\left[D : C_{kk}^{-1} \sum_{j=1}^{n} \beta_j C_{kj} \delta r, \sum_{j=1}^{n}\sum_{l=1}^{n} \beta_j \beta_l C_{jl} - C_{kk}^{-1}(\sum_{j=1}^{n} \beta_j C_{kj})^2\right] \tag{9}$$

In the above, $\phi[x : \langle x \rangle, \mathrm{var}(x)]$ is the probability density function of the Normal distribution and $\delta r_k = r_k - f_k(s)$ is the deviation of the response of neuron k from its mean across all trials and choices. Denoting with Φ the cumulative Normal distribution function we obtain:

$$\begin{aligned} P(\delta r_k|D_{n_\mathrm{T}} < 0) &= 2\phi(r_k : f_k(s), C_{kk})\Phi\left(0 : \frac{\mathbf{C}\boldsymbol{\beta}_k}{C_{kk}}\delta r_k; \boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta} - \frac{\mathbf{C}\boldsymbol{\beta}^2}{C_{kk}}\right) \\ &= \frac{2}{\sqrt{C_{kk}}}\phi\left(\frac{\delta r_k}{\sqrt{C_{kk}}} : 0, 1\right)\Phi\left(-\frac{\mathbf{C}\boldsymbol{\beta}_k}{\sqrt{C_{kk}\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta} - \mathbf{C}\boldsymbol{\beta}^2}}\frac{\delta r_k}{\sqrt{C_{kk}}} : 0, 1\right) \end{aligned} \tag{10}$$

where

$$(\mathbf{C}\boldsymbol{\beta})_k \equiv \sum_{j=1}^{n} \beta_j C_{kj}$$
$$\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta} \equiv \sum_{j=1}^{n}\sum_{l=1}^{n} \beta_j \beta_l C_{jl}$$

are as verbally described in the main text. Eq. (10) represents a skew-normal distribution (generally defined as $P(x) = 2\phi(x)\Phi(\alpha x)$ where α is a scalar determining the shape) – see Figure 2A for an example.

Having derived the general choice-triggered response distribution (eq. (10)), we can now compute the choice probability according to the following formula ([3]):

$$\mathrm{CP}_k = \int_{-\infty}^{\infty} \mathrm{d}\delta r_k \, P(\delta r_k | D > 0) \int_{-\infty}^{\delta r_k} \mathrm{d}\delta r'_k \, P(\delta r'_k | D < 0). \tag{11}$$

In the following we sketch the solution restricting ourselves to the major steps. Defining $\alpha := \frac{\mathbf{C}\boldsymbol{\beta}_k}{\sqrt{C_{kk}\boldsymbol{\beta}^\top \mathbf{C}\boldsymbol{\beta} - \mathbf{C}\boldsymbol{\beta}^2}}$ it follows from eqs (10) and (11)

$$\begin{aligned}\mathrm{CP}_k &= 4\int_{-\infty}^{\infty} \mathrm{d}x \, \phi(x)\Phi(\alpha x) \int_{-\infty}^{x} \mathrm{d}y \, \phi(y)[1 - \Phi(\alpha y)] \\ &= 4\left[\int_{-\infty}^{\infty} \mathrm{d}x \, \phi(x)\Phi(x)\Phi(\alpha x) - \int_{-\infty}^{\infty} \mathrm{d}x \, \phi(x)\Phi(\alpha x) \int_{-\infty}^{x} \mathrm{d}y \, \phi(y)\Phi(\alpha y)\right]\end{aligned}$$

where zero mean and unit variance have been omitted from ϕ and Φ for brevity. Partially integrating both terms we obtain

$$\mathrm{CP}_k = \frac{3}{2} - 2\alpha \int_{-\infty}^{\infty} \mathrm{d}x \, \phi(\alpha x)\Phi(x)^2. \tag{12}$$

We perform the integral on the right $F(\alpha) := \int_{-\infty}^{\infty} \mathrm{d}x \, \phi(\alpha x)\Phi(x)^2$ by differentiation and integrate to find:

$$\begin{aligned}\frac{\mathrm{d}F(\alpha)}{\mathrm{d}\alpha} &= -\alpha \int_{-\infty}^{\infty} \mathrm{d}x \, x^2 \phi(\alpha x)\Phi(x)^2 \\ &= -\frac{1}{\alpha}F(\alpha) - \frac{2}{\alpha}\int_{-\infty}^{\infty} \mathrm{d}x \, x\phi(\alpha x)\phi(x)\Phi(x) \\ &= -\frac{1}{\alpha}F(\alpha) - \frac{1}{\pi}\frac{1}{\alpha(1+\alpha^2)\sqrt{2+\alpha^2}}\end{aligned} \tag{13}$$

