Article

Neuron

# How Can Single Sensory Neurons Predict Behavior?

## Highlights

- Responses of single neurons correlate with heading percepts
- This can be explained by optimally decoding populations with limited information …
- … Or by suboptimally decoding populations with extensive information
- Electrophysiological data support the model with limited information

## Authors

Xaq Pitkow, Sheng Liu, Dora E. Angelaki, Gregory C. DeAngelis, Alexandre Pouget

## Correspondence

xaq@rice.edu

## In Brief

The activity of just one sensory neuron in the brain often accurately predicts what an animal will perceive in simple tests. Pitkow et al. provide a new theory of why this happens, and offer experimental data that support their theory.

Pitkow et al., 2015, Neuron *87*, 411–423
July 15, 2015 
http://dx.doi.org/10.1016/j.neuron.2015.06.033

CellPress

Neuron
Article

# How Can Single Sensory Neurons Predict Behavior?

Xaq Pitkow,[1,2,*] Sheng Liu,[1] Dora E. Angelaki,[1,2] Gregory C. DeAngelis,[3] and Alexandre Pouget[3,4]
[1]Department of Neuroscience, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA
[2]Department of Electrical and Computer Engineering, Rice University, 6100 Main MS-366, Houston, TX 77005, USA
[3]Department of Brain and Cognitive Sciences, University of Rochester, 358 Meliora Hall, Rochester, NY 14607, USA
[4]Department of Neuroscience, University de Genève, 1 Rue Michel-Servet, 1211 Geneva 4, Switzerland
*Correspondence: xaq@rice.edu
http://dx.doi.org/10.1016/j.neuron.2015.06.033

## SUMMARY

**Single sensory neurons can be surprisingly predictive of behavior in discrimination tasks. We propose this is possible because sensory information extracted from neural populations is severely restricted, either by near-optimal decoding of a population with information-limiting correlations or by suboptimal decoding that is blind to correlations. These have different consequences for choice correlations, the correlations between neural responses and behavioral choices. In the vestibular and cerebellar nuclei and the dorsal medial superior temporal area, we found that choice correlations during heading discrimination are consistent with near-optimal decoding of neuronal responses corrupted by information-limiting correlations. In the ventral intraparietal area, the choice correlations are also consistent with the presence of information-limiting correlations, but this area does not appear to influence behavior, although the choice correlations are particularly large. These findings demonstrate how choice correlations can be used to assess the efficiency of the downstream readout and detect the presence of information-limiting correlations.**

## INTRODUCTION

Individual sensory neurons in the brain are often predictive of animals' choices in simple perceptual decision-making tasks. It is said that these neurons have a significant choice probability. This remarkable fact has been demonstrated in numerous tasks and brain areas, including those dedicated to sensing visual motion (Britten et al., 1996), depth (Uka and DeAngelis, 2004; Nienborg and Cumming, 2007), and self-motion (Gu et al., 2008; Fetsch et al., 2012; Chen et al., 2013; Liu et al., 2013). Many of these cells have neural thresholds, which quantify sensitivity to stimulus variations, that are not much greater than psychophysical thresholds (Cohen and Newsome, 2009). It is therefore puzzling why pooling these signals does not predict sensitivity much greater than that exhibited by behavior. Perhaps the brain merely selects a small subset of neurons to inform its decisions (Tolhurst et al., 1983; Ghose and Harrison, 2009)—but then how could experiments so frequently encounter these extremely rare neurons that influence behavior? A proposed explanation for these puzzling observations was that response variability is correlated across neurons (Zohary et al., 1994): even with very weak correlated noise between pairs of neurons, the total information content of a neural population may saturate to a finite value as the number of neurons increases, such that optimally pooling more responses cannot improve behavioral sensitivity. Additionally, neurons are correlated not only with each other but also with the pooled signal that presumably drives the perceptual decision, which would generate high choice probabilities.

This solution (Zohary et al., 1994) was established for a very simplified model of neural responses, correlations, and decoding. Subsequent studies relaxed some of these simplifications and found consistent results for broad correlations in neural populations tuned to a one-dimensional stimulus (Sompolinsky et al., 2001). However, it was suggested that diversity in the amplitude and width of neural tuning curves would change the picture (Abbott and Dayan, 1999), and later calculations demonstrated that weak noise correlations do *not* limit information in heterogeneous neural populations: information continues to increase linearly with the number of neurons (Shamir and Sompolinsky, 2006; Ecker et al., 2011). We say that such a population has "extensive information." If correct, this would imply that correlated noise cannot explain the frequent occurrence of significant choice probabilities, for the following reason: in optimally decoded populations with extensive information, each neuron provides a tiny contribution, inversely proportional to the size of the neural pool, toward the perceptual decision. This prediction is at odds with observed choice probabilities and ratios of neural to psychophysical thresholds.

Perhaps the neural population contains vast amounts of information, but it is not all used in perception. There are many forms of such suboptimal decoding that misuse neural signals. We will show that suboptimal decoding could indeed explain both why behavioral thresholds are barely better than single neuron thresholds and why choice probabilities are so large and common.

A second explanation of these phenomena does not rely on suboptimal decoding but instead blames a subtle form of neural noise correlations (Moreno-Bote et al., 2014) that limit the information contained in a population code. These information-limiting noise correlations cause massive redundancy between neurons, which restricts behavioral thresholds to be not much

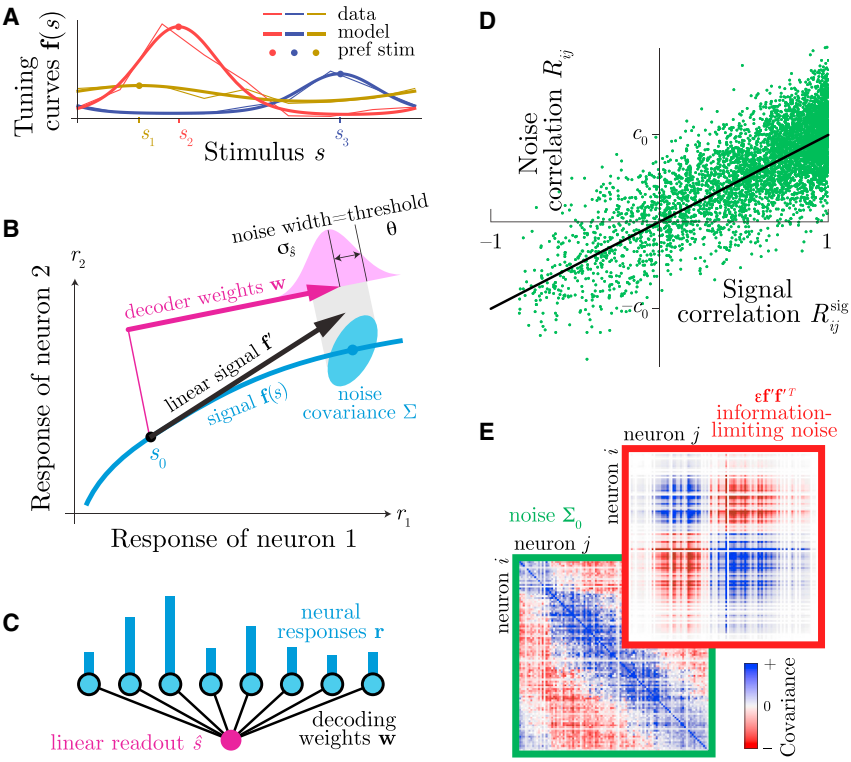


**Figure 1. Model for Neural Responses and Decoding**

(A) Three example tuning curves **f**(s) show the mean neural responses to a stimulus *s* (thin lines). These tuning curves are modeled by von Mises functions (thick curves) with parameters including the preferred stimulus $s_k$ (dots).

(B) As the stimulus *s* varies, the mean activity of all neurons traces out a curve **f**(s) (blue) through the *N*-dimensional space of neural responses. For fine local discriminations, the relevant signal direction lies along the tangent **f′** (black). Neural responses are decoded by projection onto a readout direction **w** (magenta). The amplitude of **w** is scaled to give an unbiased estimate $\hat{s}$ of the stimulus, so that a unit change in stimulus generates a unit change in the estimate. Trial-to-trial variability is expressed as multivariate Gaussian noise, with covariance Σ (ellipse). The projection of this noise along the decoding direction (pink Gaussian) has standard deviation $\sigma_{\hat{s}}$, which we define as the population threshold $\theta$. Although we illustrate responses for only two neurons here, these relationships generalize to high-dimensional response spaces.

(C) Linear decoding projects the neural responses, both noise and signal, onto a particular direction **w** to obtain an estimate $\hat{s}$ of the stimulus.

(D) Noise correlation coefficients $R_{ij}$ between distinct neurons *i* and *j* are modeled as being proportional on average to the signal correlations $R_{ij}^{sig}$, with proportionality $c_0$. This means that neurons with similar tuning tend to have more correlated fluctuations.

(E) We distinguish two components to the noise covariance Σ: information-limiting correlations that lie along the signal direction **f′** and thus have covariance $\varepsilon \mathbf{f}'\mathbf{f}'^T$ (front, matrix boxed in red), and the remaining noise with covariance $\Sigma_0$ (back, matrix boxed in green). The two forms of noise have distinct structures that are apparent in the covariance matrices. The striations in both matrices reflect the heterogeneous tuning curve amplitudes.

better than individual neural thresholds. We show that this explanation also predicts many neurons with high choice probabilities.

Thus both suboptimal decoding and information-limiting noise correlations could explain these two puzzling phenomena. Which is the correct explanation? We derive quantitative consequences of each hypothesis in order to understand the nature of neural population codes. We test these consequences in various brain areas that are responsive to vestibular signals and are activated during a heading discrimination task. We find that most of the data are more consistent with near-optimal decoding of neural responses with information-limiting noise correlations.

## RESULTS

During a vestibular heading discrimination task, animals were presented with a movement stimulus *s*: specifically, translation by a motorized platform within the horizontal plane (see Figure S1 available online). S/he must use the responses of neurons tuned to the vestibular stimulus *s* in order to estimate a direction of motion $\hat{s}$, and to discriminate whether that heading is slightly leftward or rightward of some reference heading $s_0$, which was straight forward in our task. A heading estimate is generated by pooling responses **r** of neurons in some way. In the brain areas from which we recorded, neurons are tuned to heading, with average responses, $f_k(s)$, that are characterized by a few parameters for each neuron *k*, including its preferred heading $s_k$ (Figure 1A; Experimental Procedures).

The ability of a single neuron to discriminate between similar headings is generally greatest when the tuning curve has a steep slope $f_k' = df_k/ds$ near the reference stimulus, such that the mean response changes substantially with small variations in heading. However, neural responses vary from trial to trial even when the stimulus is the same. Consequently, discriminability decreases for larger response variance $\sigma_k^2$. We define a discrimination threshold $\theta_k$ for each neuron as the signal change required to exceed one standard deviation of noise, $\theta_k = \sigma_k / f_k'$.

An animal may estimate the stimulus more reliably than single neurons by pooling signals appropriately across a population of neurons. To understand the information content of the population, it is helpful to visualize how the vector of mean responses traces out a curve, **f**(s), in the *N*-dimensional neural response space as a function of the stimulus *s* (Figure 1B). For the fine discrimination tasks we examine here, the tested stimulus range around the reference is sufficiently narrow that the mean neural responses depend nearly linearly on the stimulus (Gu et al., 2008), thus lying close to the tangent vector **f′**. Over such a narrow stimulus range, evidence from other systems suggests that most of the information can be extracted near-optimally by a linear decoder (Ma et al., 2006; Graf et al., 2011; Berens et al., 2012). We therefore model the animal's estimate $\hat{s}$ as a linear

weighting of all neural responses **r** in a population (Figure 1C), according to

$$\hat{s} = \mathbf{w}^T(\mathbf{r} - \mathbf{f}(s_0)) + s_0. \quad \text{(Equation 1)}$$

This linear decoding can be viewed as a projection of the *N*-dimensional responses onto a single dimension defined by the weight vector **w** (Figure 1B).

Noise in the neural population generates a cloud of possible responses around the mean response. The covariance $\Sigma$ of this high-dimensional response variability can be visualized as an ellipse centered on the mean response (Figure 1B). Among the many dimensions of this noise, only noise along the decoding direction **w** generates variability in the estimate $\hat{s}$; the remaining variability in the orthogonal directions has no effect on the estimate. We can define a discrimination threshold $\theta$ for the decoded estimate just like we do for single neurons, as the signal change needed to exceed one standard deviation of noise in the estimate, $\sigma_{\hat{s}}$ (Figure 1B; Supplemental Information).

Animals can be trained to give largely unbiased reports in this task (Experimental Procedures), which means that on average the animal has an accurate estimate. Although behavior should benefit from combining information from many neurons, behavioral discrimination thresholds are not substantially better than thresholds for the best single neurons (Gu et al., 2008; Cohen and Newsome, 2009). Is this because the responses of these neurons are correlated, such that they don't provide independent information? Or is the brain using their information poorly? We can distinguish these possibilities by looking across trials at the relationship between neural responses and perceptual reports.

This relationship is typically quantified by the "choice probability," which is the probability that a neural response associated with one behavioral choice is greater than a neural response associated with the other possible choice (Britten et al., 1992). As derived by Haefner et al. (2013), choice probability is influenced both by the neural correlations (via the noise covariance matrix $\Sigma$) and by the decoding weights **w** (Experimental Procedures). Here, we measure the relationship between neuron and behavior by computing "choice correlation," the Pearson correlation coefficient $C_k$ between the response $r_k$ of neuron $k$ and the estimated stimulus $\hat{s}$, $C_k = \mathrm{Corr}(\hat{s}, r_k)$ (we think of $\hat{s}$ as a continuous "choice"). This quantity has a simple, nearly affine relationship to choice probability (see Haefner et al., 2013; Experimental Procedures), but it is conceptually simpler and mathematically more convenient. Below we extend the important results of (Haefner et al., 2013) to analyze choice correlation under conditions of information-limiting correlations, suboptimal decoding, or both.

## Consequences of Suboptimal Decoding on Choice Correlations

One possible account of high choice correlations is suboptimal decoding. Perhaps the information encoded by areas representing heading is indeed extensive, growing in proportion to the number of neurons, but the downstream neural circuits fail to extract this information efficiently. This could predict psychophysical thresholds that are not vastly better than the typical single-neuron discrimination threshold. Would this mechanism also produce high choice correlations?

We examined a family of suboptimal decoders that are blind to the patterns of correlated fluctuations present in the neural population. These are decoders that are based solely on the signal strength in individual neurons, and do not take into account the correlations between neurons. For instance, one commonly used decoder of this type is known as a "factorial decoder." This decoder assumes that all neurons are independent, so the probability of a population response factorizes over neurons. It thus simply weights each response according to the individual neural sensitivities. This is a maximum likelihood decoder only when the neurons are truly independent (Földiák, 1993; Sanger, 1996; Liu et al., 2013), an assumption which is generally violated in the brain. Nonetheless, in some circumstances, the factorial decoder is nearly optimal, despite unfaithfully neglecting correlations (Wu et al., 2001). In other circumstances, such as in the presence of tuning curve diversity and smoothly varying noise correlation coefficients (Shamir and Sompolinsky, 2006; Ecker et al., 2011), this correlation-blind decoder is extremely suboptimal and throws away almost all of the information. More specifically, while the information in the population may be extensive, the information recovered by the factorized decoder is not. Instead, the information saturates to a finite value, producing a high behavioral threshold (Supplemental Information; also see Shamir and Sompolinsky, 2006, for the similar population-vector decoder). In this case, instead of cancelling the large noise fluctuations shared by many neurons, the suboptimal, correlation-blind decoder would preserve them and behavior would be largely driven by that irrelevant noise. Many neurons would be strongly correlated with behavior because they share the strong correlated noise that drives it (Figure S2). This could explain the prevalence of high choice correlations.

To be more quantitative, we model the noise correlation coefficient matrix $R$ in accordance with recent experimental studies (Cohen and Kohn, 2011; Liu et al., 2013). Specifically, we assume that noise correlations are proportional to signal correlations on average, with proportionality constant $c_0$, but with substantial heterogeneity around this trend (Figure 1D). Such noise does not limit the information content of a heterogeneously tuned population (Shamir and Sompolinsky, 2006; Ecker et al., 2011). Yet the resultant extensive information can be extracted only if the noise correlations are cancelled by appropriate weighting of the neurons, which is not the case for suboptimal, correlation-blind decoders. In the Supplemental Information we show that the choice correlations for a correlation-blind decoder are well approximated by

$$C_k^{\mathrm{cb}} \approx \sqrt{c_0}|\sin s_k|, \quad \text{(Equation 2)}$$

where $s_k$ is the preferred stimulus of the neuron relative to the reference stimulus $s_0 = 0$. This relationship reflects not the individual neural sensitivities but rather the structure of the broad noise correlations that are not removed by the decoder. This remains true even when there is large heterogeneity in tuning curves, noise correlations, or correlation-blind decoder structure (Figures S3 and S4). When correlated noise is not removed, shared fluctuations dominate the animal's resulting choice, and thus determine the choice correlation (Supplemental Information).

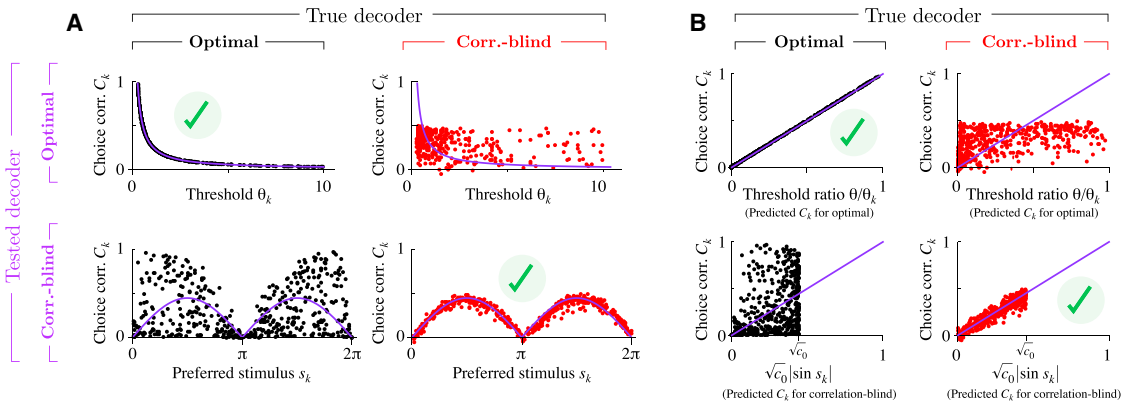


**Figure 2. Optimal and Suboptimal, Correlation-Blind Decoding of Simulated Heterogeneous Neural Populations Can Be Readily Distinguished by Their Predicted Patterns of Choice Correlations**

Top and bottom rows show the choice probabilities of optimally and suboptimally decoded neural populations, plotted against predictions of correlation-blind (Equation 2) and optimal (Equation 3) decoders. In (A), simulated choice correlations are plotted against neural threshold $\theta_k$ and preferred stimulus $s_k$, quantities with which they should have nonlinear relationships. In (B), the same simulated choice correlations are plotted against $\theta/\theta_k$ and $\sqrt{c_0}|\sin s_k|$, where a linear relationship should hold for optimal and suboptimal decoding, respectively. Green checks indicate good agreement when the prediction matches the true decoder. Note that the bottom-left panel of (B) provides a direct comparison of the two predictions in the case of hetereogeneous tuning curves, since the horizontal axis indicates choice correlations for the correlation-blind decoder while the vertical axis gives choice correlations for the optimal decoder (which is the true decoder for that simulation).

