Revealing nonlinear neural decoding by analyzing choices

Qianli Yang[1] and Xaq Pitkow[1,2,3]

[1]Rice University, Department of Electrical and Computer Engineering
[2]Baylor College of Medicine, Department of Neuroscience
[3]Baylor College of Medicine, Center for Neuroscience and Artificial Intelligence

Sensory data about most natural task-relevant variables are confounded by task-irrelevant sensory variations, called nuisance variables. To be useful, the sensory signals that encode the relevant variables must be untangled from the nuisance variables through nonlinear transformations, before the brain can use or decode them to drive behaviors. The information to be untangled is represented in the cortex by the activity of large populations of neurons, constituting a nonlinear population code. Here we provide a new way of thinking about nonlinear population codes and nuisance variables, leading to a theory of nonlinear feedforward decoding of neural population activity. This theory obeys fundamental mathematical limitations on information content that are inherited from the sensory periphery, producing redundant codes when there are many more cortical neurons than primary sensory neurons. The theory predicts a simple, easily computed quantitative relationship between fluctuating neural activity and behavioral choices if the brain uses its nonlinear population codes optimally: more informative patterns should be more correlated with choices.

1 Introduction

How does an animal use, or ‘decode’, the information represented in its brain? When the average responses of some neurons are well-tuned to a stimulus of interest, this is straightforward. In binary discrimination tasks, for example, a choice can be reached simply by a linear weighted sum of these tuned neural responses. Yet real neurons are rarely tuned to precisely one variable: variation in multiple stimulus dimensions influence their responses. As we show below, this can dilute or even abolish the mean tuning to the relevant stimulus. The brain cannot simply use linear computation, nor can we understand neural processing using linear models.

To see this problem in a simple case, imagine a simplified model of a visual neuron that includes an oriented edge-detecting linear filter followed by additive noise, with a Gabor receptive field like simple cells in primary visual cortex (Figure 1A). If an edge is presented to this model neuron, different rotation angles will change the overlap, producing a different mean. This neuron is then tuned to orientation.

However, when the edge has the opposite polarity, with black and white reversed, then the linear response is reversed also. If the two polarities occur with equal frequency, then the positive and negative responses cancel on average. The mean response of this linear neuron to any given orientation is therefore precisely constant, so the model neuron is untuned.

Notice that stimuli aligned with the neuron’s preferred orientation will generally elicit the highest or lowest response magnitude, depending on polarity. Edges with the smallest response to one polarity will also have the smallest response to its inverse. Thus, even though the mean response of this linear neuron is zero, independent of orientation, the *variance* is tuned.

To estimate the variance, and thereby the orientation itself, the brain can compute the square of the linear responses. This would allow the brain to estimate the orientation independently from polarity. This is consistent with the well-known energy model of complex cells in primary visual cortex, which use squaring nonlinearities to achieve invariance to the polarity of an edge [1]. We will return to this paradigmatic example of simple nonlinear computation throughout this article.

Generalizing from this example, we identify edge polarity as a ‘nuisance variable’ — a property in the world that alters how task-relevant stimuli appear but is, itself, irrelevant for the current task (here, perceiving orientation). Other examples of nuisance variables include the illuminant for guessing surface color, position for object recognition, expression for face identification, or pitch for speech recognition. Nuisance variables generally make it hard to extract the task-relevant variables from sense data, which is the central task of perception [2–5]. (Of course, what is a nuisance for one task might be a target variable in another task, and vice versa.)

The prevailing neuroscience view of this disentangling process is deterministic: the output of a complex (often multi-stage) nonlinear function identifies the variables of interest [2, 3, 6]. Here we take a statistical perspective: the brain learns from its history of sensory inputs which statistics of its many sense data can be used to extract the task-relevant variable. In the orientation estimation task above, the relevant statistic was not the mean but the variance.

Just because a neural population encodes information, it does not mean that the brain decodes it all. Here, *encoding* specifies how the neural responses relate to the stimulus input; similarly, *decoding* specifies how the neural responses relate to the behavioral output. To understand the brain’s computational strategy we must understand how encoding and decoding are related, *i.e.* how the brain uses the information it has. As we will see, our statistical perspective provides a simple way of testing whether the brain’s decoding strategy is efficient, based on whether neural response patterns that are informative about the task-relevant sensory input are also informative about the animal’s behavior in the task.

2 Results

2.1 Task, stimulus, neural responses, action

To specify our mathematical framework for nonlinear decoding, we model a task, a stimulus with both relevant and irrelevant variables, neural responses, and behavioral choices.

In our task, an agent observes a multidimensional stimulus $(s, \boldsymbol{n})$ and must act upon one particular relevant aspect of that stimulus, s, while ignoring the rest, $\boldsymbol{n}$. The irrelevant stimulus aspects serve as nuisance variables for the task (the letter $\boldsymbol{n}$ stands for nuisance).