The homogenous part of this differential equation implies $F(\alpha) \propto 1/\alpha$ leading to the ansatz $F(\alpha) = g(\alpha)/\alpha$. Substituting this into eq. (13) and integrating yields

$$g(\alpha) = -\frac{1}{\pi}\arctan\left(\frac{\alpha}{\sqrt{\alpha^2+2}}\right) + c$$

where c is an integration constant. Substituting $g(\alpha)$ back into F, and F into eq (12), and choosing c appropriately, we arrive at

$$\mathrm{CP}_k = \frac{1}{2} + \frac{2}{\pi}\arctan\frac{\alpha}{\sqrt{\alpha^2+2}}$$

which, after substituting in α, yields our central result in eq (1) in the main text.

4.2 Details for reconstruction of weights

In order to derive eq (4) from eq (2), we assume, without loss of generality, that $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} = 1$. This is possible since $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} = \sum_{k,l=1}^{n} \beta_k \beta_l C_{kl}$ scales with the square of the overall scale of $\boldsymbol{\beta}$ so that it can always be chosen such that $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} = 1$. Such a scaling implies no loss of generality since the overall scale of the weights is irrelevant for the behavior of the system: multiplying all weights by the same factor changes neither neuronal responses nor decisions. On the other hand, not every set of β_k obtained from eq (4) is guaranteed to obey the condition $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} = 1$. Given $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} = 1$, eq 2 becomes a linear equation in $\boldsymbol{\beta}$ that can be inverted to yield eq (4). This inversion is always possible since $\mathbf{C}$, being a covariance matrix, is invertible. Linearly scaling from the CP_k to γ_k we obtain:

$$\gamma_k = \frac{\pi}{\sqrt{2}} \sqrt{C_{kk}} (\mathrm{CP}_k - 1/2) \quad \text{and hence}$$

$$\boldsymbol{\gamma} = \mathbf{C} \frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}}} \quad \text{or}$$

$$\frac{\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}}} = \mathbf{C^{-1}} \boldsymbol{\gamma}$$

For a valid $\boldsymbol{\beta}$ to exist, $\boldsymbol{\gamma}^\top \mathbf{C^{-1}} \boldsymbol{\gamma} = 1$ needs to hold. If we had actually measured all the CPs, and all the pairwise correlations C_{ij}, and $\boldsymbol{\gamma}^\top \mathbf{C^{-1}} \boldsymbol{\gamma}$ were not 1, then this would imply that our model is wrong, or not incorporating some essential aspects of reality. However, if we have observed only a small subset of neurons in the population and hence measured only a small set of β_k and C_{jk}, then the first candidate for the mismatch may be the way we extrapolated from the few measured neurons to the entire population. In particular, variability in the underlying weights and/or noise correlations will lead to variability in the CPs. If this variability is not accounted for, but averaged away, then this will bias $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}$ to values larger or smaller than 1. For instance, adding variability to the β will generally increase $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta}$. This means that an extrapolation that averages away unobserved variability will lead to $\boldsymbol{\beta}^\top \mathbf{C} \boldsymbol{\beta} < 1$ and hence imply CPs that are larger than those that are actually observed. Fortunately, such a bias will only affect the magnitude of the implied CPs, and would not affect the structure of the read-out weights, i.e. the decoding strategy.

In order to reconstruct the weights one needs to make an assumption about the number of sensory neurons n, which typically would be unknown. However, we find that given the data-based correlation structure we use, CPs asymptote as the number of neurons increases. This happens as n approaches several hundred (see Fig. 2C for an illustration of the saturation based on different average correlation values). This means that for sufficiently large n, the reconstruction equation yields virtually identical weight profiles, and the actual number of neurons is irrelevant for reconstructing the weight profile as long as its large enough. In Suppl Figure 7 we illustrate the effect of using a range of n for the reconstruction for the case of our heterogenous population. We find that the corresponding weight profiles are indistinguishable away from the decision boundaries, and very similar close to them.