Suboptimal decoding could also account for a wide range of empirical observations regarding the average strength of choice correlations. According to Equation 2, choice correlations depend on the overall correlation scale given by the proportionality constant $c_0$ between signal and noise correlations, which is typically in the range of 0.1–0.5 (Chen et al., 2013; Liu et al., 2013). Thus, a steeper slope in the relationship between noise and signal correlations would lead to greater choice correlations in this regime, and some empirical studies have reported results that are consistent with this prediction of suboptimal decoding (Chen et al., 2013; Liu et al., 2013). In this situation, choice correlations are readily distinguishable from chance and do not decrease with the number of neurons or the overall information content in the population.

## Consequences of Optimal Decoding on Choice Correlations

Next we consider what happens when behavior is optimal given the neural response properties. For a fine discrimination task, the optimal decoder is $\mathbf{w} \propto \Sigma^{-1}\mathbf{f}'$ (Salinas and Abbott, 1994), where $\Sigma$ is the true covariance matrix of neural population responses. For this decoder, extending the result of Haefner et al. (2013), we have found that the choice correlations take the remarkably simple form of a ratio of discrimination thresholds (Experimental Procedures),

$$C_k^{\text{opt}} = \frac{\theta}{\theta_k}, \qquad \text{(Equation 3)}$$

where $\theta_k$ is the discrimination threshold of neuron $k$ and $\theta$ is the discrimination threshold of the optimal decoder of the population. A consequence of this relationship is that more informative neurons—those with lower thresholds—have a greater correlation with behavior, as often observed experimentally (Britten et al., 1996; Purushothaman and Bradley, 2005; Gu et al., 2008).

We verified that Equations 2 and 3 can be used to successfully identify the decoding strategy by simulating heterogeneous populations and a decoding strategy that is chosen to be either optimal or suboptimal. Figure 2 plots the simulated choice correlations, first against the relevant neural properties (Figure 2A) and second against the choice correlations predicted from those properties (Figure 2B). The choice correlations for optimal decoding are perfectly fit by the prediction for optimal decoding (Equation 3), and not by the prediction for suboptimal decoding (Equation 2); the reverse holds approximately for choice correlations generated by suboptimal decoding. The bottom-left panel of Figure 2B also provides a direct comparison of the two predictions, since the horizontal axis indicates predicted choice correlations for the suboptimal decoder while the vertical axis gives choice correlations for the optimal decoder (which is the true decoder for that simulation). The predictions are only weakly correlated ($r = 0.26$, $p < 0.01$ Pearson correlation test), and quite clearly distinguishable.

## Consequences of Information-Limiting Noise

Interestingly, noise correlations do not appear explicitly in Equation 3, because the optimal decoder has removed them to the extent possible. Nonetheless, their structure has enormous importance for both behavior and choice correlations because they determine the population threshold, $\theta$. Quantitative models of population codes have typically measured this by computing the Fisher information $J$, whose inverse provides a lower bound on the variance of an unbiased estimator. Since we defined the discrimination threshold $\theta$ as the standard deviation of our estimator, therefore $\theta \geq J^{-1/2}$; consequently higher information permits a lower discrimination threshold.

According to current models of population codes (Shamir and Sompolinsky, 2006; Ecker et al., 2011), when noise correlations are greater for neurons with similar tuning curves

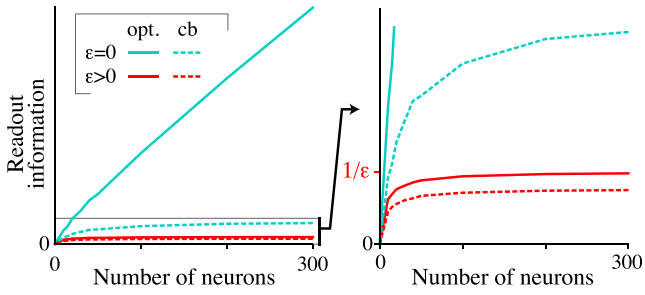


**Figure 3. Effects of Information-Limiting Noise and Correlation-Blind Decoding on the Information Extracted by a Linear Estimator (Equation 5)**

Optimal decoding of a neural population without differential correlations yields extensive information that increases without bound as more neurons are added (solid cyan). Suboptimal, correlation-blind (cb) decoding of the same population is only able to extract a limited amount of information even with an infinite number of neurons (dashed cyan). In the presence of differential correlations with variance $\varepsilon$ (Equation 4), however, information saturates to a finite value of $1/\varepsilon$ even for large populations that are decoded optimally (solid red). The suboptimal (cb) decoder does not perform much worse than the optimal decoder in this case, extracting information (dashed red) which is not much smaller than $1/\varepsilon$. The second panel is a vertically expanded view of the first panel.

(as often observed empirically, see Cohen and Kohn, 2011), the amount of Fisher information contained in a population is extensive, growing with the number of neurons in the population (Figure 3, solid cyan line). This is because when tuning curves are heterogeneous, the noise has a different structure from the signal, and thus noise and signal can be distinguished. For a population of 1,000 neurons, optimal decoding could produce a behavioral threshold roughly 30-fold ($\sqrt{1{,}000}$) smaller than the threshold of a typical neuron that is sensitive to the stimulus, and this is much smaller than generally observed (Cohen and Newsome, 2009). According to Equation 3, the choice correlations would then be correspondingly tiny, and most likely not significantly different from zero given typical measurement uncertainty. This argument rules out models in which the brain has extensive information and decodes it optimally.

However, models involving diverse neural populations might radically overestimate the amount of information in a population because they do not account for how the sensory periphery limits the information provided to a large cortical network. If information is not available at the input, then it cannot be created later by adding more neurons. Those extra cortical neurons may appear to add more signal strength, but they also inherit noise with the same structure as the signal. As a consequence, that noise is correlated in a very particular way. For the fine discrimination task that we examined, the signal is encoded in the change in the mean rate, $\mathbf{f}'$, so the relevant information-limiting noise covariance is proportional to $\mathbf{f}'\mathbf{f}'^T$, which we have described elsewhere as "differential correlations" (Moreno-Bote et al., 2014). The total noise then has a covariance that can be modeled as

$$\Sigma = \Sigma_0 + \varepsilon \mathbf{f}'\mathbf{f}'^T \qquad \text{(Equation 4)}$$

where $\Sigma_0$ is a covariance matrix of noise that does not limit information and $\varepsilon$ represents the variance of the information-limiting noise (Figure 1E).

No matter how the population is decoded, information-limiting noise prevents the variance of any unbiased linear estimator from falling below $\varepsilon$. The sum of noise covariances in Equation 4 manifests as a sum of noise variances in the decoded estimate $\sigma^2_{\hat{s}} = \sigma^2_{0\hat{s}} + \varepsilon$, where $\sigma^2_{0\hat{s}}$ is the variance that would have been obtained without information-limiting noise (Supplemental Information). Since $\sigma^2_{0\hat{s}} \geq 0$, therefore $\sigma^2_{\hat{s}} \geq \varepsilon$. This same relationship can be expressed in terms of (linear) Fisher information terms, $J = 1/\sigma^2_{\hat{s}}$ and $J_0 = 1/\sigma^2_{0\hat{s}}$, yielding

$$J = \frac{1}{1/J_0 + \varepsilon} \qquad \text{(Equation 5)}$$

Thus information-limiting noise prevents the decoded information from exceeding $1/\varepsilon$ (Figure 3, solid red line).

While no decoder can exceed this limit, it is of course possible to do worse. The quality of a decoder is determined by how efficiently it eliminates the noise that is not information limiting. If $\varepsilon$ is small relative to $\sigma^2_{0\hat{s}}$, then performance will be greatly enhanced by learning decoding weights that eliminate as much noise as possible. However, if $\varepsilon >> \sigma^2_{0\hat{s}}$, there is relatively little to be gained by fine-tuning the decoding weights. This is why, in the absence of information-limiting correlations ($\varepsilon = 0$), a suboptimal correlation-blind decoder loses the vast majority of available information (Figure 3, dashed cyan curve; Supplemental Information), yet in the presence of information-limiting noise the same decoder loses only a modest fraction of the information that is available (Figure 3, dashed red curve). This demonstrates that large population codes with limited information are redundant and exhibit considerable robustness to suboptimal decoding: a broad range of decoders may all produce similar near-optimal performance.

Despite the importance of information-limiting correlations, they are difficult to estimate directly, requiring large simultaneous recordings with many trials. There are two reasons for this. First, the information-limiting component can be very small yet have enormous effects on population information. Second, the fine details of the correlation patterns matter greatly. Extrapolating the full noise correlations from a sparse subset of pairwise correlation measurements is extremely difficult, and mistakes can radically change the estimated information content of a neural population (Moreno-Bote et al., 2014). Fortunately, we show below that there are indirect consequences of this information-limiting noise that are observable with only single-neuron measurements: choice correlations should be observably large and should obey the predictions of optimal decoding (Equation 3).

Choice correlations are influenced by both suboptimal decoding and information-limiting noise correlations according to the weighted sum

$$C_k \approx \alpha C_k^{\text{opt}} + \sqrt{1-\alpha} C_k^{\text{sub}} \qquad \text{(Equation 6)}$$

where $\alpha = \varepsilon J$ is the fraction of the uncertainty in the stimulus $\hat{s}$ caused by the information-limiting noise (Supplemental

Information). We emphasize that the decoder producing Equation 6 is not somehow both optimal and suboptimal. Instead, it is suboptimal for all $\alpha < 1$, but its choice correlations are a weighted sum of $C_k^{opt}$ and $C_k^{sub}$, the choice correlations for purely noise-limited optimal decoding or purely suboptimal decoding (of any type, not just correlation blind) of a population with extensive information ($\varepsilon = 0$), respectively. As long as the behavioral threshold is primarily limited by noise and not by losses from suboptimal decoding, then $\alpha$ will be near 1. There will then be an inverse relationship between a neuron's threshold and its influence on behavior, regardless of neural correlations or the form of the decoder weights.

## Suboptimal Decoders Can Produce Choice Correlations that Are Scaled Versions of Optimal Choice Correlations

Surprisingly, one can identify circumstances in which the choice correlations have the same pattern as for optimal decoding, that is, $C_k = \beta\, C_k^{opt}$, but with $\beta > 1$. This arises when behavior is driven by another, more informative, source of sensory signals besides the neural population under study. If responses of this more-sensitive population are correlated with the observed population, then the observed neurons will exhibit choice correlations. As detailed in the Supplemental Information, $\beta > 1$ occurs only for suboptimal decoding, and can be explained quantitatively if the observed population is mostly ignored while the other more-sensitive population is decoded efficiently.

Intuition for this result can be gained by examining results for a pair of idealized neurons, $x$ and $y$, and then generalizing to two large populations. Imagine that the behavior is determined solely by the activity of neuron $x$, given by $x = s + n$ for signal $s$ and noise $n$. Naturally, this decoded neuron will then be perfectly correlated with behavior. Since the neural threshold $\theta_x$ is the same as the behavioral threshold $\theta$, the choice correlation is accurately described by $C_x = 1 = \theta/\theta_x$. Now imagine that neuron $y$ carries the same stimulus-related signal and the same noise on every trial, except that the noise is multiplied by a factor of 2, $y = s + 2n$. Even though $y$ is not decoded, it is still perfectly correlated with behavior since it is perfectly correlated with $x$. Yet because its neural threshold is twice as large due to greater noise, $\theta_y = 2\theta_x$, its choice correlation will be twice as large as an optimally decoded neuron with the same threshold: $C_y = 1 = 2(\theta/\theta_y)$.

The situation for two larger populations is analogous, albeit with the relevant signal and noise being distributed among many neurons along the direction of $\mathbf{f}'$. When population $x$ is decoded near-optimally, its noise is correlated with behavior according to Equation 3. A second, undecoded population $y$ can have partially overlapping information, for instance inherited from overlapping subsets of upstream neurons (Figure 4A). This corresponds to partially correlated information-limiting noise. Figure 4B shows how this noise corresponds to trial-to-trial shifts in the hills of activity, with different variances in each population and a nontrivial covariance between them. When this covarying information-limiting noise is a larger fraction of the signal in population $y$ compared to population $x$ (i.e., its noise is larger in units of $\mathbf{f}'$), then choice correlations for neurons in population $y$ will have a larger overall scale, but the same pattern, as choice correlations for neurons in population $x$ (Figure 4C).

An important observation about this scenario is that choice correlations proportional but not equal to the optimal prediction (Equation 3) occur only with some degree of suboptimal decoding. This can be seen in Figure 4D, which shows the efficiency (color scale) of a decoder that ignores the less informative of the two populations, where efficiency is the ratio of information actually decoded to the information that could have been decoded. This efficiency depends on the relative information content in the two populations (horizontal axis) as well as on $\beta$ (vertical axis), which specifies how much choice correlations are amplified in the undecoded population relative to the prediction from optimal decoding (Equation 3). Whenever the efficiency is 1 (white area of Figure 4D), then $\beta = 1$ and the choice correlations match the optimal predictions. Conversely, whenever the choice correlations have $\beta \neq 1$, the efficiency is less than 1. The counterintuitive situation in which choice correlations are greater than expected ($\beta > 1$) likely arises when there are two correlated populations of neurons that carry task-related information and the population with greater noise variance is decoded suboptimally (e.g., ignored).

## Inferring Decoding Quality from Neural Data

To determine whether the information in a neural population is limited by noise or by suboptimal decoding, we can use linear regression to fit the measured choice correlations against those predicted from optimal and suboptimal decoders, while carefully accounting for uncertainties in each measurement (Experimental Procedures; Minka, 1999). We consider the two natural forms of suboptimality described above: correlation-blind decoding, and a decoder that ignores one population. These decoders generate choice correlations of the form $C_k = \beta C_k^{opt} + \gamma C_k^{cb}$ (Equation 6; Supplemental Information). We fit these coefficients separately to allow for both of these forms of suboptimality. The coefficient $\beta$ should reveal the fraction of behavioral variance caused by information-limiting noise in the recorded population (Supplemental Information).

To validate this approach, we simulated virtual neural populations and their virtual behavioral outputs, and used our method to try to recover the true decoder properties under realistic experimental conditions. These four model systems were (1) optimal decoding and information-limiting noise; (2) suboptimal, correlation-blind decoding of a population with extensive information; (3) suboptimal, correlation-blind decoding with information-limiting noise; and (4) two subpopulations with correlated information-limiting noise where only the more informative subpopulation is decoded while we recorded from the other subpopulation (Experimental Procedures). For (3), we set parameters such that $\alpha = 0.9$ in Equation 6, meaning that 90% of the variance of the stimulus estimate was due to information limiting correlations. We then simulated recordings from small subsets of the virtual neurons, including measurement error, and estimated choice correlations, neural thresholds, tuning curves, and their corresponding uncertainties. The thresholds and tuning data were used to predict choice correlations according to Equations 2 and 3 separately, and these predictions were combined through linear regression to find the coefficients, $\beta$ and $\gamma$, attached to the optimal and suboptimal choice correlations. Figure 5 shows these coefficients plotted separately (Figure 5A) and together

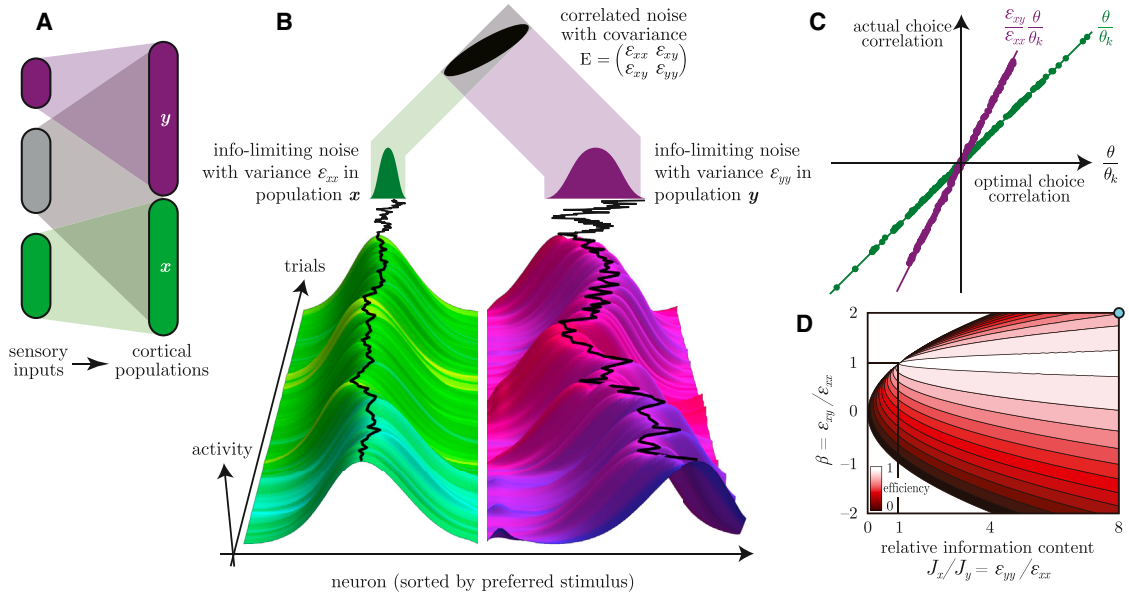


**Figure 4. Choice Correlations for Two Populations, *x* (green) and *y* (purple), with Correlated Information-Limiting Noise**

(A) Schematic of one way that the two large cortical populations can inherit this form of noise by receiving some shared sensory input (gray ellipse) in addition to their own private sensory inputs.