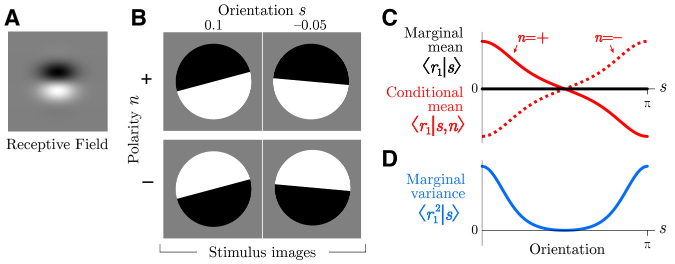

Figure 1: Simple nonlinear code for orientation induced by two polarities. (**A**) Receptive field for a linear neuron. (**B**) Four example images, each with an orientation $s \in [0, \pi)$ and a polarity $\boldsymbol{n} \in \{-1, +1\}$. (**C**) Even though the mean response of the linear neuron is tuned to orientation if polarity were specified (conditional mean, red), the mean response is untuned when the polarity is unknown and could take any value (marginal mean, black). (**D**) Tuning is recovered by the marginal variance even if the polarity is unknown (blue).

Together, these stimulus properties determine a complete sensory input that drives some responses $\boldsymbol{r}$ in a population of N neurons according to the distribution $p(\boldsymbol{r}|s, \boldsymbol{n})$.

We consider a feedforward processing chain for the brain, in which the neural responses $\boldsymbol{r}$ are nonlinearly transformed downstream into other neural responses $\boldsymbol{R}(\boldsymbol{r})$, which in turn are used to create a perceptual estimate of the relevant stimulus $\hat{s}$:

$$(s, \boldsymbol{n}) \to \boldsymbol{r} \to \boldsymbol{R} \to \hat{s} \tag{1}$$

We model the brain’s estimate as a linear function of the downstream responses $\boldsymbol{R}$. Ultimately these estimates are used to generate an action that the experimenter can observe. Here we assume that the task is local or fine-scale estimation: the subject must directly report its estimate $\hat{s}$ for the relevant stimuli near a reference s_0. We measure performance by the variance of this estimate, $\sigma^2_{\hat{s}}$.

We assume that we have recorded activity only from some of the upstream neurons, so we don’t have direct access to $\boldsymbol{R}$, only $\boldsymbol{r}$. Nonetheless we would like to learn something about the downstream computations used in decoding. In this paper we show how to use the statistics of cofluctuations in $\boldsymbol{r}$ and $\hat{s}$ to estimate the quality of nonlinear decoding.

2.2 Signal and noise

The population response, which we take here to be the spike counts of each neuron in a specified time window, reflects both *signal* and *noise*, where signal is the repeatable stimulus-dependent aspects of the response, and noise reflects trial-to-trial variation. Conventionally in neuroscience, the signal is often thought to be the stimulus dependence of the *average* response, *i.e.* the tuning curve $\boldsymbol{f}(s) = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s) = \langle \boldsymbol{r}|s\rangle$ (angle brackets denote an average over all responses given the condition after the vertical bar). Below we will broaden this conventional definition to allow the signal to include any stimulus-dependent statistical property of the population response.

Noise is the non-repeatable part of the response, characterized by the variation of responses to a fixed stimulus. It is convenient to distinguish *internal* noise from *external* noise. Internal noise is internal to the animal, and is described by response distribution $p(\boldsymbol{r}|s,\boldsymbol{n})$ when everything about the stimulus is fixed. This could also include uncontrolled variation in internal states [7–10], like attention, motivation, or wandering thoughts. External noise is variability generated by the external world, or nuisance variables, such as the positions of all dots in a random dot kinematogram [11] or the polarity of an edge (Figure 1). External noise leads to a neural response distribution $p(\boldsymbol{r}|s)$ where only the relevant variables are held fixed. Both types of noise can lead to uncertainty about the true stimulus.

Trial-to-trial variability can of course be correlated across neurons. Neuroscientists often measure two types of second-order correlations: signal correlations and noise correlations [12–20]. Signal correlations measure shared variation in responses $\boldsymbol{r}$ averaged over the set of stimuli s: $\rho_{\text{signal}} = \text{Corr}(\boldsymbol{r})$. (Internal) noise correlations measure shared variation that persists even when the stimulus is completely identical, nuisance variables and all: $\rho_{\text{noise}}(s,\boldsymbol{n}) = \text{Corr}(\boldsymbol{r}|s,\boldsymbol{n})$.

For multidimensional stimuli, however, these are only two extremes on a spectrum, depending on how many stimulus aspects are fixed across the trials to be averaged. We propose an intermediate type of correlation: *nuisance correlations*. Here we fix the task-relevant stimulus variable(s) s, and average over the nuisance variables $\boldsymbol{n}$: $\rho_{\text{nuisance}} = \text{Corr}(\boldsymbol{r}|s)$. Just as signal correlations don't mean correlations between signals, nuisance correlations are not correlations between nuisance variables, but rather between neural responses induced by the external noise or nuisance variation. Of course nuisance correlations will be task-dependent, since the task determines which variables are nuisance and which are relevant [21, 22].