Instead of basing the inversion on the first-order approximation eq (2), it can also be based on the exact equation for the CP, eq (1), to yield:

$$\beta_k = \sum_{l=1}^{n} \{C^{-1}\}_{kl} \sqrt{C_{ll}} \sqrt{\frac{2}{1 + (\tan \eta)^2}} \quad \text{where} \quad \eta = \frac{\pi}{2} \left(\mathrm{CP}_l - \frac{1}{2} \right) \tag{14}$$

However, the improvement in accuracy is small for realistic values of CP (see also Figure 2B).

4.3 Simulation details

For all simulations, the optimal weight profile was constructed from the correlation matrix and the difference in population response for the two stimuli that were to be discriminated ([2]):

$$\boldsymbol{\beta} = C^{-1}\big[\mathbf{r}(s_1) - \mathbf{r}(s_2)\big]$$

where $\mathbf{r}$ is the population response to stimuli s_1 and s_2, respectively. For Figure 4 we assumed a population of identical direction-selective neurons only differing in their preferred direction for the random-dot task. We assumed van-Mises tuning curves with width $\kappa = 3$. In the case of zero signal (zero motion coherence) in the stimulus, the one considered by us, all neurons are assumed to have the same mean response and same variability.

For our realistic simulation shown in Figure 5 we used a heterogenous population of van-Mises tuning curves whose amplitude was drawn from a Poisson distribution with mean 20. Furthermore, we assumed response variability to be proportional to the mean response. The correlation matrix was based on linear fits (separate in each quadrant) to the data reported by [9] (see Fig. 4A) with noise with variance 0.01 added independently to each entry (while keeping C symmetrical and positive definite). For the average model, the simulated data from the subset of 64 neuron pairs was fit by linear functions to obtain the average correlation matrix, and by polynomials of degree 8 to obtain the average CP profile.

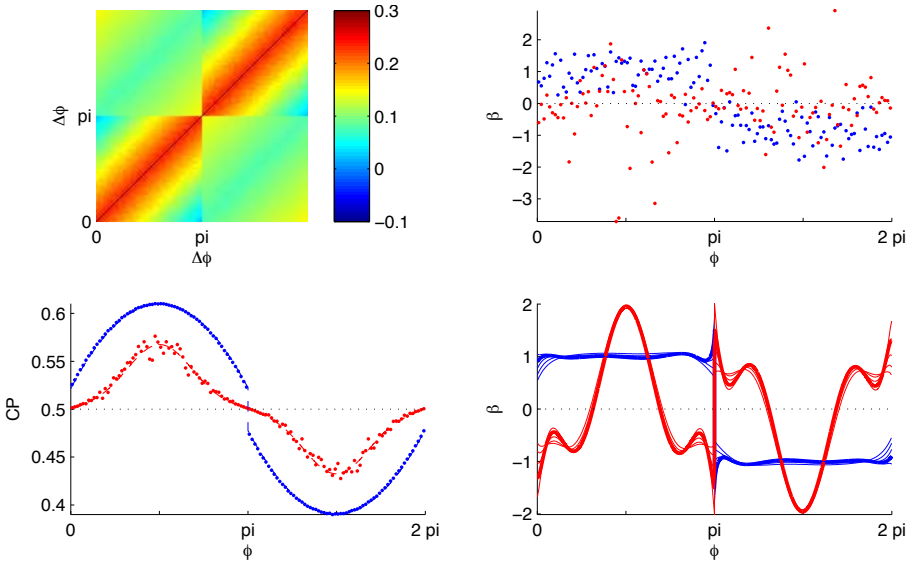

Figure 7: **Reconstruction of weights from limited data in the case of heterogenous weights and noisy correlation structure from CPs and correlations with heterogenous weights and limited data. A:** Noisy correlation structure around means as in Figure 4. **B:** Weights β chosen randomly around a uniform mean (blue:) and linearly-optimally (red:) based on the noisy correlation structure in A. **C:** Implied CPs associated with correlation structure in A and weights in B. Dashed red line is a parametric fit of the noisy CPs. **D:** Reconstructed weights assuming the mean of the correlation structure in A and the fits to the noisy CPs in C. Colors correspond to panel B. The thick line indicates the weights reconstructed assuming the correct number of neurons (here 4096). The thin lines correspond to reconstructions assuming a population size of 512, 1024, 2048, 4096, and 8192 neurons, respectively.

5 Supplementary Materials

5.1 Weight reconstruction

Figure 7 shows the reconstructed weight profile depending on the assumed number of neurons in the population n. We find that the reconstructed weights profile is largely independent of the assumed population size. Other than in the number of neurons (4096 instead of 8192), this example is identical to the one used in the main text (Figure 5).