(B) Illustration of activity in both cortical populations as a consequence of correlated information-limiting noise. Information-limiting noise causes the neural activity in each population to fluctuate from trial to trial (greenish and purplish surfaces). As shown here, these fluctuations are visualized most readily for homogeneous neural populations with pure information-limiting noise—fluctuations that look exactly as if the stimulus itself had shifted (black curves). Over many trials, this variability has a distribution (shown above the neural activity) whose width is determined by $\varepsilon$ and inversely related to the information content. If the information-limiting noise is identical between the two populations, then the fluctuations will have the same extent, but if the populations have at least partially distinct sources of information, then the information-limiting noise in each will be partially correlated (black ellipse, Equation 7). This example has a variance of $\varepsilon_{xx} = 1$ in the green population, and a higher variance $\varepsilon_{yy} = 8$ in the purple population, with covariance $E \propto \left(\begin{smallmatrix} 1 & 2 \\ 2 & 8 \end{smallmatrix}\right)$. For these parameters, half of the variance in the purple population copies the fluctuations in the green population, but with twice the size.

(C) If only population *x* is decoded—and decoded optimally—then its choice correlations are given exactly by Equation 3 (green line). Choice correlations for the nondecoded population *y* (purple) are proportional to those optimal predictors, but with a proportionality $\varepsilon_{xy}/\varepsilon_{xx}$ (Supplemental Information) that can be greater or less than 1. This analytical result (solid lines) is supported by simulations of neurons in two populations (filled symbols; Supplemental Information).

(D) Contour map of the decoder efficiency, which is the fraction of the total information in the population that is actually extracted by the decoder. For a suboptimal decoder that extracts essentially all of the information $J_x$ in population *x* but uses none of the responses from population *y*, the efficiency can range from highly suboptimal ($J_x/J \approx 0$) to nearly optimal ($J_x/J \approx 1$) depending on the covariance structure *E*. For the covariance depicted in (B) and (C), this decoder achieves an efficiency of 80% (cyan dot).

(Figure 5B). In all four model systems, this method was able to recover the true values of both coefficients within the simulated experimental uncertainty. Figure 5C shows scatterplots of the predicted versus measured choice correlations, for three different predictors: optimal decoding (the psychophysical/neural threshold ratio, Equation 3), suboptimal correlation-blind decoding (Equation 2), and the linear combination of the two that uses best-fit weights $\beta$ and $\gamma$. As expected, the best predictors are the ones that match the actual decoding model (Figures 5C and 5D). This demonstrates that our procedure successfully identifies whether neural activity is decoded optimally or suboptimally and can recover the fraction of behavioral variance caused by information-limiting noise.

## Neural Response Properties during Vestibular Heading Discrimination Are Consistent with Near-Optimal Decoding

We now apply these theoretical insights to experimental data. In particular, we examine response properties and choice correlations of neurons recorded in multiple cortical and subcortical brain areas during a vestibular heading discrimination task (Gu et al., 2007, 2008). Monkeys were translated forward and slightly leftward or rightward on a motorized platform, and were trained to report their perceived heading, left or right relative to straight forward, by making an eye movement to one of two targets. During this task, single neurons were recorded from the vestibular and cerebellar nuclei (VN/CN), the dorsal medial superior temporal (MSTd) area, and the ventral intraparietal (VIP) area. For each brain area, neural responses and behavioral choices were analyzed to extract choice correlations, neural thresholds, and behavioral thresholds (Experimental Procedures). In a separate stimulus condition, monkeys were translated along eight equally spaced headings in the horizontal plane during a visual fixation task, and neural responses were used to generate heading tuning curves and extract heading preferences. Measurements of pairwise noise correlations were also extracted from these data, as described previously (Gu et al., 2011; Chen et al., 2013; Liu et al., 2013). Together,

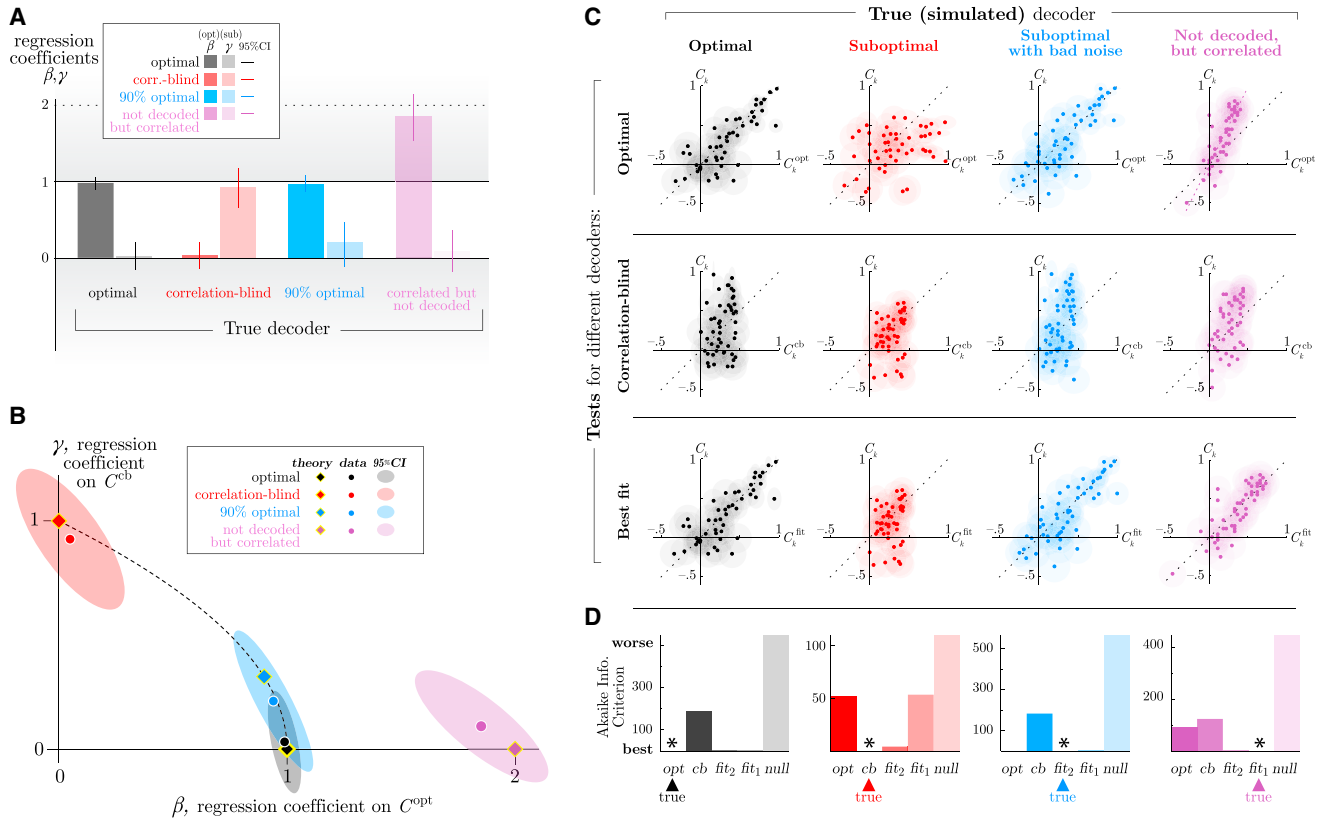


**Figure 5. Inferring Decoding Quality from Simulations with Realistically Limited Amounts of Data**

We use numbers of neurons (50) and trials (30 repetitions of each stimulus) that are comparable to those obtained in our experiments. Choice correlations are fit to predictions from optimal (Equation 3) and correlation-blind (Equation 2) decoding, and the resultant regression coefficients ($\beta$, $\gamma$) are plotted separately as a pair of bars in (A) and jointly as a point in (B), indicating how much of variance in behavior can be explained by optimal or correlation-blind decoding. In particular, $\beta = 1$ when decoding is optimal, with its quality limited by noise rather than by suboptimal decoding. Four example populations and decoders are shown: optimal (black), correlation blind (red), correlation blind with information-limiting noise (blue), and an undecoded population that is correlated with an optimally decoded population. For each example, the recovered coefficients (filled circles) for the decoded populations fall within 95% confidence intervals (shaded ellipses in B) of the parameters expected for the true decoder (filled diamonds). Thus, the optimal decoder is revealed as optimal (black), and the correlation-blind decoder is revealed as clearly suboptimal for a population with extensive information (red) but nearly optimal for a population with limited information (blue). The parametric curve $(\alpha, \sqrt{1-\alpha})$ (dashed) shows coefficients expected for models with a fraction $\alpha$ of total uncertainty caused by information-limiting noise. For the undecoded population, coefficients fall near the theoretical values (2,0) used in the simulation (purple), correctly implying that the population cannot account for the quality of the decoded output. (C) Choice correlations for individual simulated neurons are plotted against those predicted for optimal decoding ($C_k^{opt} = \theta/\theta_k$, the psychophysical/neural threshold ratio; Equation 3), correlation-blind decoding ($C_k^{cb} = \sqrt{c_0}|\sin s_k|$; Equation 2), and correlation-blind decoding combined with substantial information-limiting noise. For this last decoder, the information-limiting noise is responsible for a fraction $\beta$ of the behavioral variance, where $\beta$ is determined by fitting ($C_k^{fit} \approx \beta C_k^{opt} + \gamma C_k^{cb}$). Shaded ellipses again represent 95% confidence intervals for each data point. (D) Relative quality of statistical models, assessed using the corrected Akaike Information Criterion (AICc, see Methods). Smaller AICc are much better, because the relative probability of model *i* given the data depends exponentially on the AICc according to $p_i \propto e^{-AICc_i/2}$. We compare AICc for five models that predict choice correlations: purely optimal decoding (*opt*), a correlation-blind decoder that ignores correlated noise (*cb*), a two-parameter fit based on a weighted sum of the *opt* and *cb* predictors ($fit_2$), a one-parameter fit using only a scaled version of the *opt* predictor ($fit_1$), and a null model attributing all variability to random errors (*null*). In each example shown here, the model with the lowest AICc (asterisk) is in fact the correct one (triangle marked "true"). Since $fit_1$ is a special case of $fit_2$, and *opt* is a special case of $fit_1$, then if the best explanation of the choice correlation data is indeed the optimal predictor, then the $fit_1$ and $fit_2$ models also provide good explanations, although they are penalized for having extra parameters (black, blue). On the other hand, the undecoded population is not optimally decoded, but its choice correlations can be explained as a scaled version of the optimal choice correlations, and is thus best explained by $fit_1$ or $fit_2$ (purple).

these data were used to test the theoretical predictions derived above.

Again we used linear regression to fit coefficients for optimal and suboptimal predictors of choice correlation to determine whether choice correlations were better predicted by optimal or suboptimal decoding. For VN/CN data obtained from four animals, the resultant regression coefficients on the suboptimal predictors were near zero, whereas those on optimal predictors were close to one (Figures 6A and 6B). We conclude that the information available in the vestibular and cerebellar nuclei is used near-optimally, and that behavioral performance is limited primarily by correlated noise, not suboptimal decoding. Likewise, weights for area MSTd in one monkey were consistent with optimal decoding, whereas choice correlations for MSTd in the other monkey were inconclusive.

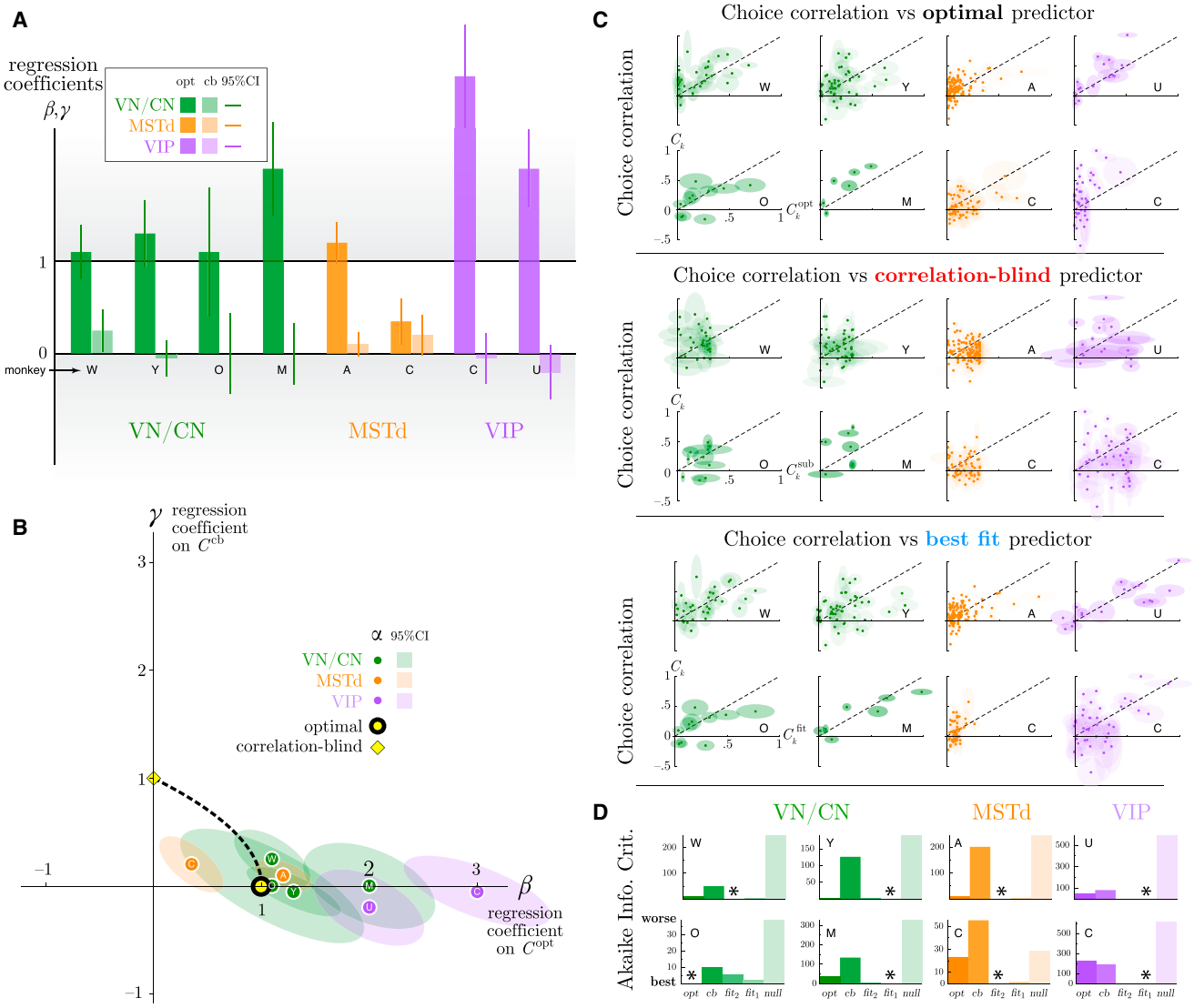


**Figure 6. Inferring the Decoding Quality from Experimental Data**

Plotted as in Figure 5. Neural populations were recorded from VN/CN (green), MSTd (orange), and VIP (purple). Choice correlations are fit to predictions from optimal (Equation 3) and correlation-blind (Equation 2) decoding, and the resultant regression coefficients ($\beta$, $\gamma$) are plotted separately as a pair of bars in (A) and jointly as a point in (B). Each datum combines all recordings from one brain area in one monkey, whose initials are shown below the bars or inside the point. (C) Scatterplots of measured choice correlations for individual neurons are plotted against choice correlations predicted from the optimal (*opt*) and correlation-blind (*cb*) models, as well as the best-fitting linear combination of those two predictors ($fit_2$). Each plot has the initial of the monkey from whom the data was recorded. The plots are arranged in pairs of rows only for compactness. (D) The quality of these three models, plus two additional models $fit_1$ and *null*, was assessed using the AICc, as described for Figure 5. Based on these measures of model quality, data from the heading-discrimination task are largely consistent with animals using VN/CN and MSTd information near-optimally, while neglecting less-informative responses in VIP.

Data from area VIP in two animals revealed near-zero coefficients for the correlation-blind predictor, but interestingly showed coefficients on the optimal predictor that were substantially greater than one (Figures 6A and 6B). As described above, this suggests that the information available in VIP is insufficient to account for behavioral performance. Instead, this finding is consistent with a model in which VIP is not decoded for heading discrimination but nonetheless contains information that is correlated with another area that is decoded near-optimally (see Figure 4). This account is also consistent with recent preliminary results showing that reversible chemical inactivation of VIP does not impair heading discrimination (Klier et al., 2013), despite the fact that VIP neurons exhibit large choice correlations (Chen et al., 2013). Similarly, another recent study of the parietal cortex (area LIP) also reports high choice correlations, yet inactivating this region again has no discernible effect on visual motion discrimination (Yates et al., 2014).

As expected from optimal decoding, we found that the ratio of psychophysical to neural thresholds (Equation 3) did a reasonable job of predicting the measured choice correlations (Figure 6C, top panel) for neurons in VN/CN and MSTd. We emphasize that this prediction has no free parameters. In contrast, the prediction of suboptimal decoding (Equation 2) was poorly matched to data across all areas (Figure 6C, middle panel). The best linear combination of predictors using the regression weights above could provide a better explanation than either predictor alone, but the improvement over the parameter-free optimal decoding predictor was small for VN/CN in all animals and MSTd in one monkey (Figure 6C, bottom panel). To measure model quality, we computed the Akaike Information Criterion for the optimal model (*opt*), the suboptimal (correlation-blind) model (*cb*), and the best linear combination of those two models ($fit_2$). For completeness, we also tested two additional models, one that fits the best scaling of the optimal predictor ($fit_1$), and a null model (*null*) that attributes all variation in measured $C_k$ to random chance (Figure 6D). According to this statistical measure, our data were strong enough to significantly differentiate between the models.

We also examined three additional neural response properties for possible deviations from optimality. First, for optimal linear computation, choice correlations should be zero for neurons with very high thresholds. A few of these uninformative neurons did have choice correlations that differed from zero by more than two standard deviations of measurement uncertainty, but no more than expected by chance ($p = 0.10$, t test on $C_k/\sigma_{C_k}$ for the 19 neurons with unmeasurably large thresholds). Second, on the other end of the sensitivity spectrum, optimal computation requires that no neuron has a better threshold than the behavior. Indeed, our data, like those of a previous study (Cohen and Newsome 2009), reveal that no neurons have a threshold lower than the animal's behavioral threshold, as long as neural thresholds are properly corrected for use of a neuron-antineuron pair (a correction that was not applied in Cohen and Newsome, 2009). Third, there should be no substantially negative choice correlations; if a neuron prefers one stimulus polarity (leftward or rightward heading), it should not drive behavior toward the opposite choice. Although we did observe some negative choice correlations, the associated 95% confidence intervals exclude zero from below for only 4/339 recorded cells, a proportion which is not significant ($p = 0.96$). All of these lines of evidence are consistent with the idea that, on this simple heading discrimination task, the brain uses the vestibular information in its neural populations near-optimally.