Critically, but confusingly, some so-called 'noise' correlations and nuisance correlations actually serve as signals. This happens whenever the statistical pattern of trial-by-trial fluctuations depends on the stimulus, and thus contain information. For example, a stimulus-dependent noise covariance functions as a signal. There would still be true noise, *i.e.* irrelevant trial-to-trial variability that makes the signal uncertain, but it would be relegated to higher-order fluctuations [23] such as the variance of the response covariance (Figure 2D, Table 1). Stimulus-dependent correlations, principally due to nuisance variation, lead naturally to nonlinear population codes, as we will explain below.

2.3 Nonlinear encoding by neural populations

Most accounts of neural population codes actually address *linear* codes, in which the mean response is tuned to the variable of interest and completely captures all signal about it [24–28]. We call these codes linear because the neural response property needed to best estimate the stimulus near a reference (or even infer the entire likelihood of the stimulus, Supplement S1.2) is a linear function of the response. Linear codes for different variables may arise early in sensory processing, or after many stages of computation [2, 5].

If any of the relevant signal can only be extracted using nonlinear functions of the neural responses, then we say that the population code is nonlinear.

It is illuminating to take a statistical view: unlike a linear code, the information is not encoded in mean neural responses but instead by higher-order statistics of responses [16, 29]. These functional and statistical views are naturally linked because estimating higher-order statistics requires nonlinear operations. For instance, information from a stimulus-dependent covariance $Q(s) = \left\langle \boldsymbol{r}\boldsymbol{r}^{\top}|s\right\rangle$ can be decoded by quadratic operations $\boldsymbol{R} = \boldsymbol{r}\boldsymbol{r}^{\top}$ [22, 30, 31]. Table 1 compares the relevant neural response properties for linear and nonlinear codes.

A simple example of a nonlinear code is the exclusive-or (XOR) problem. Given the responses of two binary neurons, r_1 and r_2, we would like to decode the value of a task-relevant signal $s = \text{XOR}(r_1, r_2)$ (Figure 2A). We don't care about the specific value of

	linear	nonlinear	quadratic
raw data	$\boldsymbol{r}$	$\boldsymbol{R}(\boldsymbol{r})$	$\boldsymbol{r}\boldsymbol{r}^\top$
signal	$\text{Mean}(\boldsymbol{r}\mid s)$	$\text{Mean}(\boldsymbol{R}\mid s)$	$\text{Mean}(\boldsymbol{r}\boldsymbol{r}^\top\mid s)$
noise	$\text{Cov}(\boldsymbol{r}\mid s)$	$\text{Cov}(\boldsymbol{R}\mid s)$	$\text{Cov}(\boldsymbol{r}\boldsymbol{r}^\top\mid s)$

Table 1: Neural response properties relevant for linear and nonlinear codes. In each case, the brain must estimate the stimulus from a single example of neural data, but the relevant function of that data is linear for linear codes, and nonlinear for nonlinear codes. The noise and signal can be quantified by the corresponding covariance and stimulus-dependent changes in the corresponding means (*i.e.* the tuning curve slope).

r_1 by itself, and in fact r_1 alone tells us nothing about s. The same is true for r_2. The signal is actually reflected in the trial-by-trial *correlation* between r_1 and r_2: when they are the same then $s = -1$, and when they are opposite then $s = +1$. The correlation, and thus the relevant variable s, can be estimated nonlinearly from r_1 and r_2 as $\hat{s} = -r_1 r_2$.

Some experiments have reported stimulus-dependent internal noise correlations that depend on the signal, even for a completely fixed stimulus without any nuisance variation [32–36]. Other experiments have turned up evidence for nonlinear population codes by characterizing the nonlinear selectivity directly [6, 37, 38].

More typically, however, stimulus-dependent correlations arise from external noise, leading to what we call nuisance correlations. In the introduction (Figure 1) we showed a simple orientation estimation example in which fluctuations of an unknown contrast eliminate the orientation tuning of mean responses, relegating the tuning to variances. Figure 2B–E shows a slightly more sophisticated version of this example, where instead of two image polarities, we introduce spatial phase as a continuous nuisance variable. This again eliminates mean tuning, but introduces nuisance covariances that are orientation tuned.

One might object that although the nuisance covariance is tuned to orientation, a subject cannot compute the covariance on a single trial because it does not experience all possible nuisance variables to average over. This objection stems from a conceptual error that conflates the tuning (signal) with the raw sense data (signal+noise). In linear codes, the subject does not have access to the tuned mean response $\langle\boldsymbol{r}|s\rangle$, just a noisy single-trial version of the mean, namely $\boldsymbol{r}$. Analogously, the subject does not need access to the tuned covariance, just a noisy single-trial version of the covariance, $\boldsymbol{r}\boldsymbol{r}^\top$ (Table 1). In this simple example, the nuisance variable of spatial phase ensures that quadratic statistics contains relevant information.