5.2 Comparison of quadrant-wise constant with realistic correlations

Figure 8 illustrates why the reconstruction of read-out weights is virtually impossible in large neuronal populations while the same is not true for a more realistic correlation structure (shown in Fig 8, equivalent to the one used in the main text part of the paper). Fig 8 shows the two sets of weights that we compare: in blue are pool-wise constant weights that correspond to a simple average over the responses of all neurons within a pool. The profile in red corresponds to a read-out that only considers a small fraction of the neurons – those that are aligned with the task (i.e. the neurons that are typically the most informative for the task). In Fig 8 the implied CPs are shown: in blue and red based on the correlation structure in panel A, and in magenta and cyan based on the case of zero correlations. The case of zero correlations is equivalent to every quadrant-wise constant correlation

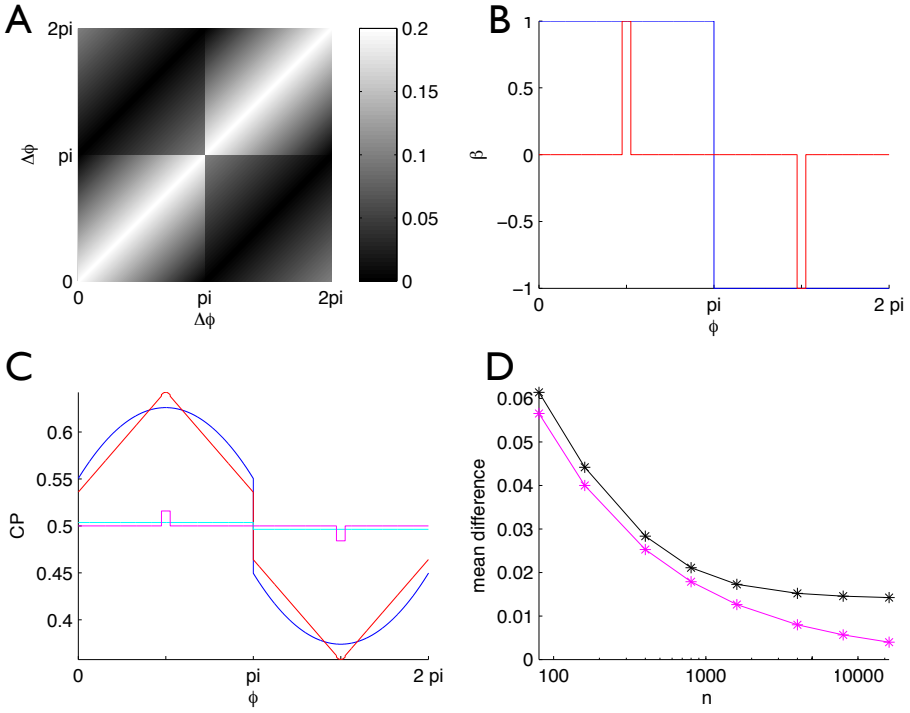

Figure 8: **Comparison of weight reconstruction from quadrant-wise constant and non-constant correlation structure. A:** Realistic correlation structure equivalent to that used in Figure 4. **B:** Two sets of weights β shown: simple average across all neurons within each pool (blue) and average over a small subset of neurons whose preferred direction is aligned with the task (red). **C:** Implied CPs for correlation structure shown in A (blue and red) and for zero correlations (cyan and magenta). **D:** Absolute difference in implied CPs for both weight profiles – averaged across all neurons.

structure in that the implied CPs for different values of correlations in the quadrants will differ by a constant offset. The magenta line in Fig 8 shows the mean absolute difference between the cyan and the magenta lines in panel C depending on the pool size n. For $n \to \infty$, the magenta line tends to 0 which means that in large populations the CPs implied by the blue weights in panel B and by the red weights in panel C become identical, making it impossible to infer which of the two profiles the brain is using. On the other hand, the black line in Fig 8, which shows the mean absolute difference between the red and the blue line in panel C, asymptotes at a positive value. This means that even in arbitrarily large populations, their will be an appreciable difference between the CP profiles allowing us to infer which profile is used by the brain. As this example indicates, the implied difference in CP may be quite small. However, since the CP profile that we need to reconstruct is only one-dimensional, a realistic amount of data (on the order of 100 neurons) is able to distinguish between the red and the blue line in Fig 8C – especially, when the symmetries of the task are taken into consideration reducing the problem to one in which the difference between the CPs only for the range $0 < \phi < \pi/2$ need to be considered.