To summarize, our results indicate that areas MSTd and VN/CN provide redundant codes that are read out near-optimally, whereas our data suggest that information in area VIP is not used efficiently for this task. From the pattern and scale of choice correlations in VIP, we infer that VIP must be highly redundant with VN/CN, MSTd, or another unrecorded area, in which case the brain loses very little information by ignoring VIP.

## DISCUSSION

The high information content in current models of neural population codes (Shamir and Sompolinsky, 2006; Ecker et al., 2011) would lead to a huge difference between neural thresholds and behavioral thresholds and immeasurably tiny choice correlations if the brain used all of that information efficiently. Since choice correlations are often not small, we consider two alternatives: either the models are incorrect with regard to the high information content of population codes, or the brain is highly suboptimal in extracting the information. Our analysis revealed that these two alternatives have distinguishable consequences, and we tested for these consequences in different brain regions involved in vestibular judgments of self-motion. Our results show that, at least for the simple discrimination task considered here, the first alternative is more likely: the brain has limited information and is able to extract it near-optimally. This is a crucial property of neural coding that must be considered in future theories and experiments.

Our results imply that the information encoded in neural populations is highly redundant and therefore robust to suboptimal decoding. The information-limiting noise correlations cannot be removed, because they look just like the signal. In their presence, there is little advantage to optimally removing all of the remaining components of correlated noise. As a result, a broad range of decoding weights can be near-optimal as long as they produce less variance than the information-limiting correlations. Interestingly, since many decoders would then produce nearly indistinguishable outputs on a trial-by-trial basis, and thus nearly indistinguishable patterns of choice probabilities (Equation 6), it may not be possible to use those choice probabilities to uniquely identify decoding weights from experimental data using the approach of Haefner et al. (2013) (Supplemental Information).

In an early study of correlations between neural activity and behavior (Zohary et al., 1994), the authors wrote that "[t]he covariation of single-neuron responses and psychophysical decisions, an observation that strains credulity at first glance, is a logical consequence of weakly correlated noise within the pool of sensory neurons leading to the decision." Our study shows that this insight remains essentially true: correlated noise can create significant choice correlations. However, weak positive noise correlations do not always produce large choice correlations; this is only true for the model they consider with homogeneous neurons and homogeneous noise correlations. In general, the only noise correlations that lead to significant choice correlations when decoding is near-optimal are those that mimic the effect of the stimulus on population activity, i.e., information-limiting correlations (Moreno-Bote et al., 2014).

Despite these challenges in extracting decoding weights from neural and behavioral data, our results show that one can still fruitfully compare the patterns of choice correlations expected under different hypotheses. By analyzing choice correlations generated by optimal linear decoding, correlation-blind decoding, or a decoder using only a subset of all neurons, we are able to draw strong conclusions about information processing in the brain. Due to measurement noise in the data, it remains possible that neural processing is even better described by some other class of suboptimal decoders that we did not consider. However, we presented theoretical arguments in favor of limited information. First, if cortical populations have extensive information, then the brain would need to throw away almost all of it to explain behavioral performance. Second, since cortical

populations are generally much larger than the population of peripheral sensory neurons but not much noisier, the extensive information model attributes more information to the cortex than to the sensors—and this is prohibited by the data-processing inequality. In addition to these theoretical arguments, the predictions of the limited-information model provide a good match to data. Thus we conclude that the vestibular code for heading is redundant, that the information in VN/CN is used near-optimally, and that although VIP contains robust vestibular information, it is likely not decoded efficiently for this heading discrimination task. This would predict that deactivating VIP should have little effect on the performance of the animal—which is indeed what we have recently found experimentally (Lakshminarasimhan et al., 2014).

In this study, we evaluated whether choice correlations are consistent with optimal or suboptimal decoding. These results depend on the particular class of suboptimal decoders we considered in the extensive information case, namely the correlation-blind decoders. This class is both biologically plausible, well-established in the literature, and quite general, although not all-encompassing. As long as neurons are broadly tuned, noise correlations resemble signal correlations, and the suboptimal decoder does not remove these broad correlations, then the resultant pattern of choice correlations will be close to a sinusoidal function of the preferred direction of the neurons (Equation 2; Figure S4). Nonetheless, it is possible for a suboptimal decoder to produce patterns of choice correlations that differ substantially from Equation 2. Indeed, we can always concoct some combination of correlations and suboptimal decoders that would be consistent with fine details of our measurements. However, this would require unpalatable fine-tuning of the model, and thus would be rejected in favor of the simple limited-information model we offer here, which fits the data quite well.

Our results also depend on models of neural signals and neural noise. Our predictions of choice correlations under suboptimal decoding rely on the structure of noise correlations. We modeled the dominant noise correlations as proportional to signal correlations, a trend that is generally supported by data (Liu et al., 2013). Note that task-dependent changes in correlation amplitude (Cohen and Newsome, 2008) or increases in differential correlations (Bondy and Cumming, 2013) will not change the overall shape of choice correlations, and thus will not qualitatively alter the fact that observed choice correlations are inconsistent with decoders that ignore correlations. This remained true after doubling the time window in which signals were integrated, or equivalently increasing all neural thresholds by $\sqrt{2}$ (Supplemental Information).

In summary, we have presented a theory of how choice correlations depend on the information content of a neural population. A large cortical population that has access to only limited information from its sensors will exhibit a specific form of noise correlations that are difficult to detect yet have an enormous impact on the neural code. One consequence of these information-limiting noise correlations is that a large class of decoders can then extract information near-optimally. Many neurons will then be correlated with behavior because they are correlated with each other. Our data provide evidence that this is the situation in the vestibular system during heading discrimination tasks. These theoretical and experimental conclusions highlight the importance of understanding the detailed structure of noise correlations, for these details fundamentally change how the brain can use and process sensory information.

## EXPERIMENTAL PROCEDURES

### Subjects and Apparatus

Eight rhesus monkeys (*Macaca mulatta*, 4–6 kg) were chronically implanted with an eye coil, a head-restraint ring, and a plastic grid of holes through which guide tubes were passed for electrophysiological recordings (Meng et al., 2005; Gu et al., 2006). All surgical and experimental procedures were approved by the Institutional Animal Care and Use Committee at Washington University and were performed in accordance with institutional and NIH guidelines. Motion stimuli were delivered using a six-degree-of-freedom motion platform (Moog 6DOF2000E), as described previously (Gu et al., 2006).

### Vestibular Heading Discrimination Task

Animals were trained to perform a fine heading discrimination task around psychophysical threshold. During neural recordings in the discrimination task, seven logarithmically spaced headings (±6.4°, ±2.6°, ±1°, and 0° relative to straight ahead) were presented in a block of randomly interleaved trials, while animals maintained fixation on a head-fixed target at the center of the display (2° × 2° electronic window). The range and spacing of headings were chosen carefully to obtain near-maximal psychophysical sensitivity while allowing neural sensitivity to be reliably estimated for most neurons. The motion trajectory (30 cm displacement) was 2 s in duration and followed a Gaussian velocity profile (SD, 0.5 s; peak velocity, 45 cm/s), with a corresponding biphasic linear acceleration profile (±0.1G = ±0.98 $ms^{-2}$).

At the end of each trial of the discrimination task, the fixation point disappeared, two choice targets appeared, and the monkey was trained to make a saccade to the left or right target to report his perceived heading (leftward or rightward relative to an internal standard of straight ahead). Correct choices were rewarded with a drop of water or juice. For the ambiguous straight-forward heading direction (0°), rewards were delivered randomly on half of the trials. If fixation was broken at any time during the 2 s motion stimulus, the trial was aborted and the data were discarded. If neural isolation was lost before completion of at least 10 repetitions of the discrimination task, that neuron was excluded from quantitative analysis. In our sample, cells were held long enough to be tested with at least 20 repetitions of each distinct stimulus for 77% (75/97) of neurons in VN/CN, 95% (174/183) of neurons in MSTd, and 69% (41/59) of neurons in VIP.

### Neural Recordings

We recorded extracellularly the activity of single neurons in the VN/CN, MSTd, and VIP using epoxy-coated tungsten microelectrodes (FHC, 5–7 MΩ impedance for VN/CN, 1–2 MΩ for MSTd and VIP). To target recordings to the VN and CN, we first identified the abducens nuclei bilaterally in initial experiments with each animal. We then used the location of the abducens nuclei to guide electrode penetrations into the CN and VN (Meng et al., 2005; Liu et al., 2013). Area MSTd was located using a combination of magnetic resonance imaging (MRI) scans, stereotaxic coordinates (~15 mm lateral and ~3–6 mm posterior to AP-0), white-/graymatter transitions, and physiological response properties. In some penetrations, electrodes were further advanced into the retinotopically organized area MT. Most recordings concentrated on the posterior/medial portions of MSTd, corresponding to more eccentric, lower hemifield receptive fields in the underlying area MT (Gu et al., 2006, 2007). To localize area VIP, we first identified the medial tip of the intraparietal sulcus and then moved laterally until there was no longer directionally selective visual response in the multiunit activity. At the anterior end, visually responsive neurons gave way to purely somatosensory cells in the fundus. At the posterior end, direction-selective neurons gave way to visual cells that were not selective for motion (Chen et al., 2011).

### Computing Choice Correlations and Thresholds

Behavioral and neural thresholds for the heading task were defined as standard deviations of cumulative Gaussians fits to psychometric or

neurometric functions obtained from ROC analysis (Green and Swets, 1966; Britten et al., 1992; Gu et al., 2008). Choice probabilities (CPs) were also obtained by ROC analysis on neural responses, with balanced $z$-scoring across conditions (Britten et al., 1996; Kang and Maunsell, 2012). Finally, experimental choice probabilities were converted into choice correlations according to $C_k = (\pi/\sqrt{2})(\mathrm{CP}_k - 1/2)$ (Haefner et al., 2013; see below).

### Modeling Neural Responses and Correlations

We model neuronal responses $\mathbf{r}$ as having bell-shaped tuning curves $\mathbf{f}(s)$ and variances equal to the mean. For simulations with extensive information, we constructed noise covariance matrices $\Sigma$ with correlation coefficients $R$ that are on average proportional to the similarity of the pair's tuning, with proportionality $c_0$ (Cohen and Kohn, 2011; Chen et al., 2013; Liu et al., 2013) (Supplemental Information).

For simulations with information-limiting correlations, we added a component to the covariance given by $\varepsilon \mathbf{f}'\mathbf{f}'^T$ (Moreno-Bote et al., 2014) with information-limiting variance $\varepsilon$. For two subpopulations of neurons, we include such noise in each subpopulation separately, with correlations across them. The result is a covariance matrix of those two information-limiting components,

$$E = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} \\ \varepsilon_{xy} & \varepsilon_{yy} \end{pmatrix}. \qquad \text{(Equation 7)}$$

The Supplemental Information derives the full noise covariance for this information-limiting noise.

### Linear Decoding

We model decoding as unbiased linear estimation of the stimulus from the population activity (Equation 1; Figures 1C and 1D). The continuous estimate $\hat{s}$ is converted into a binary behavioral choice around the reference stimulus $s_0$ according to $\mathrm{sgn}(\hat{s}–s_0)$. Properties of linear decoders, including thresholds, information content, and dependence on noise correlations, are described in the Supplemental Information.

### Choice Correlation

Choice correlation $C_k$ is the Pearson correlation coefficient between the neural response and a continuous behavioral "choice" $\hat{s}$. Under the model assumptions, the choice correlation for neuron $k$ is given by

$$C_k = \mathrm{Corr}(\hat{s}, r_k) = \frac{\langle \hat{s} r_k \rangle - \langle \hat{s} \rangle \langle r_k \rangle}{\sqrt{\left(\langle \hat{s}^2 \rangle - \langle \hat{s} \rangle^2\right)\left(\langle r_k^2 \rangle - \langle r_k \rangle^2\right)}} = \frac{(\Sigma \mathbf{w})_k}{\sqrt{\Sigma_{kk}\, \mathbf{w}^T \Sigma \mathbf{w}}}. \qquad \text{(Equation 8)}$$

Previous work (Haefner et al., 2013) showed that the choice probability $\mathrm{CP}_k$ for neuron $k$ is nearly linearly related to the quantity $(\Sigma \mathbf{w})_k / \sqrt{\Sigma_{kk} \mathbf{w}^T \Sigma \mathbf{w}}$, which we recognize as the choice correlation $C_k$:

$$\mathrm{CP}_k \approx \frac{1}{2} + \frac{\sqrt{2}}{\pi} C_k. \qquad \text{(Equation 9)}$$

We can also directly calculate the correlation between neural response and a binary choice $\mathrm{sgn}(\hat{s})$ instead of a continuous choice $\hat{s}$, and find it is exactly proportional to (Equation 8), $\mathrm{Corr}(\mathrm{sgn}(\hat{s}), r_k) = \mathrm{Corr}(\hat{s}, r_k)\sqrt{2/\pi} \approx 0.8\, \mathrm{Corr}(\hat{s}, r_k)$. Given the greater conceptual simplicity of choice correlation (Equation 8), and its near-equivalence to choice probability, we prefer to use it in our analyses.

### Choice Correlations for Optimal Decoding

A locally optimal linear decoding $\hat{s}_{\mathrm{opt}}$ of a neural population uses the weights $\mathbf{w}_{\mathrm{opt}} \propto \Sigma^{-1}\mathbf{f}'$ (Salinas and Abbott, 1994). Substituting these weights into Equation 8, we find choice correlations given by

$$C_k = \frac{\left(\Sigma \Sigma^{-1} \mathbf{f}'\right)_k}{\sqrt{\Sigma_{kk} \mathbf{f}'^T \Sigma^{-1} \mathbf{f}'}} = \frac{f_k'}{\sigma_k} \sigma_{\hat{s}} = \frac{\theta}{\theta_k}, \qquad \text{(Equation 10)}$$

where neural threshold $\theta_k = \sigma_k / f_k'$ and population threshold $\theta = \sigma_{\hat{s}} = \sqrt{1/\mathbf{f}'^T \Sigma^{-1} \mathbf{f}'}$ are the standard deviation of estimators based on a single neuron response and the full population response (Supplemental Information).

### Choice Correlations for Suboptimal Decoding

The factorized decoder ignores correlations, using weights $w_k \propto (f_k'/\sigma_k^2)$ where $\sigma_k^2 = f_k$ for Poisson neurons. In the Supplemental Information we calculate properties of this decoder and more general correlation-blind decoders, leading to Equation 2. We also calculate consequences for a second class of suboptimal decoders, namely reading out only a subpopulation. In the presence of information-limiting noise (Equation 7), this leads to $C_k = \beta C_k^{\mathrm{opt}}$ where $\beta = \varepsilon_{xy}/\varepsilon_{xx}$ (Figure 4C; Supplemental Information).

### Fitting Model Predictions to Data

Choice correlations were fit to the optimal and suboptimal predictors, Equations 3 and 2 respectively, using Bayesian multiple linear regression with heteroscedastic errors in variables (Minka, 1999; Supplemental Information). To assess the quality of the fits, we used the corrected Akaike Information Criterion (AICc; Burnham and Anderson, 2004).

### Simulations

Simulated neural populations had 500 neurons with baseline-shifted von Mises tuning curves of the form

$$f(s) = b + a \exp[\kappa(\cos(s - s_k) - 1)]. \qquad \text{(Equation 11)}$$

Tuning properties were sampled independently with replacement from maximum likelihood parameters for experimentally recorded neurons (Supplemental Information). Median and central quartiles on those tuning parameters were approximately as follows: $a$ = 24 ± 20 Hz; $b$ = 0 for 35% of cells, $b$ = 13 ± 10 Hz for the rest; $\kappa$ = 1 ± 0.5 radians$^{-2}$. Preferred stimuli $s_k$ were drawn randomly from a uniform distribution over $[0,2\pi)$.

The suboptimal correlation-blind decoder was a factorial decoder unless otherwise specified.

For simulations with finite data (Figure 5), 50 neurons were sampled from the simulated population. Tuning curves were re-estimated from ten responses to eight stimulus directions, as described above for real data. Simulated measurements of choice correlations were drawn from a Gaussian distributon with the true mean $C_k$ and variance given above by $\sigma_{C_k}^2 = (1 - C_k^2)^2/(t-1)$ (Supplemental Information), for $t$ = 30 trials.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures and Supplemental Experimental Procedures and can be found with this article at http://dx.doi.org/10.1016/j.neuron.2015.06.033.

## AUTHOR CONTRIBUTIONS

X.P., G.C.D., and A.P. conceived the study. G.C.D., D.E.A., and S.L. designed the experiments. X.P. derived the equations. X.P. and S.L. analyzed the data. X.P., D.E.A., G.C.D., and A.P. wrote the paper.

## ACKNOWLEDGMENTS

Thanks to Aihua Chen and Yong Gu for providing access to their neural recordings, and to Kaushik Lakshminarasimhan for helpful conversations. This work was supported by NIH grant T32DC009974 and a McNair Foundation grant to X.P., by NIH grant EY016178 to G.C.D., by NIH grants EY017866 and DC004260 to D.E.A., and by a James McDonnell Foundation grant and SNF grant 31003A-143707 to A.P.

Received: May 1, 2014
Revised: March 29, 2015
Accepted: June 23, 2015
Published: July 15, 2015

## REFERENCES


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

Berens, P., Ecker, A.S., Cotton, R.J., Ma, W.J., Bethge, M., and Tolias, A.S. (2012). A fast and simple population code for orientation in primate V1. J. Neurosci. *32*, 10618–10626.

Bondy, A., and Cumming, B. (2013). Top down signals influence the distribution of noise correlations amongst sensory neurons. Soc. Neurosci.

Britten, K.H., Shadlen, M.N., Newsome, W.T., and Movshon, J.A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. J. Neurosci. *12*, 4745–4765.

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S., and Movshon, J.A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis. Neurosci. *13*, 87–100.

Burnham, K.P., and Anderson, D.R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociol. Methods Res. *33*, 261–304.

Chen, A., DeAngelis, G.C., and Angelaki, D.E. (2011). Representation of vestibular and visual cues to self-motion in ventral intraparietal cortex. J. Neurosci. *31*, 12036–12052.