2.4 Decoding and choice correlations

To study how neural information is used or decoded, past studies have examined whether neurons that are sensitive to sensory inputs also reflect an animal's behavioral outputs or choices [39–47]. However, this choice-related activity is hard to interpret, because it may reflect decoding of the recorded neurons, or merely correlations between them and other neurons that are decoded instead [48].

In principle, we could discount such indirect relationships with complete recordings of all neural activity. This is currently impractical for most animals, and even if we could record from all neurons simultaneously, data limitations would prevent us from fully disambiguating how neural activities directly influence behavior.

To understand key principles of neural computation, however, we may not care about all detailed patterns of decoding weights and their underlying synaptic connectivity. Instead we may want to know only certain properties of the brain's strategies. One important property is the efficiency with which the brain decodes available neural information as it generates an animal's choices.

Conveniently, testable predictions about choice-related activity can reveal the brain's decoding efficiency, in the case of linear codes [28]. Next we review these predictions, and then generalize them to nonlinear codes.

2.5 Choice correlations predicted for optimal linear decoding

We define 'choice correlation' C_{r_k} as the correlation coefficient between the response r_k of neuron k and the stimulus estimate (which we view as a continuous 'choice') $\hat{s}$, given a fixed stimulus s:

$$C_{r_k} = \text{Corr}(r_k, \hat{s}|s) \tag{2}$$

This choice correlation is a conceptually simpler and more convenient measure than the more conventional statistic, 'choice probability' [49], but it has almost identical properties (Methods 4.2) [28, 48].

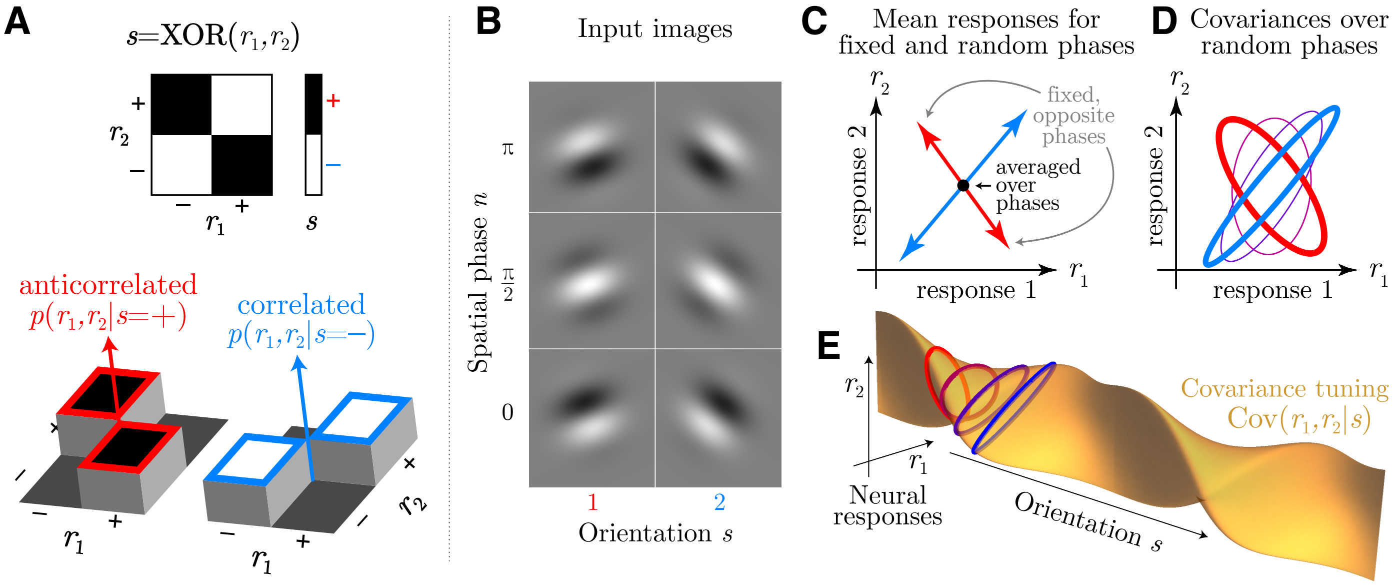

Figure 2: Nonlinear codes. **A**: Simple example in which a stimulus s is the XOR of two neural responses (top). Conditional probabilities $p(r_1, r_2|s)$ of those responses (bottom) show they are anti-correlated when $s = +1$ (red) and positively correlated when $s = -1$ (blue). This stimulus-dependent correlation between responses creates a nonlinear code. The remaining panels show that a similar stimulus-dependent correlation emerges in orientation discrimination with unknown spatial phase. **B**: Gabor images with two orientations and three spatial phases. **C**: Mean responses of linear neurons with Gabor receptive fields are sensitive to orientation when phase is fixed (arrows), but point in different directions for different spatial phases. When phase is an unknown nuisance variable, this mean tuning therefore vanishes (black dot). **D**: The response covariance Cov$(r_1, r_2|s)$ between these linear neurons is tuned to orientation even when averaging over spatial phase. Response covariances for four orientations are depicted by ellipses. **E**: A continuous view of the covariance tuning to orientation for a pair of neurons.