Chen, A., Deangelis, G.C., and Angelaki, D.E. (2013). Functional specializations of the ventral intraparietal area for multisensory heading discrimination. J. Neurosci. *33*, 3567–3581.

Cohen, M.R., and Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nat. Neurosci. *14*, 811–819.

Cohen, M.R., and Newsome, W.T. (2008). Context-dependent changes in functional circuitry in visual area MT. Neuron *60*, 162–173.

Cohen, M.R., and Newsome, W.T. (2009). Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. J. Neurosci. *29*, 6635–6648.

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

Fetsch, C.R., Pouget, A., DeAngelis, G.C., and Angelaki, D.E. (2012). Neural correlates of reliability-based cue weighting during multisensory integration. Nat. Neurosci. *15*, 146–154.

Földiák, P. (1993). The 'Ideal Homunculus': statistical inference from neural population responses. In Computation in Neural Systems, F.H. Eeckman and J.M. Bower, eds. (Norwell, MA: Kluwer Academic Publishers), pp. 55–60.

Ghose, G.M., and Harrison, I.T. (2009). Temporal precision of neuronal information in a rapid perceptual judgment. J. Neurophysiol. *101*, 1480–1493.

Graf, A.B., Kohn, A., Jazayeri, M., and Movshon, J.A. (2011). Decoding the activity of neuronal populations in macaque primary visual cortex. Nat. Neurosci. *14*, 239–245.

Green, D.M., and Swets, J.A. (1966). Signal Detection Theory and Psychophysics (New York: Wiley).

Gu, Y., Watkins, P.V., Angelaki, D.E., and DeAngelis, G.C. (2006). Visual and nonvisual contributions to three-dimensional heading selectivity in the medial superior temporal area. J. Neurosci. *26*, 73–85.

Gu, Y., DeAngelis, G.C., and Angelaki, D.E. (2007). A functional link between area MSTd and heading perception based on vestibular signals. Nat. Neurosci. *10*, 1038–1047.

Gu, Y., Angelaki, D.E., and Deangelis, G.C. (2008). Neural correlates of multisensory cue integration in macaque MSTd. Nat. Neurosci. *11*, 1201–1210.

Gu, Y., Liu, S., Fetsch, C.R., Yang, Y., Fok, S., Sunkara, A., DeAngelis, G.C., and Angelaki, D.E. (2011). Perceptual learning reduces interneuronal correlations in macaque visual cortex. Neuron *71*, 750–761.

Haefner, R.M., Gerwinn, S., Macke, J.H., and Bethge, M. (2013). Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nat. Neurosci. *16*, 235–242.

Kang, I., and Maunsell, J.H. (2012). Potential confounds in estimating trial-to-trial correlations between neuronal response and behavior using choice probabilities. J. Neurophysiol. *108*, 3403–3415.

Klier, E.M., Liu, S., Gu, Y., DeAngelis, G.C., and Angelaki, D.E. (2013). Ventral intraparietal (VIP) inactivation does not affect mulitsensory heading perception (San Diego: Annual Meeting of the Society for Neuroscience).

Lakshminarasimhan, K., Liu, S., Klier, E.M., Gu, Y., DeAngelis, G.C., Pitkow, X., and Angelaki, D.E. (2014). Dissecting the contributions of MSTd and VIP to heading perception. Meeting abstract, COSYNE 2014, Salt Lake City, UT, USA.

Liu, S., Gu, Y., DeAngelis, G.C., and Angelaki, D.E. (2013). Choice-related activity and correlated noise in subcortical vestibular neurons. Nat. Neurosci. *16*, 89–97.

Ma, W.J., Beck, J.M., Latham, P.E., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. Nat. Neurosci. *9*, 1432–1438.

Meng, H., Green, A.M., Dickman, J.D., and Angelaki, D.E. (2005). Pursuit–vestibular interactions in brain stem neurons during rotation and translation. J. Neurophysiol. *93*, 3418–3433.

Minka, T. (1999). Linear regression with errors in both variables: a proper Bayesian approach (MIT Media Lab Note).

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

Nienborg, H., and Cumming, B.G. (2007). Psychophysically measured task strategy for disparity discrimination is reflected in V2 neurons. Nat. Neurosci. *10*, 1608–1614.

Purushothaman, G., and Bradley, D.C. (2005). Neural population code for fine perceptual decisions in area MT. Nat. Neurosci. *8*, 99–106.

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

Sanger, T.D. (1996). Probability density estimation for the interpretation of neural population codes. J. Neurophysiol. *76*, 2790–2793.

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

Sompolinsky, H., Yoon, H., Kang, K., and Shamir, M. (2001). Population coding in neuronal systems with correlated noise. Phys. Rev. E Stat. Nonlin. Soft Matter Phys. *64*, 051904.

Tolhurst, D.J., Movshon, J.A., and Dean, A.F. (1983). The statistical reliability of signals in single neurons in cat and monkey visual cortex. Vision Res. *23*, 775–785.

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

Wu, S., Nakahara, H., and Amari, S. (2001). Population coding with correlation and an unfaithful model. Neural Comput. *13*, 775–797.

Yates, J., Katz, L., Park, I.M. Pillow, J., and Huk, A.C. (2014). Dissociated functional significance of choice-related activity across the primate dorsal stream. Meeting abstract, COSYNE 2014, Salt Lake City, UT, USA.

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

Neuron, Volume *87*

## Supplemental Information

# How Can Single Sensory Neurons Predict Behavior?

Xaq Pitkow, Sheng Liu, Dora E. Angelaki, Gregory C. DeAngelis, and Alex Pouget

# How can single sensory neurons predict behavior?
*Supplemental material*

Xaq Pitkow[1,2], Sheng Liu[1], Dora E. Angelaki[1,2], Greg C. DeAngelis[3], and Alexandre Pouget[3,4]

[1]Baylor College of Medicine, Department of Neuroscience
[2]Rice University, Department of Electrical and Computer Engineering
[3]University of Rochester, Department of Brain and Cognitive Sciences
[4]University de Genève, Department of Basic Neuroscience

## Contents



## List of Figures

## S.1 Experimental setup

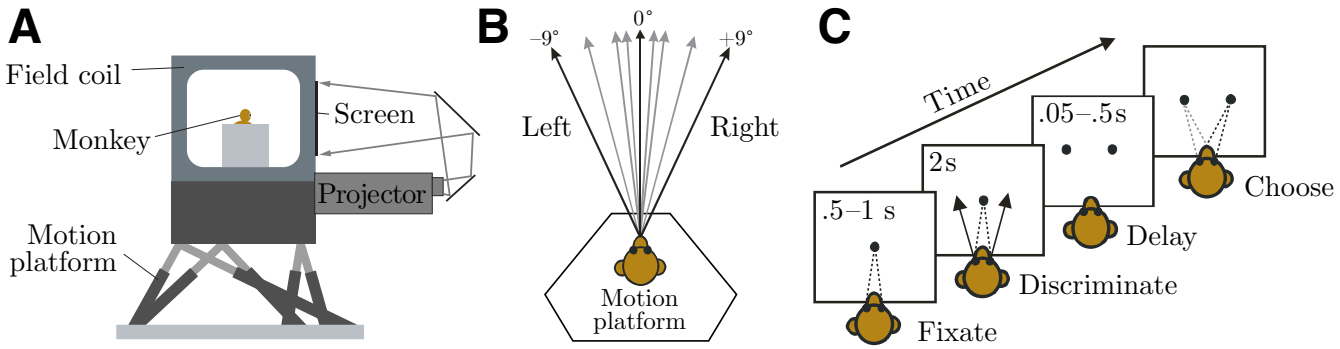


**Figure S1:** Experimental paradigm, related to Figure 1. (**A**) The monkey is seated on a motion platform facing a visual display screen. (**B**) A vestibular self-motion percept is induced by translating the platform in some heading direction. (**C**) In the behavioral task, after the monkey fixates on a target, the target disappears and the motion stimulus begins, following a gaussian speed profile over a period of two seconds. After the motion is completed, two visual targets appear. The monkey reports whether the heading was left or right of 0° (straight ahead) by making a saccade to the left or right visual targets. If correct, the monkey is rewarded with a small amount of juice. On trials with a heading of 0°, the monkey was rewarded randomly half of the trials. After the trial, the motion platform resets to its original position.

## S.2 Computing choice probabilities and thresholds from data

Behavioral performance was quantified by plotting the proportion of 'rightward' choices as a function of heading (the azimuth angle of translation relative to straight ahead). Psychometric data were fit with a cumulative Gaussian function, $P(\text{right}|s) = \frac{1}{2}\text{erfc}\big((\mu - s)/\sigma\sqrt{2}\big)$ where $P(\text{right}|s)$ is the proportion of choices of rightward headings, $s$ is the actual stimulus direction, $\mu$ is the mean of the Gaussian (corresponding to the point of subjective equality) and $\sigma$ is the standard deviation. Psychophysical threshold was defined as the standard deviation of the Gaussian fit, $\sigma$, which corresponds to 68% correct performance (assuming no bias).

For the analyses of neural responses, we used mean firing rates calculated during the middle 400ms interval of each stimulus presentation. To characterize neural sensitivity, we used ROC analysis to compute the ability of an ideal observer to discriminate between two oppositely-directed headings (e.g., $-6.4°$ versus $+6.4°$) based solely on the firing rate of the recorded neuron and a presumed 'antineuron' with opposite tuning [1]. ROC values were plotted as a function of heading, resulting in neurometric functions that were also fit with a cumulative Gaussian function. Neural threshold was defined as the standard deviation of the fitted Gaussian, but increased by a factor of $\sqrt{2}$ to account for the extra information from the antineuron. For a handful of insensitive neurons for which the estimated thresholds were very large, values were truncated to an upper limit of 300°. Because neural and psychophysical thresholds were measured simultaneously during each experimental session, the two could be directly and quantitatively compared.

To quantify the relationship between neural responses and the monkey's perceptual decisions, we also computed 'choice probabilities' using ROC analysis [2]. For each heading, neural responses were sorted into two groups based on the choice that the animal made at the end of each trial: 'preferred' choices refer to decisions that favor the preferred heading of the recorded neuron, whereas 'null' choices refer

to decisions in the opposite direction. ROC values were calculated from these response distributions, yielding a choice probability (CP) for each heading, as long as the monkey made at least 3 choices in favor of each direction. To combine across different headings, we computed a grand CP by balanced $z$-scoring of responses in different conditions, combining $z$-scored response distributions across conditions, and then performing ROC analysis on those distributions [3]. The statistical significance of CPs (i.e., whether they were significantly different from the chance level of 1/2) was determined by a permutation test (1000 permutations).

Other definitions of threshold use a particular performance criterion such as 75% correct, but this merely introduces a proportionality constant for each threshold that cancels in the ratio of (Eq. 3).

## S.3 Analysis procedures for fitting predictions to data

### S.3.1 Multiple linear regression with heteroscedastic errors-in-variables

Eq. 6 shows that the choice correlations can be modeled as a weighted sum of choice correlations corresponding to optimal and suboptimal decoding. The weighting factor is given by the fraction of total uncertainty caused by information-limiting noise. To evaluate the degree of optimality, we perform linear regression of the choice correlations on the optimal and suboptimal predictors. The dependent variable and explanatory variables each have uncertainty, and moreover that uncertainty is different for different data points. Consequently we must perform multiple linear regression with heteroscedastic errors-in-variables. No closed-form solution is available in this case, but we generalize the approach of [4] and compute a log-likelihood directly, and then numerically optimize it to find the maximum marginal likelihood solution.

The linear regression model is $(y_0+\delta y) = \boldsymbol{a}^\top(\boldsymbol{x}_0+\delta\boldsymbol{x})$, where $\boldsymbol{x}_0$ and $y_0$ are the true explanatory and dependent variables, with measurement errors $\delta_x$ and $\delta_y$, and $\boldsymbol{a}$ is the target linear weighting. Combining these variables and errors into the data vector $\boldsymbol{z} = (\boldsymbol{x}, y) = (\boldsymbol{x}_0+\delta_{\boldsymbol{x}}, y_0+\delta_y)$ with given noise covariance $Z$, we have the probability distribution

$$P(\boldsymbol{z}|\boldsymbol{a}, Z, \boldsymbol{x}) = |2\pi Z|^{-1/2} \exp\left[-\tfrac{1}{2}(\boldsymbol{z} - A^\top\boldsymbol{x})^\top Z^{-1}(\boldsymbol{z} - A^\top\boldsymbol{x})\right] \tag{S.1}$$

where $A = (I_x, \boldsymbol{a})$ and $I_{\boldsymbol{x}}$ is the identity matrix with the dimensionality of $\boldsymbol{x}$. We don't know the true value of $\boldsymbol{x}$, so we marginalize over this unknown value, using a prior approximated as a gaussian with mean $\overline{\boldsymbol{x}}_0$ and covariance $X_0$ of the empirical observations $\boldsymbol{x}$:

$$P(\boldsymbol{z}|\boldsymbol{a}, Z) = \int d\boldsymbol{x}\, P(\boldsymbol{x}) P(\boldsymbol{z}|\boldsymbol{a}, Z, \boldsymbol{x}) \tag{S.2}$$

which gives

$$\log P(\boldsymbol{z}|\boldsymbol{a}, Z) = -\frac{1}{2}\sum_i \left[\log|2\pi U| + (\boldsymbol{z} - A^\top\overline{\boldsymbol{x}}_0)^\top U^{-1}(\boldsymbol{z} - A^\top\overline{\boldsymbol{x}}_0)\right] \tag{S.3}$$

where $U = AX_0A^\top + Z$. If each data point $\boldsymbol{z}_i$ has its own covariance $Z_i$, then we can compute the total log-likelihood over a as a sum of such terms,

$$\log P(\{\boldsymbol{z}\}|\boldsymbol{a}, \{Z\}) = \sum_i \log P(\boldsymbol{z}_i|\boldsymbol{a}, Z_i) \tag{S.4}$$

$$= -\frac{1}{2}\sum_i \left[\log|2\pi U_i| + (\boldsymbol{z}_i - A^\top\overline{\boldsymbol{x}}_0)^\top U_i^{-1}(\boldsymbol{z}_i - A^\top\overline{\boldsymbol{x}}_0)\right] \tag{S.5}$$

We maximize this expression numerically to find the best linear model for the data,

$$\hat{\boldsymbol{a}} = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, P(\{\boldsymbol{z}\}|\boldsymbol{a}, \{Z\}) \tag{S.6}$$

Confidence intervals on $\hat{\boldsymbol{a}}$ are computed from the covariance of the likelihood function, given by the negative inverse Hessian of this expression evaluated at the maximum,

$$\Sigma_{\boldsymbol{a}} = -\left[\frac{\partial^2}{\partial \boldsymbol{a}^2} P(\{\boldsymbol{z}\}|\boldsymbol{a}, \{Z\})\bigg|_{\hat{\boldsymbol{a}}}\right]^{-1} \tag{S.7}$$

95% confidence intervals along any single parameter are given by the $2\sigma$ range along that coordinate, $\pm 2(\Sigma_{\boldsymbol{a}})_{ii}$. 95% confidence regions in two dimensions are given by the ellipse for which $(\boldsymbol{a} - \hat{\boldsymbol{a}})^\top \Sigma_{\boldsymbol{a}} (\boldsymbol{a} - \hat{\boldsymbol{a}}) = \Gamma_1^{-1}(1 - .95) \approx 6.17$ where $\Gamma_\alpha^{-1}$ is an inverse incomplete gamma function [5].

In our application, $y$ is the choice correlation $C_k$ and the explanatory vector $\boldsymbol{x}$ is $\left(\theta/\theta_k, \sqrt{c_0}|\sin s_k|\right)$. The prior for $\boldsymbol{x}$ was taken to be a normal distribution

$$P(\boldsymbol{x}) \approx \mathcal{N}\left(\begin{pmatrix} .19 \\ .24 \end{pmatrix}, \begin{pmatrix} .043 & -.005 \\ -.005 & .028 \end{pmatrix}\right) \tag{S.8}$$

specified by the empirical means and covariances of the explanatory data.

We compute uncertainties in choice correlation $C_k$ and predictors $\theta/\theta_k$ and $\sqrt{c_0}|\sin s_k|$ by bootstrapping: we resample with replacement from the neural and behavioral data, recompute these quantities for each resampling, and calculate the variance over the ensemble of resamplings.

To compute these likelihoods we approximated measurement errors in $C_k$ as normally distributed around the true values, with variances $\sigma^2_{C_k}$. Variances for experimentally measured choice correlations and threshold ratios (Eq. 3) were computed by 1000-fold bootstrapping. Variances for the suboptimal predictors (Eq. 2) were obtained from the second derivative of the log-likelihood used to fit the tuning curves.

For model simulations, we generated normally-distributed measurement errors with variance $\sigma^2_{C_k} = (1 - C_k^2)^2/(t - 1)$ for $t$ trials, which is the leading order expression for the variance of a correlation coefficient estimated from a finite number of bivariate gaussian samples [6].

### S.3.2 Robustness to time window of decoding

The predictions of optimal decoding depend on how we measure neural and behavioral thresholds. Ideally, we would like to compute neural thresholds from the same neural signals that drive the behavior, but we don't know precisely what those signals are. For example, some past analyses comparing neurons and behavior [1] integrated neural signals over time windows that likely extend past the animal's decision time, artificially inflating neural sensitivity relative to behavior [15]. Our analysis assumed that the relevant neural signals are the spike counts in a 400 msec time window around the peak movement velocity, rather than the full 2sec movement duration or the central 1sec where the gaussian movement profile is most substantial. However, many neurons in VN/CN and VIP have some sensitivity to acceleration which is small at that time [8, 16], so their threshold may be lower slightly before our chosen time window. Consequently, we tested how sensitive our results would be to a global decrease in neural threshold by $\sqrt{2}$, such as might be caused by advancing or doubling the temporal integration window. The result was an equal decrease in the weight given to the optimal decoding predictor without a concomitant change in the suboptimal predictor weight. Since most inferred optimal weights were slightly greater than 1, this change brings the weights slightly below but still within the 95% confidence intervals around 1. Thus, our conclusions are robust to moderate errors in estimating neural sensitivity.

## S.4 Tuning curves

We characterize neural responses $\boldsymbol{r}$ during discrimination tasks by their stimulus-dependent mean $\boldsymbol{f}(s)$, also known as the tuning curve, and the covariance of their noise given the stimulus, $\Sigma(s)$: $\boldsymbol{r} \sim \mathcal{N}(\boldsymbol{f}(s), \Sigma(s))$.

The tuning curves are approximated by baseline-shifted von Mises functions,

$$f_k(s;\psi_k) = b_k + a_k \exp\left[\kappa_k(\cos\left(s - s_k\right) - 1)\right] \tag{S.9}$$

with parameters $\psi_k = \{b_k, a_k, \kappa_k, s_k\}$ representing baseline firing rate, modulation amplitude, width, and preferred heading respectively (Figure 1A).

We fit this tuning curve model using neural responses to multiple repetitions of eight azimuthal headings were recorded. The parameters $\psi_k = \{b_k, a_k, \kappa_k, s_k\}$ were determined using maximum likelihood estimation assuming Poisson statistics for spike count $r$ in a time window $\Delta_t = 1$sec in response to eight uniformly spaced azimuth angles $s \in \{\pi/4, \ldots, 2\pi\}$

$$\psi_k = \underset{\psi}{\operatorname{argmax}} \sum_s \left(-f(s;\psi)\Delta t + r \log f(s;\psi)\Delta t\right) \tag{S.10}$$

In the maximization, the width parameter was subject to the constraint $\kappa < (8/\pi)^2$ corresponding to the requirement that the tuning curve width be no narrower than the spacing between the eight tested headings.

## S.5 Noise correlations for the extensive noise model

We assume that noise correlation coefficients $R$, in the absence of information-limiting noise, are proportional to the signal correlation coefficients *on average*,

$$\bar{R} = (1 - c_0)I + c_0 R^{\text{sig}} \tag{S.11}$$

where $R^{\text{sig}}$ is the correlation coefficient matrix between mean neural responses $\boldsymbol{f}(s)$ over a uniform distribution of inputs $p(s) = 1/2\pi$. This model is consistent with prior descriptions of measured neural correlations [7–9]. For the tuning curves given above by (Eq. S.9), the signal correlations are given by

$$R^{\text{sig}}_{ij} = \frac{\langle f_i f_j \rangle - \langle f_i \rangle \langle f_j \rangle}{\sqrt{\left(\langle f_i^2 \rangle - \langle f_i \rangle^2\right)\left(\langle f_j^2 \rangle - \langle f_j \rangle^2\right)}} \tag{S.12}$$

$$= \frac{I_0\left(\sqrt{\kappa_i^2 + \kappa_j^2 + 2\kappa_i\kappa_j\cos\left(s_i - s_j\right)}\right) - I_0(\kappa_i)I_0(\kappa_j)}{\sqrt{\left(I_0(2\kappa_i) - I_0(\kappa_i)^2\right)\left(I_0(2\kappa_j) - I_0(\kappa_j)^2\right)}} \tag{S.13}$$

where $I_0$ is a modified Bessel function of the first kind. Note that this expression depends on the difference between the preferred headings of a pair of neurons, $s_i - s_j$ (Figure 1B).

To introduce diversity into the resultant covariances, we first generated a mean covariance matrix that scaled with firing rates according to $\left(\overline{\Sigma}_0\right)_{ij} = \sqrt{f_i f_j}\bar{R}_{ij}$. We then drew a covariance matrix from a Wishart distribution with mean $\overline{\Sigma}_0$ and $2N$ degrees of freedom, where $N$ is the number of neurons:

$$\Sigma_0 \sim \mathcal{W}(\overline{\Sigma}_0, 2N)/2N \tag{S.14}$$

The diversity in the covariance matrix $\Sigma_0$ induced corresponding diversity in the correlation coefficient matrix according to $R_{ij} = \left(\Sigma_0\right)_{ij} / \sqrt{f_i f_j}$ (Figure 1B), such that the correlation coefficient between $R$ and $R^{\text{sig}}$ was around 0.8 for $N = 500$. Figure S4B shows that generating even greater diversity in correlations does not change our main results.

## S.6 Linear decoding and information

A linear decoder (Eq. 1) produces an estimate $\hat{s}$ as a weighted sum of neural responses, $\hat{s} = \boldsymbol{w}^\top(\boldsymbol{r} - \boldsymbol{f}(s_0))$ for reference.

Unbiased decoding requires that our linear estimator is accurate on average, $\langle \hat{s} \rangle_{p(\boldsymbol{r}|s)} = s$, leading to the constraint that $\boldsymbol{w}^\top \boldsymbol{f}' = 1$. The variance of this estimator is given by

$$\sigma_{\hat{s}}^2 = \boldsymbol{w}^\top \Sigma \boldsymbol{w} \tag{S.15}$$

Optimal decoding under these assumptions is achieved by decoding weights given by $\boldsymbol{w} \propto \Sigma^{-1}\boldsymbol{f}'$ [10], where $\boldsymbol{f}' = d\boldsymbol{f}/ds$ and the proportionality is determined by the condition $\boldsymbol{w}^\top \boldsymbol{f}' = 1$ required for unbiased decoding. This optimal decoder then gives a variance of

$$\sigma_{\hat{s}}^2 = \boldsymbol{w}^\top \Sigma \boldsymbol{w} = \frac{\boldsymbol{f}'^\top \Sigma^{-1} \Sigma \Sigma^{-1} \boldsymbol{f}'}{(\boldsymbol{f}'^\top \Sigma \boldsymbol{f}')^2} = \frac{1}{\boldsymbol{f}'^\top \Sigma \boldsymbol{f}'} \tag{S.16}$$

which is precisely equal to the inverse of the linear Fisher information, $J = \boldsymbol{f}'^\top \Sigma \boldsymbol{f}'$.

More generally, we define the linear Fisher information of a (possibly suboptimal) estimator to be the inverse variance of that estimator, $J = 1/\sigma_{\hat{s}}^2$. When decoding is optimal, the linear Fisher information of the decoder is equal to the linear Fisher information for the whole population.

One can also consider an estimator $\hat{s}(r_k)$ based only on the $k$th single neuron response $r_k$, which has variance

$$\sigma_{\hat{s}(r_k)}^2 = \frac{\sigma_k^2}{f_k'^2} \tag{S.17}$$

where $f_k' = df_k'/ds$ is the derivative of the tuning curve with respect to the stimulus and $\sigma_k^2$ is the variance of the neuron's response.

To compute the variance of an unbiased estimator in the presence of information-limiting noise, we use the decomposition of (Eq. 4) to obtain

$$\sigma_{\hat{s}}^2 = \boldsymbol{w}^\top (\Sigma_0 + \varepsilon \boldsymbol{f}' \boldsymbol{f}'^\top) \boldsymbol{w} = \sigma_{0\hat{s}}^2 + \varepsilon \tag{S.18}$$

where $\sigma_{0\hat{s}}^2 = \boldsymbol{w}^\top \Sigma_0 \boldsymbol{w}$ is the estimator variance for $\varepsilon = 0$. Here we used the fact that $\boldsymbol{w}\top \boldsymbol{f}' = 1$ for unbiased decoding, so $\boldsymbol{w}^\top (\varepsilon \boldsymbol{f}' \boldsymbol{f}')^\top \boldsymbol{w} = \varepsilon$. Using the inverse relationship between linear Fisher information and estimator variance, we obtain

$$J = \frac{1}{1/J_0 + \varepsilon} \tag{S.19}$$

where $J_0 = 1/\sigma_{0\hat{s}}^2$ is the information that would have been obtained from that decoder if $\varepsilon = 0$. Thus information-limiting noise prevents the decoded information from exceeding $1/\varepsilon$.

For a fine discrimination task like ours, information-limiting correlations are differential correlations, proportional to $\varepsilon \boldsymbol{f}' \boldsymbol{f}'^\top$ [11], as described in (Eq. 4). More generally, for a coarse discrimination task, the difference in mean neural activity between two stimuli can be denoted $\Delta \boldsymbol{f}$, and the relevant information-limiting correlations are then proportional to $\Delta \boldsymbol{f} \Delta \boldsymbol{f}^\top$. The choice correlations will again be a weighted sum satisfying (Eq. 6). A single noise covariance matrix can also contain information-limiting correlations for many different tasks simultaneously [11].

## S.7 Choice correlations for correlation-blind decoders

In the main text we stated that choice correlations for the correlation-blind decoder were given by $C_k = \sqrt{c_0}|\sin s_k|$ (Eq. 2). We can understand this equation by examining both the suboptimal decoding weights and the noise correlation structure. For the family of suboptimal, correlation-blind decoders we consider, the weight assigned to each neuron tends to have the same sign as its tuning slope at the reference stimulus. Recall that the noise correlations (Eqs. S.11–S.13) are strongest for neurons with similar preferred stimuli, so neurons with preferred directions near the reference stimulus will be positively correlated with other neurons whose preferred stimuli are also near the reference stimulus — but many of these preferences will fall on the other side of the reference and thus have opposite signs of decoding weights. These neurons' response fluctuations are therefore largely cancelled by the decoder, and, thus, cells with preferences near the reference will have weak correlations with the output (Figure S2A). In contrast, neurons with preferred stimuli farther from the reference will be correlated with many other neurons weighted with the same sign, and not with neurons on the other side of the reference that are weighted by the opposite sign (Figure S2B). As a result, there is little cancellation of noise within this group and their collective fluctuations successfully drive the output. This explains why (Eq. 2) shows choice correlations determined by the distance between each cell's preferred heading and the reference stimulus (Figure S2C).

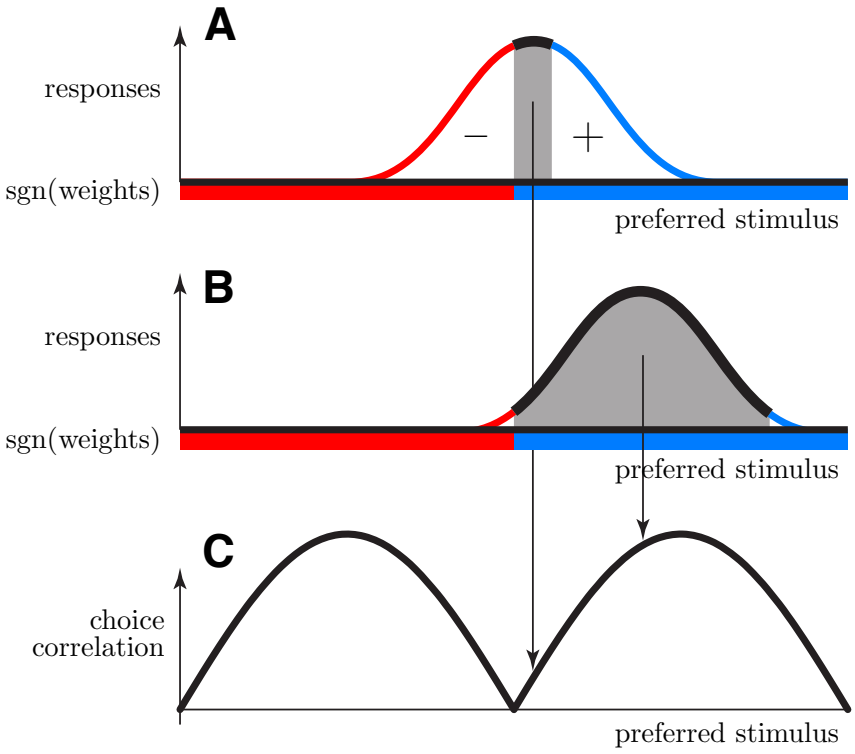


**Figure S2:** Explanation of choice correlations for correlation-blind decoders (Eq. 2), related to Methods. These suboptimal decoders weight neural responses with a sign (shown in red or blue) given by the difference between the preferred stimulus $s_k$ and the reference $s_0$. (**A**) Neurons near the reference have broadly correlated responses but nearly half of these responses are weighted with opposite signs, and thus cancel. (**B**) Neurons far from the reference also have broadly correlated responses but these are mostly weighted with a consistent sign, and thus do not cancel. (**C**) The choice correlations are greatest for neurons whose noise fluctuations are reinforced by many correlated neurons as in (B), rather than cancelling as in (A).

### S.7.1 Choice correlations for the factorial decoder

The factorized decoder uses $\boldsymbol{w} = F^{-1}\boldsymbol{f}'$ where $F$ is a diagonal matrix with response variances along the diagonal; in component form, $w_k \propto f_k'/\sigma_k^2$ where $\sigma_k^2 = f_k$ for Poisson neurons. Substitution into the expression for choice correlation

$$C_k = \frac{(\Sigma \boldsymbol{w})_k}{\sqrt{\Sigma_{kk}\boldsymbol{w}^\top \Sigma \boldsymbol{w}}} \tag{S.20}$$

and using the covariance $\Sigma = F^{1/2} R F^{1/2}$ yields

$$C_k^{\text{fact}} = \frac{(R\boldsymbol{q})_k}{\sqrt{\boldsymbol{q}^\top R \boldsymbol{q}}} \tag{S.21}$$

where $\boldsymbol{q} = F^{-1/2}\boldsymbol{f}'$ is the vector of signal-to-noise ratios for each neuron.

The numerator $(R\boldsymbol{q})_k$ is a weighted sum over the signal-to-noise ratios of many neurons. For broad correlations, this averages away the tuning curve diversity, leaving only the mean $(R\overline{\boldsymbol{q}})_k$, where an overbar designates an average over the tuning parameters $\{b, a, \kappa\}$. Similarly, the quadratic form $\boldsymbol{q}^\top R \boldsymbol{q}$ is approximately $\overline{\boldsymbol{q}}^\top R \overline{\boldsymbol{q}} + \sigma_{\boldsymbol{q}}^2 \operatorname{Tr} R$, and the former term dominates since it scales as $N^2$ rather than $\operatorname{Tr} R = N$. Moreover, if the heterogeneity in correlations arises from a Wishart distribution (Eq. S.14) then variability in $R$ averages away just like the diversity in $\boldsymbol{q}$, leaving us with

$$C_k^{\text{fact}} = \frac{(\bar{R}\overline{\boldsymbol{q}})_k}{\sqrt{\overline{\boldsymbol{q}}^\top \bar{R} \overline{\boldsymbol{q}}}} \tag{S.22}$$

### S.7.2 Self-averaging for the factorial decoder

To demonstrate this self-averaging property of $C_k$, it is sufficient to show that the ratio of variance $\overline{\delta C_k^2}$ to squared mean $\overline{C}_k^2$ approaches zero. Here we actually consider self-averaging of the shape of the choice correlations, so that we may consider only the numerator, defining $X_k = (R\boldsymbol{q})_k$, and show that

$$\frac{\overline{\delta X_k^2}}{\overline{X}_k^2} \to 0 \tag{S.23}$$

as $N$ grows large. Assuming that the diversity of $R$ and $\boldsymbol{q}$ are independent, we have

$$\overline{X}_k = \sum_j \bar{R}_{jk} \overline{\boldsymbol{q}}_j \sim O(N) \tag{S.24}$$

The squared mean is

$$\overline{X}_k^2 = \sum_{ij} \bar{R}_{ik}\bar{R}_{jk}\overline{q}_i\overline{q}_j \tag{S.25}$$

$$= \sum_{i=j=k} 1 1 \overline{q}_k^2 + \sum_{i=j\neq k} \bar{R}_{ik}^2 \overline{q}_i^2 + 2\sum_{i=k\neq j} \bar{R}_{jk} 1 \overline{q}_j \overline{q}_k + \sum_{i,j\neq k} \bar{R}_{ik}\bar{R}_{jk}\overline{q}_i\overline{q}_j \tag{S.26}$$

The second moment of $X_k$ is

$$\overline{X_k^2} = \sum_{ij} \overline{R_{ik}R_{jk}}\,\overline{q_i q_j} \tag{S.27}$$

$$= \sum_{i=j=k} 1 1 \overline{q_k^2} + \sum_{i=j\neq k} \overline{R_{ik}^2 q_i^2} + 2\sum_{i=k\neq j} \bar{R}_{jk} 1 \overline{q}_j \overline{q}_k + \sum_{i,j\neq k} \overline{R_{ik}R_{jk}}\,\overline{q}_i\overline{q}_j \tag{S.28}$$

where we have used the assumption that the heterogeneity in $q_i$ for different neurons is independent. Subtracting these two expressions, we obtain the variance

$$\overline{\delta X_k^2} = \overline{\delta q_k^2} + \sum_{i=j\neq k} \left( \overline{R_{ij}^2 q_i^2} - \bar{R}_{ij}^2 \overline{q}_i^2 \right) + 0 + \sum_{i,j\neq k} \left( \overline{R_{ik} R_{jk}} - \bar{R}_{ik} \bar{R}_{jk} \right) \overline{q}_i \overline{q}_j \tag{S.29}$$

Note that $\overline{R_{ik}R_{jk}} - \bar{R}_{ik}\bar{R}_{jk}$ is the covariance between elements of $R$. Assuming that diversity in $R$ is generated according to a Wishart distribution with $2N$ degrees of freedom, as in (Eq. S.14), this covariance is given by $\left(\bar{R}_{ij}\bar{R}_{kk} + \bar{R}_{ik}\bar{R}_{jk}\right)/2N \sim O(1/N)$. Using the scaling of each term, and the number of terms in each sum, we can now calculate how the variance scales with $N$:

$$\overline{\delta X_k^2} \sim O(1) + \sum^{O(N)} O(1) + \sum^{O(N^2)} O(1/N) \tag{S.30}$$

$$\sim O(N) \tag{S.31}$$

Since the mean scales as $O(N)$, the ratio of the variance to the squared mean scales as $O(N)/O(N^2) \sim O(1/N)$, which approaches zero for large $N$. This shows that the pattern of choice correlations is self-averaging, and we can neglect heterogeneity in computing this pattern.

Note that the self-averaging over $R$ is a consequence of the Wishart distribution, which concentrates around its mean for large $N$. With more heterogeneous covariance matrices, the choice correlations are not necessarily self-averaging over $R$, although they do remain self-averaging with respect to heterogeneity in $\boldsymbol{q}$. Even in this case, the choice correlations for suboptimal, correlation-blind decoders remain well-described by the suboptimal predictor of (Eq. 2), with some additional scatter around the strong sinusoidal trend (Figure S4).

### S.7.3 Approximate choice correlations for the factorial decoder

To compute choice correlations for the self-averaged system, we next exploit the particular structure of neural tuning and noise correlations described in (Eqs. S.9,S.13).

The smooth trend in $R_{ij}^{\text{signal}}$ can be approximated as an offset von Mises function (Eq. S.9) for moderately broad tuning widths $\kappa_i \lesssim 1$,

$$\bar{R}_{ij}^{\text{signal}} \approx 2\frac{\exp\left[\frac{1}{2}\overline{\kappa}(1+\cos\left(s_i - s_j\right))\right] - 1}{\exp\overline{\kappa} - 1} - 1 \tag{S.32}$$

where $\overline{\kappa}$ is the mean tuning width. Since (Eq. S.13) arose as the overlap between two tuning curves, the width of $\bar{R}_{ij}^{\text{signal}}$ is double that of the constituent tuning curves, which explains why the width parameter is $\overline{\kappa}/2$.