Intuitively, if an animal is decoding its neural information efficiently, then those neurons encoding more information should be more correlated with the choice. Mathematically, one can show that choice correlations indeed have this property when decoding is optimal [28]:

$$C^{\text{opt}}_{r_k} = \sqrt{\frac{J_{r_k}}{J}} \tag{3}$$

where J and J_{r_k} are, respectively, the linear Fisher Information [23] based on the entire population $\boldsymbol{r}$ or on neuron k's response r_k (Methods 4.2). This relationship holds for a locally optimal linear estimator,

$$\hat{s} = \boldsymbol{w} \cdot \boldsymbol{r} + c \tag{4}$$

regardless of the structure of noise correlations.

Another way to test for optimal linear decoding would be to measure whether the animal's behavioral discriminability matches the discriminability for an ideal observer of the neural population response. Yet this approach is not feasible, as it requires one to measure simultaneous responses of many, or even all, relevant neurons. In contrast, the optimality test (Eq 3) requires measuring only single neuron responses, which is vastly easier. Neural recordings in the vestibular system are consistent with optimal decoding according to this prediction [28].

2.6 Nonlinear choice correlations for optimal decoding

However, when nuisance variables wash out the mean tuning of neuronal responses, we may well find that a single neuron has both zero choice correlation and zero information about the stimulus. The optimality test would thus be inconclusive.

This situation is exactly the same one that gives rise to nonlinear codes. A natural generalization of Equation 3 can reveal the quality of neural computation on nonlinear codes. We simply define a '*nonlinear* choice correlation' between the stimulus estimate $\hat{s}$ and nonlinear functions of neural activity $\boldsymbol{R}(\boldsymbol{r})$:

$$C_{R_k} = \text{Corr}(R_k(\boldsymbol{r}), \hat{s}|s) \tag{5}$$

(Methods 4.2), where $R_k(\boldsymbol{r})$ is a nonlinear function of the neural responses. If the brain optimally decodes the information encoded in the nonlinear statistics of neural activity, according to the simple nonlinear extension to Eq 4,

$$\hat{s} = \boldsymbol{w} \cdot \boldsymbol{R}(\boldsymbol{r}) + c \tag{6}$$

then the nonlinear choice correlation satisfies the equation

$$C^{\text{opt}}_{R_k(\boldsymbol{r})} = \sqrt{\frac{J_{R_k(\boldsymbol{r})}}{J}} \tag{7}$$

where $J_{R_k(\boldsymbol{r})}$ is the linear Fisher Information in $R_k(\boldsymbol{r})$ (Methods 4.2.2).

As an example of this relationship, we return to the orientation example. Here the response covariance $\Sigma(s) = \text{Cov}(\boldsymbol{r}|s)$ depends on the stimulus, but the mean $\boldsymbol{f} = \langle \boldsymbol{r}|s \rangle = \langle \boldsymbol{r} \rangle$ does not. In this model, optimally decoded neurons would have no linear correlation with behavioral choice. Instead, the choice should be driven by the product of the neural responses, $\boldsymbol{R}(\boldsymbol{r}) = \text{vec}(\boldsymbol{r}\boldsymbol{r}^\top)$, where vec$(\cdot)$ is a vectorization that flattens an array into a one-dimensional list of numbers. Such quadratic computation is what the energy model for complex cells is thought to accomplish for phase-invariant orientation coding [1]. Figure 3 shows linear and nonlinear choice correlations for pairs of neurons, defined as $C_{r_i r_j} = \text{Corr}(r_i r_j, \hat{s}|s)$. When decoding is linear, linear choice correlations are strong while nonlinear choice correlations are near zero (Figure 3A,B). When the decoding is quadratic, here mediated by an intermediate layer that multiplies pairs of neural activity, the nonlinear choice correlations are strong while the linear ones are insignificant (Figure 3C,D).

2.7 Which nonlinearity?

If the brain's decoder optimally uses all available information, choice correlations will obey the prediction of Eq. 7 even if the specific nonlinearities used by the brain differ from those selected for evaluating choice correlations (Methods 4.2.3). The prediction will hold as long as the brain's nonlinearity can be expressed as a linear combination of the tested nonlinearities (Methods 4.2.3). Figure 4 shows a situation where information is encoded by quadratic and cubic sufficient statistics of neural responses, while a simulated brain decodes them near-optimally using a generic neural network rather than a set of nonlinearities matched to those sufficient statistics. Despite this mismatch we can successfully identify that the brain is near-optimal by applying Eq 7, even without knowing the simulated brain's true nonlinear transformations.