The mean signal-to-noise ratios $\overline{q}_i$ are also well approximated by derivatives of von Mises functions,

$$\overline{q}_i \approx -\mu\overline{\kappa}\sin s_i \exp\left[\overline{\kappa}(\cos s_i - 1)\right] \tag{S.33}$$

where $\mu = \left\langle a^{1/2} \right\rangle$ is the average amplitude of the signal-to-noise ratio, and we have set the reference stimulus $s_0 = 0$ for convenience.

From these forms we can understand why $(\bar{R}\overline{\boldsymbol{q}})_k$ is sinusoidal in $s_k - s_0$. The average correlation matrix is circulant, so $\bar{R}_{ij} = \bar{R}_{i-j}$. Circulant matrices are diagonalized in the Fourier basis, with eigenvalues given by the frequency spectrum $\tilde{\bar{R}}_\nu$ for frequency $\nu$. This is given by modified Bessel functions of the first kind $I_\nu$:

$$\tilde{\bar{R}}_\nu = \mathcal{F}\left[\bar{R}_{i-j}\right] \sim I_\nu(\overline{\kappa}/2) \tag{S.34}$$

These functions fall off super-exponentially rapidly with frequency, at least as $\tilde{\bar{R}}_\nu \sim (\overline{\kappa}/4)^\nu/\nu!$ (Amos 1974), and are thus strongly dominated by the lowest frequencies, particularly when tuning curves are broad, with $\overline{\kappa} \lesssim 1$. We assume that the decoder, being unbiased, weights left and right choices antisymmetrically, which implies that the decoder weighting profile is an odd function of the neural sensitivity $\boldsymbol{f}'_k$; therefore so is $\overline{\boldsymbol{q}}$. This cancels all symmetric patterns in the numerator $\bar{R}\overline{\boldsymbol{q}}$ of (Eq. S.22). Combined with the rapid falloff in the frequency spectrum, the dominant antisymmetric noise pattern is $\sin s_k$, with the next pattern at least 8 times weaker. Only if the decoder attenuates this dominant lowest frequency to a level much smaller than the other frequencies will the resultant correlations differ substantially from the sinusoidal prediction of (Eq. 2). Correlation-blind decoders that weight neural responses based solely on their single-neuron properties cannot generically remove this lowest frequency.

To be more explicit, now we approximate the sum $(\bar{R}\overline{\boldsymbol{q}})_k$ by an integral,

$$(\bar{R}\overline{\boldsymbol{q}})_k = \sum_i \left((1-c_0)\delta_{ki} + c_0 \bar{R}^{\text{signal}}_{ki}\right)\overline{q}_i \tag{S.35}$$

$$\approx (1-c_0)\overline{q}_k + \frac{Nc_0}{2\pi}\int ds_i\, R(s_k - s_i)\overline{q}(s_i) \tag{S.36}$$

$$\approx \frac{Nc_0}{2\pi}\int ds_i \left(\frac{\exp\left[\frac{1}{2}\overline{\kappa}(1+\cos(s_k-s_i))-1\right]}{\exp\overline{\kappa}-1} - 1\right)(-\mu\overline{\kappa}\sin s_i \exp\left[\overline{\kappa}(\cos s_i - 1)\right]) \tag{S.37}$$

$$= -2\frac{Nc_0\mu\overline{\kappa}}{2\pi}\frac{e^{-\overline{\kappa}/2}}{e^{\overline{\kappa}}-1}\int ds_i\, \sin s_i \exp\left[\overline{\kappa}\cos s_i + \tfrac{1}{2}\overline{\kappa}\cos(s_k - s_i)\right] \tag{S.38}$$

where some terms have been eliminated since integration over an odd function gives zero and terms proportional to $N$ dominate those of order 1. The exponent can be expanded to give

$$\overline{\kappa}\cos s_i + \tfrac{1}{2}\overline{\kappa}\cos(s_k - s_i) = \overline{\kappa}\cos s_i + \tfrac{1}{2}\overline{\kappa}\left(\cos s_i \cos s_k - \sin s_i \sin s_k\right) \tag{S.39}$$

$$= \overline{\kappa}\left(1 + \tfrac{1}{2}\cos s_k\right)\cos s_i + \tfrac{1}{2}\overline{\kappa}\sin s_k \sin s_i \tag{S.40}$$

which has the form $a\cos x + b\sin x$. The integral in (Eq. S.38) can be expressed as a Bessel function $I_1$,

$$\frac{1}{2\pi}\int dx\, \sin x e^{a\cos x + b\sin x} = b\frac{I_1(\sqrt{a^2+b^2})}{\sqrt{a^2+b^2}} \approx \frac{b}{2} \tag{S.41}$$

where the approximation holds for $\sqrt{a^2+b^2} = \overline{\kappa}\sqrt{(1+\frac{1}{2}\cos s_k)^2 + (\frac{1}{2}\sin s_k)^2} \lesssim 1$ (Press 2007). Thus we have the numerator of (Eq. S.22),

$$\left(\bar{R}\overline{\boldsymbol{q}}\right)_k \approx -\frac{1}{2}Nc_0\mu\overline{\kappa}^2\frac{e^{-\overline{\kappa}/2}}{e^{\overline{\kappa}}-1}\sin s_k \tag{S.42}$$

We can compute the denominator in the same fashion:

$$\overline{\boldsymbol{q}}\bar{R}\overline{\boldsymbol{q}} \approx -\frac{1}{2}Nc_0\mu\overline{\kappa}^2\frac{e^{-\overline{\kappa}/2}}{e^{\overline{\kappa}}-1}\frac{N}{2\pi}\int ds_k\, \overline{q}(s_k)\sin s_k \tag{S.43}$$

with the integral part given by

$$\frac{1}{2\pi}\int ds_k\, \overline{q}(s_k)\sin s_k = \frac{1}{2\pi}\int ds_k\, \left(-\mu\overline{\kappa}e^{\overline{\kappa}(\cos s_k-1)}\sin s_k\right)\sin s_k \tag{S.44}$$

$$= -\mu\overline{\kappa}e^{-\overline{\kappa}}\frac{1}{2\pi}\int ds_k\, e^{\overline{\kappa}\cos s_k}\sin^2 s_k \tag{S.45}$$

$$= -\mu\overline{\kappa}e^{-\overline{\kappa}}\frac{I_1(\overline{\kappa})}{\overline{\kappa}} \tag{S.46}$$

$$\approx -\mu\overline{\kappa}e^{-\overline{\kappa}}\frac{1}{2} \tag{S.47}$$

Substituting the integral, we find

$$\overline{\boldsymbol{q}}\bar{R}\overline{\boldsymbol{q}} \approx \frac{1}{4}N^2c_0\mu^2\overline{\kappa}^3\frac{e^{-3\overline{\kappa}/2}}{e^{\overline{\kappa}}-1} \tag{S.48}$$

Now we can combine the numerator and denominator, to obtain

$$C_k^{\text{fact}} = \frac{(\bar{R}\overline{\boldsymbol{q}})_k}{\sqrt{\overline{\boldsymbol{q}}\bar{R}\overline{\boldsymbol{q}}}} \approx \frac{-\frac{1}{2}Nc_0\mu\overline{\kappa}^2\frac{e^{-\overline{\kappa}/2}}{e^{\overline{\kappa}}-1}}{\sqrt{\frac{1}{4}N^2c_0\mu^2\overline{\kappa}^3\frac{e^{-3\overline{\kappa}/2}}{e^{\overline{\kappa}}-1}}}\sin s_k = -\sqrt{c_0\frac{\overline{\kappa}/2}{\sinh\overline{\kappa}/2}}\sin s_k \tag{S.49}$$

We can expand the sinh to produce

$$C_k^{\text{fact}} \approx -\sqrt{c_0}\sin s_k \tag{S.50}$$

The typical experiment measures choice probability by identifying a positive response with whatever stimulus polarity (e.g., rightward or leftward) produces a higher response for the particular neuron recorded. This means that the positive choice differs for different neurons. Applying the same convention to our mathematical setup introduces a factor of $\operatorname{sgn} f_k' = \operatorname{sgn}(\sin s_k)$ to each choice probability, producing our final prediction for the suboptimal factorial decoder,

$$C_k^{\text{fact}} \approx \sqrt{c_0}\,|\sin s_k| \tag{S.51}$$

From (Eq. 3), we see that choice correlations predicted from optimal decoding depend on neural thresholds but not on the preferred stimulus, whereas predictions from suboptimal decoding (Eq. 2) depend on the preferred stimulus and not explicitly on the neural threshold. Yet, even in (Eq. 3), the choice correlations depends on the preferred stimulus because neural tuning curves tend to have their steepest slope, and thus lowest threshold $\theta_k$, for stimulus values offset from the preferred stimulus. As a consequence, we might worry that the predictions for optimal decoding (Eq. 3) could be satisfied even for a population of neurons decoded suboptimally, leading us to mischaracterize a suboptimal system as optimal.

For a homogeneous population of neurons that is decoded by the factorial decoder (Figure S3A–C), the patterns of choice correlation predicted from optimal decoding are roughly similar to those observed for a suboptimal correlation-blind decoder. However, when tuning curve amplitudes are sufficiently heterogeneous, stimulus preferences are less correlated with neural thresholds, so the relationship between choice correlation and threshold is weaker (Figure S3D–F). And when neural properties are modeled to match experimentally measured diversity in multiple tuning parameters, the relationship between choice correlation and neural threshold is extremely weak for the suboptimal decoder (Figure S3G–I). Thus, under natural diversity, the two predictions for choice correlations, Eqs. 2 and 3, are highly distinct.

Interestingly, (Eq. 2) holds even in the presence of substantial variability in correlations, baseline neural activity, amplitude heterogeneity, and tuning curve width (Figure S3), as well as in the presence

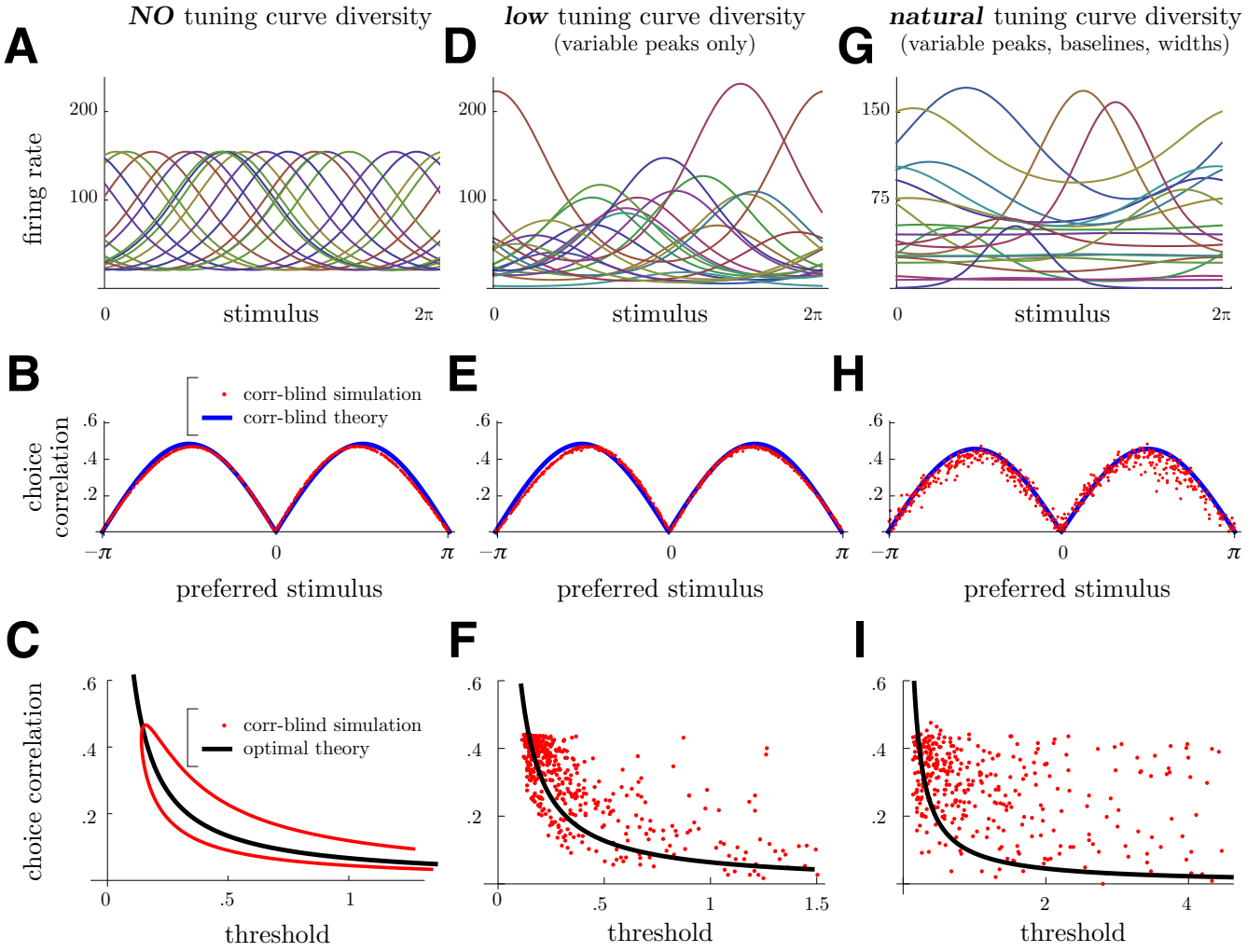


**Figure S3:** (related to Figure 2) As neural heterogeneity increases, the choice correlations predicted for optimal and suboptimal decoding become increasingly distinct. With homogeneous tuning curves (**A**), the preferred heading and neural threshold are closely related, as the neural threshold is smallest when the stimuli to be discriminated lie along the steep flank of the tuning curve. (**B**) The choice correlations for a simulated population read out using a suboptimal correlation-blind decoder (red) are well described by the predicted sinusoid (blue). (**C**) However, the optimal decoding predictor (black) is also able to fit those same choice correlations (red) reasonably well, especially in the presence of typical measurement noise. This is because choice correlations for the optimal decoder are inversely related to the neural discrimination threshold, which, again, has a direct relationship to the preferred stimulus for a homogeneous population. When there is neural diversity only in tuning curve amplitude (**D**), similar though more variable relations hold (**E**,**F**). With the large diversity in amplitude, baseline, and width observed in measured tuning curves (**G**), choice correlations for a correlation-blind decoder remain well described by the correlation-blind predictor. This is true even with correlations that do not have noise correlations that are precisely proportional to the signal correlations (Eq. S.11) but rather exhibit some diversity around that trend (Eq. S.14) (**H**). In contrast, with all of this heterogeneity, those suboptimal choice correlations become poorly fit by the optimal decoding predictor (**I**).

of random deviations (Eq. S.14) around the trend relating noise correlations to signal correlations (Eqs. S.11–S.13). The same result holds for other correlation-blind decoders that weight neurons solely by some (possibly stochastic) function of their individual sensivities (Figure S4A), and even when the relationship between noise correlations and signal correlations is very weak (Figure S4B).

The correlation-blind decoder generates choice correlations similar to the leading antisymmetric eigenvector of $R$. For correlations with considerable diversity around $\bar{R}$, this eigenvector won't look like a perfect sinusoid, but like a sinusoid with some added scatter, as seen in the choice correlations of Figure S4B.

### S.7.4 Choice correlations for suboptimal decoding with information-limiting correlations

Suboptimally decoding neural responses that have information-limiting noise correlations (Eq. 4) leads to choice correlations that can be expressed as a sum of two terms:

$$C_k = \frac{(\Sigma \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} = \frac{(\Sigma_0 \boldsymbol{w} + \varepsilon \boldsymbol{f}' \boldsymbol{f}'^\top \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} \tag{S.52}$$

For unbiased decoding, $\boldsymbol{w}^\top \boldsymbol{f}' = 1$ (see Linear decoding above). Some manipulation gives

$$C_k = \frac{(\Sigma_0 \boldsymbol{w})_k}{\sigma_{0k} \sigma_{0\hat{s}}} \frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} \frac{\sigma_{0k}}{\sigma_k} + \frac{f'_k}{\sigma_k} \sigma_{\hat{s}} \frac{\varepsilon}{\sigma^2_{\hat{s}}} \tag{S.53}$$

where $\Sigma_{0k} = (\Sigma_0)_{kk} \approx \sigma_k$ for small information-limiting noise variance (which has a large effect on information despite the small variance), and where $\sigma_{0\hat{s}}$ is the standard deviation of the estimate produced by the same suboptimal decoder $\boldsymbol{w}$ in the absence of information-limiting correlations, i.e. when the noise covariance is $\Sigma_0$. Since $\sigma^2_{\hat{s}} = \sigma^2_{0\hat{s}} + \varepsilon$, we find that

$$\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} = \sqrt{1 - \varepsilon/\sigma^2_{\hat{s}}} = \sqrt{1 - \alpha} \tag{S.54}$$

with $\alpha = \varepsilon/\sigma^2_{\hat{s}}$. Substituting these into (Eq. S.53) we find that the choice correlation is a weighted sum of the choice correlations for optimal and suboptimal decoding

$$C_k \approx C^{\text{sub}}_k \sqrt{1 - \alpha} + C^{\text{opt}}_k \alpha \tag{S.55}$$

where $C^{\text{sub}}_k$ and $C^{\text{opt}}_k$ are, respectively, the choice correlations for suboptimal decoding with $\varepsilon = 0$, and optimal decoding using $\boldsymbol{w} \propto \Sigma^{-1} \boldsymbol{f}'$ where the covariance included enough information-limiting noise (Eq. 4) to match the observed performance.