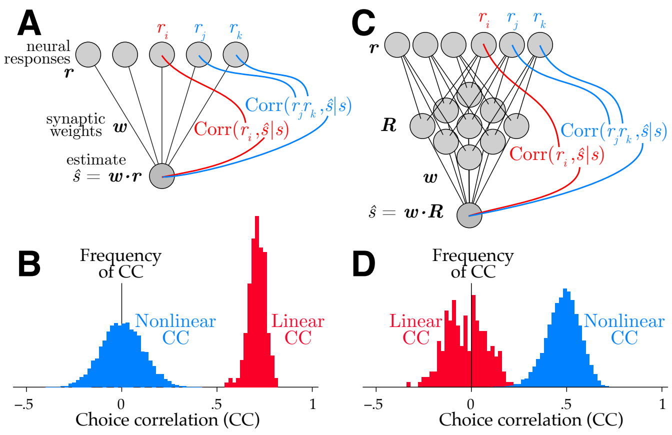

Figure 3: Linear and nonlinear choice correlations successfully distinguish network structure. A linearly decoded population (**A**) produces nonzero linear choice correlations (**B**), while the nonlinear choice correlations are randomly distributed around zero. The situation is reverse for a nonlinear network (**C**), with insignificant linear choice correlations but strong nonlinear ones (**D**). Here the network implements a quadratic nonlinearity, so the relevant choice correlations are quadratic as well, $C_{jk} = \mathrm{Corr}(r_j r_k, c)$.

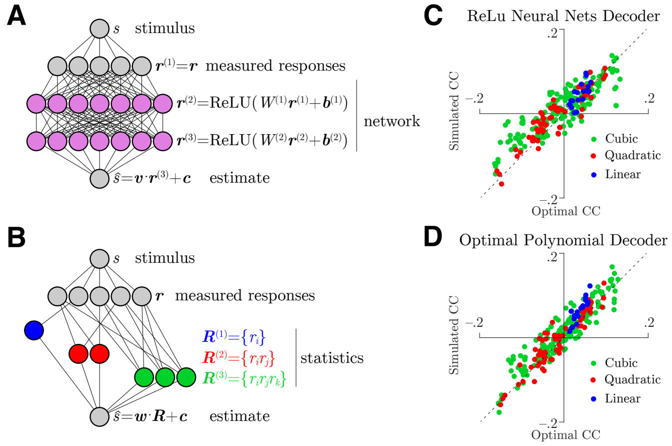
A
s stimulus
$r^{(1)}=r$ measured responses
$r^{(2)}=\mathrm{ReLU}(W^{(1)}r^{(1)}+b^{(1)})$
$r^{(3)}=\mathrm{ReLU}(W^{(2)}r^{(2)}+b^{(2)})$
network
$\hat{s}=v \cdot r^{(3)}+c$ estimate
B
s stimulus
r measured responses
$R^{(1)}=\{r_i\}$
$R^{(2)}=\{r_i r_j\}$
$R^{(3)}=\{r_i r_j r_k\}$
statistics
$\hat{s}=w \cdot R+c$ estimate
C
ReLu Neural Nets Decoder
.2
–.2
.2
–.2
Simulated CC
Optimal CC
Cubic
Quadratic
Linear
D
Optimal Polynomial Decoder
.2
–.2
.2
–.2
Simulated CC
Optimal CC
Cubic
Quadratic
Linear

Figure 4: Identifying optimal nonlinear decoding by a generic neural network using nonlinear choice correlation. **A**: Responses encode stimulus information in polynomial sufficient statistics up to cubic, simulated brain using ReLu nonlinearities trained to extract that information **B**: Simulated brain using matched polynomial nonlinearities to extract the information from responses. **C,D**: Choice correlations with polynomial nonlinear statistics show the network computations are consistent with optimal nonlinear decoding, regardless if the tested statistics match(Network **B**) or not match(Network **A**) the actual decoder(Methods 4.2.3).

2.8 Redundant codes

It might seem unlikely that the brain uses optimal, or even near-optimal, nonlinear decoding. Even if it does, there are an enormous number of high-order statistics for neural responses, so the information content in any one statistic could be tiny compared to the total information in all of them. For example, with N neurons there are on the order of N^2 quadratic statistics, N^3 cubic statistics, and so on. With so many statistics contributing information, the choice correlation for any single one would then be tiny according to the ratio in Eq 7, and would be indistinguishable from zero with reasonable amounts of data. Past theoretical studies have described nonlinear (specifically, quadratic) codes with extensive information that grows proportionally with the number of neurons [16, 30]. This would indeed imply immeasurably small choice correlations for large, optimally decoded populations.

A resolution to these concerns is information-limiting correlations [27]. The past studies that derive extensive nonlinear information treat large cortical populations in isolation from the smaller sensory population that would naturally provide its input [16, 30]. However, when a network inherits information from a much smaller input population, the expanded neural code becomes highly redundant: the brain cannot have more information than it receives. Noise in the input is processed by the same pathway as the signal, and this generates noise correlations that can never be averaged away [27].

Previous work [27] characterized linear information-limiting correlations for fine discrimination tasks by decomposing the noise covariance into $\Sigma = \Sigma_0 + \epsilon \boldsymbol{f}' \boldsymbol{f}'^{\top}$, where ϵ is the variance of the information-limiting component and Σ_0 is noise that can be averaged away with many neurons.