## S.8 Infinite suboptimality of the factorial decoder

The variance of the unbiased factorial decoder is given by

$$\sigma^2_{0k} = \frac{\boldsymbol{f}'^\top F^{-1} \Sigma F^{-1} \boldsymbol{f}'}{(\boldsymbol{f}'^\top F^{-1} \boldsymbol{f}')^2} = \frac{\boldsymbol{q}^\top R \boldsymbol{q}}{\|\boldsymbol{q}\|^4} \tag{S.56}$$

As above, the diversity in this ratio is self-averaging, giving

$$\sigma^2_{0\hat{s}} = \frac{\overline{\boldsymbol{q}}^\top \bar{R} \overline{\boldsymbol{q}}}{\|\overline{\boldsymbol{q}}\|^4} \tag{S.57}$$

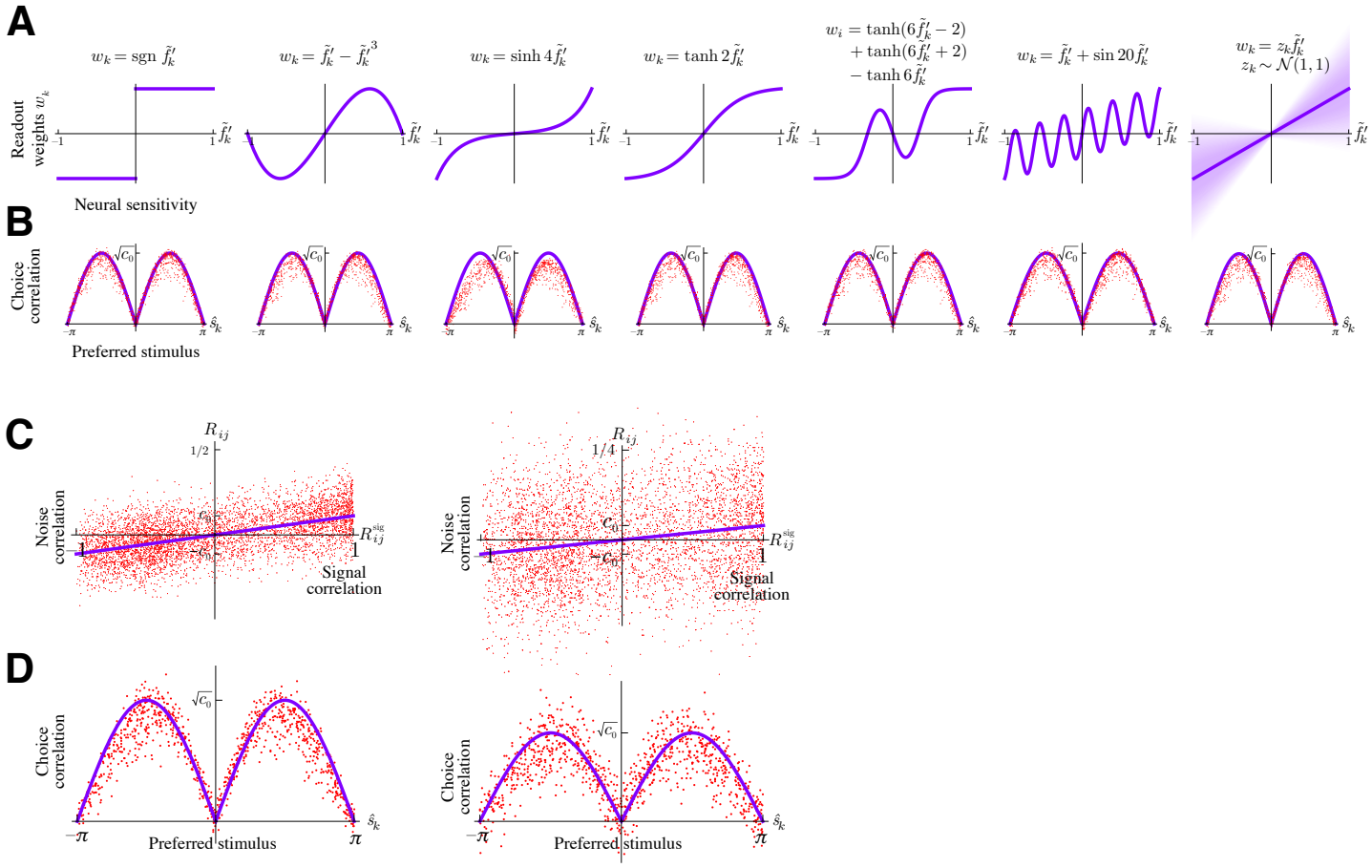


**Figure S4:** (Related to Methods) A large family of suboptimal decoders and correlation matrices generate choice correlations that follow the predictions of (Eq. 2), $C_k \approx \sqrt{c_0}|\sin s_k|$. (**A**) Seven different example suboptimal decoders. Each decoder weights neuron $k$ according to some function $h$ of that neuron's relative sensitivity to the relevant stimulus, $w_k = h(\tilde{f}_k')$, where the relative sensitivity $\tilde{f}_k'$ is the slope of the tuning curve $f_k'$, normalized to range from $-1$ to 1 by $\tilde{f}_k' = f_k'/\max_j |f_j'|$. This is a natural description of the class of correlation-blind decoders, which can only determine the decoding weight on the signal strength in each individual neuron, and not on the correlated neural activity patterns across the population. (**B**) Each of these decoders produces a similar pattern of choice correlations (dots), that is well approximated by (Eq. 2) (curves). This is true even when the function is partially random (rightmost panels) because the randomness tends to average itself away. (**C**) Noise correlations $R$ exhibit considerable diversity (dots) around the trend (Eq. 9, line) that relates them to the signal correlations $R^{\text{sig}}$. Left panel has a slope of $c_0 = 0.11$ and a correlation of 0.5 between $R$ and $R^{\text{sig}}$; right panel has $c_0 = 0.04$ and correlation of 0.2. (**D**) The suboptimal prediction (Eq. 2) holds even with these diverse correlations. Further details for (C) and (D): Since at large $N$ the Wishart distribution described in the Methods (Eq. S.14) produces full-rank matrices that concentrate around their mean, to model highly diverse correlations we generated random log-normal covariance matrices [12]. We first generated symmetric matrices according to $X_{ij} = X_{ji} \sim \mathcal{N}\left((\text{Log}\,\overline{\Sigma})_{ij}, \sigma_X^2\right)$ where Log is the matrix logarithm, $\overline{\Sigma}_{ij} = \sqrt{f_i f_j}\bar{R}_{ij}$ is an average covariance matrix reflecting the relation between noise correlation $\bar{R}$ and the signal correlation $\bar{R}^{\text{sig}}$, and the variance parameter $\sigma_X^2$ determines diversity around this trend. We then applied $\Sigma = \text{Exp}\,X$ where Exp is the matrix exponential. This ensures that the result has controllable diversity while remaining positive definite.

From (Eq. S.48) we see that the numerator scales as $N^2$, and the norm of a vector scales with $N^{1/2}$. Consequently, we find that the variance scales as $N^2/(N^{1/2})^4 \sim$ *constant*. In contrast, the variance of the optimal linear estimator of the same population falls to zero as $N^{-1}$ [13,14]. Not only is the factorial decoder suboptimal, but it is radically so, throwing away almost all the information and saturating to a finite value instead of falling asymptotically toward zero.

## S.9 Choice correlations for decoding a subpopulation with limited information

We now show that if neurons are divided into two populations that receive limited but overlapping information, and only one population is decoded optimally while the other is ignored, then choice correlations in the latter population can be proportional to the optimal choice correlations (Eq. 3) with a proportionality constant greater than 1.

We denote the tuning curves of the two populations by $\boldsymbol{x}$ and $\boldsymbol{y}$, so the full tuning curves are $\boldsymbol{f} = (\boldsymbol{x}, \boldsymbol{y})$, with neural responses on a single trial given by $\boldsymbol{r} = (\boldsymbol{r}_x, \boldsymbol{r}_y)$. For fine discrimination tasks, information-limiting noise in each population separately takes the form of differential correlations [11], $\varepsilon_{xx}\boldsymbol{x}'\boldsymbol{x}'^{\top}$ and $\varepsilon_{yy}\boldsymbol{y}'\boldsymbol{y}'^{\top}$. If these two populations share information, then there will also be correlations between brain regions of the form $\varepsilon_{xy}\boldsymbol{x}'\boldsymbol{y}'^{\top}$ (depicted in Figure 4A). We can express the full noise covariance $\Sigma$ as a rank-two perturbation $UEU^{\top}$ to a covariance $\Sigma_0$ that does not limit information, where

$$U^{\top} = \begin{pmatrix} -\boldsymbol{x}'- & -0- \\ -0- & -\boldsymbol{y}'- \end{pmatrix} \tag{S.58}$$

is a $2 \times N$ matrix and

$$E = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} \\ \varepsilon_{xy} & \varepsilon_{yy} \end{pmatrix} \tag{S.59}$$

is a $2 \times 2$ covariance matrix of the populations' information-limiting correlations. This can also be expressed in block matrix form as

$$\Sigma = \Sigma_0 + UEU^{\top} = \begin{pmatrix} X_0 & Z_0^{\top} \\ Z_0 & Y_0 \end{pmatrix} + \begin{pmatrix} \varepsilon_{xx}\boldsymbol{x}'\boldsymbol{x}'^{\top} & \varepsilon_{xy}\boldsymbol{x}'\boldsymbol{y}'^{\top} \\ \varepsilon_{xy}\boldsymbol{y}'\boldsymbol{x}'^{\top} & \varepsilon_{yy}\boldsymbol{y}'\boldsymbol{y}'^{\top} \end{pmatrix} \tag{S.60}$$

where $X_0$ and $Y_0$ are covariance matrices of $\boldsymbol{r}_x$ and $\boldsymbol{r}_y$ that do not limit information, and $Z_0$ is a cross-covariance between $\boldsymbol{r}_x$ and $\boldsymbol{r}_y$ that again does not limit information. The total linear Fisher information in both populations together can be expressed as

$$J = \boldsymbol{f}'^{\top}\Sigma^{-1}\boldsymbol{f}' \approx \mathbf{1}^{\top}E\mathbf{1} = \frac{\varepsilon_{xx} + \varepsilon_{yy} - 2\varepsilon_{xy}}{\varepsilon_{xx}\varepsilon_{yy} - \varepsilon_{xy}^2} \tag{S.61}$$

where $\mathbf{1} = (1, 1)^{\top}$ and the approximation holds in the limit of many neurons with a $\Sigma_0$ that permits extensive information.

Under the condition that only population $\boldsymbol{x}$ is decoded, yet is decoded optimally, then we have the unbiased weight vector $\boldsymbol{w} = (X_0^{-1}\boldsymbol{x}', \mathbf{0})/J_x$ where $J_x \approx 1/\varepsilon_{xx}$ is the linear Fisher information in population $\boldsymbol{x}$ alone. We can now calculate the choice correlations in the $\boldsymbol{y}$ population, according to

$C_y = (\Sigma \boldsymbol{w})_k / \sqrt{\Sigma_{kk} \boldsymbol{w}^\top \Sigma \boldsymbol{w}}$, which yields

$$C_k = \frac{Z_0 X_0^{-1} \boldsymbol{x}' / J_{0x} + \varepsilon_{xy} \boldsymbol{y}'}{\sqrt{\Sigma_{kk}(J_{0x}^{-1} + \varepsilon_{xx})}} \tag{S.62}$$

$$\approx \frac{\varepsilon_{xy}}{\varepsilon_{xx}} \frac{\theta}{\theta_k} \tag{S.63}$$

The approximation holds in the limit of large populations (hundreds of neurons) where $\Sigma_0$ does not limit information. The choice probabilities in population $\boldsymbol{y}$ can be larger than those in the optimally decoded population (Figure 4C) precisely when there is information-limiting noise in $\boldsymbol{y}$ that is partially correlated with the information-limiting noise in $\boldsymbol{x}$, but with larger amplitude. Under such conditions, population $\boldsymbol{y}$ alone has less information than $\boldsymbol{x}$, so reading out population $\boldsymbol{x}$ alone can be only moderately suboptimal (Figure 4D).

There is a subtle technical point about how to reconcile (Eq. S.63), which shows that $C_k \approx \beta C_k^{\text{opt}}$ with $\beta$ that can be greater than 1, with Eqs. 6 or S.55, which state that $C_k \approx \alpha C_k^{\text{opt}} + \sqrt{1-\alpha} C_k^{\text{sub}}$ with a coefficient on $C_k^{\text{opt}}$ that is less than or equal to 1. At first glance they seem inconsistent, because when $\beta > 1$ one might expect to see a term $\sqrt{1-\beta} C_k^{\text{sub}}$ which cannot be valid. To resolve this, notice that the noise structure (Eq. S.60) induces choice correlations $C_k^{\text{sub}}$ that are proportional to $C_k^{\text{opt}}$ for each population (with different proportionality constants $\zeta$ for each population that are determined by the correlations and decoder). As a result, the two terms in (Eq. 6) can be combined into one:

$$C_k \approx \alpha C_k^{\text{opt}} + \sqrt{1-\alpha} C_k^{\text{sub}} \tag{S.64}$$

$$\approx \alpha C_k^{\text{opt}} + \sqrt{1-\alpha} \zeta C_k^{\text{sub}} \tag{S.65}$$

$$\approx \beta C_k^{\text{opt}} \tag{S.66}$$

which recovers (Eq. S.63). Thus while $\beta$ measures the scale of choice correlations, $\alpha$ represents the fraction of uncertainty caused by information-limiting noise rather than suboptimal decoding and remains between 0 and 1. Even though they may be confounded by some measurements, $\alpha$ and $\beta$ are conceptually distinct.

### S.9.1 Properties of choice correlations with optimal decoding

Our analytical results reveal several properties of choice correlation that defy conventional wisdom.

One common expectation is that choice correlation should be smaller for larger populations since information is distributed across more neurons. However, if the code is redundant and decoded near-optimally, then the choice correlations do not vary with population size. This is because information saturates with the number of neurons, so the ratio of the population threshold to an individual neuron's threshold saturates too. According to (Eq. 3), the choice correlations are equal to this ratio, and therefore they saturate as well.

A recent study [20] demonstrated how to reconstruct the weights that the brain gives to different neurons from choice probability measurements and correlated noise. Despite the elegance of this result, one consequence of our findings is that such a full reconstruction may not be feasible in the presence of information-limiting correlations. Due to the robustness of the resultant neural code, many different decoding weight profiles can extract nearly all of the available information (Fig. 4). Under conditions in which behavior is limited by noise ($\alpha$ near 1), (Eq. 6) shows that choice correlations will be dominated by the optimal term $\alpha C_k^{\text{opt}}$ regardless of the particular structure of the decoding weights, while the suboptimal term $\sqrt{1-\alpha} C_k^{\text{sub}}$ will be small. Any resultant deviations of choice correlations from the

optimal predictor are therefore likely to be subtle and difficult to detect in the presence of measurement noise. Since many weight profiles would generate almost the same choice probabilities, it may not be possible to use choice probabilities to compute decoding weights from experimental data using the approach of [20].

Another expectation based on previous results is that choice correlations should be greater when noise correlations are larger [8, 18]. This is true only for a specific pattern of correlations, or for certain types of decoders. For suboptimal decoding of highly informative populations, choice correlations do increase with the overall scale of correlated noise (Eq. 2). However, our results instead favor near-optimal decoding, for which most correlations play little role: only information-limiting correlations increase the choice correlations according to (Eq. 3). Since these special noise correlations may constitute only a small portion of the total measured noise covariance, choice correlations could potentially rise even while the total noise correlation falls.

A consequence of optimal decoding that emerges from (Eq. 3) is that choice correlations should depend on behavioral performance: neurons with similar stimulus sensitivities should have higher choice correlations in animals that have higher behavioral thresholds. Conversely, if different brain regions seem to have different choice correlations, one must take care to account for differences in psychophysical sensitivity between animals before concluding that the difference between brain areas is meaningful. The surest diagnostic of differences between brain regions would be to measure choice correlations from multiple areas within a single animal.

Finally, the results of [20] imply that if decoding is optimal, choice correlation for neuron $k$ should be proportional to that neuron's sensitivity, $1/\theta_k$. This is not sufficient to demonstrate optimality, however: such a pattern of choice correlations indicates optimal decoding if and only if the proportionality constant equals the threshold of population activity, $\theta$, which is the same as the behavioral threshold when decoding is optimal. All other values of the proportionality constant reflect suboptimal decoding (Figure 4).

## S.10 References


[1] Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. Journal of Neuroscience 12: 4745–4765.

[2] Green DM, Swets JA (1966) Signal detection theory and psychophysics. John Wiley.

[3] Kang I, Maunsell JH (2012) Potential confounds in estimating trial-to-trial correlations between neuronal response and behavior using choice probabilities. Journal of Neurophysiology 108: 3403–3415.

[4] Minka T (1999) Linear regression with errors in both variables: A proper bayesian approach. MIT Media Lab Technical report : 1–14.

[5] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical Recipes 3rd edition: The Art of Scientific Computing. Cambridge University Press.

[6] Kenney J (1940) Mathematics of Statistics, Part Two. Chapman and Hall.

[7] Cohen MR, Kohn A (2011) Measuring and interpreting neuronal correlations. Nature Neuroscience 14: 811–819.

[8] Liu S, Gu Y, DeAngelis GC, Angelaki DE (2013) Choice-related activity and correlated noise in subcortical vestibular neurons. Nature Neuroscience 16: 89–97.

[9] Chen A, DeAngelis GC, Angelaki DE (2013) Functional specializations of the ventral intraparietal area for multisensory heading discrimination. Journal of Neuroscience 33: 3567–3581.

[10] Salinas E, Abbott L (1994) Vector reconstruction from firing rates. Journal of Computational Neuroscience 1: 89–107.

[11] Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A (2014) Information-limiting correlations. Nature Neuroscience 17: 1410–1417.

[12] Schwartzman A (2013) Lognormal distributions and geometric averages of positive definite matrices. arXiv stat.ME: 1407.6383.

[13] Shamir M, Sompolinsky H (2006) Implications of neuronal diversity on population coding. Neural Computation 18: 1951–1986.

[14] Ecker AS, Berens P, Tolias AS, Bethge M (2011) The effect of noise correlations in populations of diversely tuned neurons. Journal of Neuroscience 31: 14272–14283.

[15] Cohen MR, Newsome WT (2009) Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. Journal of Neuroscience 29: 6635–6648.

[16] Chen A, DeAngelis GC, Angelaki DE (2011) Representation of vestibular and visual cues to self-motion in ventral intraparietal cortex. Journal of Neuroscience 31: 12036–12052.

[17] Drugowitsch J, DeAngelis GC, Klier EM, Angelaki DE, Pouget A (2014) Optimal multisensory decision-making in a reaction-time task. eLife 3: e03005.

[18] Nienborg H, Cumming BG (2009) Decision-related activity in sensory neurons reflects more than a neurons causal effect. Nature 459: 89–92.

[19] Wohrer A, Machens C (2013) Percept formation from neural populations in sensory decision-making tasks. arXiv q-bio.NC: 1303.1939.

[20] Haefner RM, Gerwinn S, Macke JH, Bethge M (2013) Inferring decoding strategies from choice probabilities in the presence of correlated variability. Nature Neuroscience 16: 235–242.