For *nonlinear* population codes, it is not just the mean that encodes the signal, $\boldsymbol{f}(s) = \langle \boldsymbol{r}|s \rangle$, but rather the nonlinear statistics $\boldsymbol{F}(s) = \langle \boldsymbol{R}(\boldsymbol{r})|s \rangle$. Likewise, the noise does not comprise only second-order covariance of $\boldsymbol{r}$, $\mathrm{Cov}(\boldsymbol{r}|s)$, but rather the second-order covariance of the relevant nonlinear statistics, $\Gamma = \mathrm{Cov}(\boldsymbol{R}|s)$ (Section 2.2). Analogous to the linear case, these correlations can be locally decomposed as

$$\Gamma = \mathrm{Cov}(\boldsymbol{R}(\boldsymbol{r})|s) = \Gamma_0 + \epsilon \boldsymbol{F}' \boldsymbol{F}'^{\top} \tag{8}$$

where ϵ is again the variance of the information-limiting component, and Γ_0 is any other covariance which can be averaged away in large populations. The information-limiting noise bounds the estimator variance $\sigma^2_{\hat{s}}$ to no smaller than ϵ even with optimal decoding.

Neither additional neurons nor additional decoded statistics can improve performance beyond this bound. As a direct consequence, when there are many fewer sensory inputs than cortical neurons, many distinct statistics $R_k(\boldsymbol{r})$ will carry redundant information. Under these conditions, many ratios J_{R_k}/J (Eq 7) can be measurably large even for optimal nonlinear decoding (Figure 5).

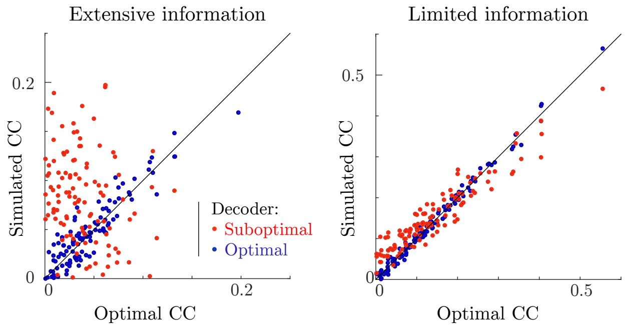

Figure 5: Information-limiting noise makes a network more robust to suboptimal decoding. (Left) A simulated optimal decoder produces choice correlations that match our optimal predictions (blue, on diagonal). In contrast, a suboptimal decoder, such as one that is blind to higher-order correlations ($\boldsymbol{w} \propto \boldsymbol{F}'$), exhibits a suboptimal pattern of choice correlations (red, off-diagonal) in the presence of noise Γ_0 that permits the population to have extensive information. (Right) When information is limited, the same decoding weights are less detrimental, and thus exhibit a similar pattern of choice correlations as an optimal decoder.

2.9 Decoding efficiency revealed by choice correlations

Even if decoding is not strictly optimal, Eq. 7 can be satisfied due to information-limiting correlations. Decoders that seem substantially suboptimal because they fail to avoid the largest noise components in Γ_0 can be nonetheless dominated by the bound from information-limiting correlations. This will occur whenever the variability from suboptimally decoding Γ_0 is smaller than ϵ. Just as we can decompose the nonlinear noise correlations into information-limiting and other parts, we can decompose nonlinear choice correlations into corresponding parts as well, with the result that

$$C_R^{\text{sub}} \approx \alpha C_R^{\text{opt}} + \zeta_R \tag{9}$$

where ζ_R depends on the particular type of suboptimal decoding (Supporting Information S7). The slope α between choice correlations and those predicted from optimality is given by the fraction of estimator variance explained by information-limiting noise, $\alpha = \epsilon/\sigma^2_{\hat{s}}$. This slope therefore provides an estimate of the efficiency of the brain's decoding.

Figure 5 shows an example of a decoder that would be highly suboptimal without considering redundancy, but is nonetheless close to optimal when information limits are inherited.

In realistically redundant models that have more cortical neurons than sensory neurons, many decoders could be near-optimal, as we recently discovered in experimental data for a linear population code [28]. However, even in redundant codes there may be substantial inefficiencies, especially for unnatural tasks [50].

3 Discussion

This study introduced a theory of nonlinear population codes, grounded in the natural computational task of separating relevant and irrelevant variables. The theory considers both encoding and decoding — how stimuli drive neurons, and how neurons drive behavioral choices. It showed how correlated fluctuations between neural activity and behavioral choices could reveal properties of the brain's decoding. Unlike previous theories [16, 30], ours remains consistent with biological constraints due to the large cortical expansion of sensory representations by incorporating redundancy through information-limiting correlations. Crucially, this theory provides a remarkably simple test to determine if downstream nonlinear computation decodes all that is encoded.

Alternative methods to estimate whether animals use their information efficiently rely upon comparing behavioral performance to performance of an ideal observer that can access the entire population. Even with impressive advances in neurotechnology, this challenge remains out of reach for large populations. In contrast, our proposed method to test for optimal decoding has a vastly lower experimental burden. It requires only that a few cells be recorded simultaneously while an animal performs a fine estimation or discrimination task.

On the other hand, this simple test does not offer a complete description of neural transformations.

It instead tests one important hypothesis about their functional role — that the brain performs optimal decoding. The theory also provides a practical way of estimating decoding efficiency. The brain may not be optimal, but instead may be satisfied by a more modest decoding efficiency. In this case, more work is needed to understand what suboptimalities the brain tolerates for satisfactory performance.

3.1 Which nonlinearities should we test?

If all neural signals are decoded optimally, then all choice correlations for any function of those signals should also be consistent with optimal decoding, since they contain the same information. Yet for the wrong or incomplete nonlinearities that do not disentangle the task-relevant variables from the nuisance variables, the test may be inconclusive, just as it was for linear decoding of a nonlinear code (Figure 4): the chosen nonlinear functions may not extract linearly decodable information nor have any choice correlation.

The optimal nonlinearities would be those that collectively extract the sufficient statistics about the relevant stimulus, which will depend on both the task and the nuisance variables. In complex tasks, like recognizing object from images with many nuisance variables, most of the relevant information lives in higher-order statistics, and therefore require more complex nonlinearities to extract. In such high-dimensional cases, our proposed test is unlikely to be useful. This is because our method expresses stimulus estimates as sums of nonlinear functions, and while that is universal in principle [51], that is not a compact way to express the complex nonlinearities of deep networks. Alternatively, with good guidance from trained neural network models our method could potentially judge whether those nonlinearities provide a good description of neural decoding. This decoding perspective would complementing studies that argue for a good match between encoding by convolutional neural networks [6].

The best condition to apply our optimality test is in tasks of modest complexity but still possessing fundamentally nonlinear structure. Some interesting examples where our test could have practical relevance include motion detection using photoreceptors [52], visual search with distractors (XOR-type tasks) [31, 53], sound localization in early auditory processing before the inferior colliculus [54], or context switching in higher-level cortex [55].

Our test for optimal nonlinear decoding really amounts to testing for optimal linear decoding of nonlinear functions of recorded neural data. If we had access to some putative downstream neurons that computed these nonlinear functions, we could just test whether the brain linearly decoded those neurons optimally. Yet that would circumvent the most interesting and crucial nonlinear aspects of neural computation. Alternatively, if we could record from neurons at different levels of the processing chain, we could try to characterize that nonlinear recoding between them directly, without reference to a behavioral choice. But this would not easily relate these computations to their functional role. The method proposed here allows us to skip these intermediate steps and directly test the optimality of all accumulated downstream nonlinearities.

3.2 Nonlinear decoding or switched linear decoding?

Could the brain avoid nonlinear decoding just by switching between different linear decoders depending on the current nuisance variable $\boldsymbol{n}$, so that $\hat{s} = \boldsymbol{w}(\hat{\boldsymbol{n}}) \cdot \boldsymbol{r}$? The switching variable itself would have to be inferred from sensory data, which requires marginalizing over the task variable; this takes us back to the original problem, but with task and nuisance variables reversed. Even so, switched linear decoding would actually be equivalent to nonlinear decoding whenever $\hat{\boldsymbol{n}}$ is estimated from neural responses: $\hat{s} = \boldsymbol{w}(\hat{\boldsymbol{n}}(\boldsymbol{r})) \cdot \boldsymbol{r} = f(\boldsymbol{r})$.

A discrimination task with a changing class boundary [12, 56, 57] is, in principle, a nonlinear task. But if the class boundary is changed too slowly, perhaps changing only on different days, then the brain may well re-learn its weights rather than performing some nonlinear decoding of recent activity. A better experimental design for revealing nonlinear computation for task context would be randomly changing the tasks, either cued [58] or even uncued [55], on a short enough time scale that the recent neural activity affects the class boundary.

3.3 Limitations of the approach

For efficient decoding in a learned task, the optimality test (7) is necessary but not sufficient. If the brain neglects some of informative sufficient statistics, and we don't test these neglected statistics either, then we could find the brain is consistent with our optimal decoding test, yet still be suboptimal. Only if the test is passed for *all* statistics will the test be conclusive. For an extreme example, a single neuron might pass the test, but if other neurons don't, then the brain is not using its information well. On a broader scale, one might find that all individual responses r_k pass the optimality test, while products of responses $r_j r_k$ fail. This would be consistent with linear information being used well while distinct quadratic information is present but unused; on the other hand this outcome would not be consistent with quadratic statistics that are uninformative but decoded anyway, since that would increase the output variance beyond that expected from the linear information. In future work we will demonstrate how we can use nonlinear choice correlations to identify properties of suboptimal decoders [59].

Our approach is currently limited to feedforward processing, which unquestionably oversimplifies cortical processing. Nonetheless, feedforward models do a fair job of capturing the representational structure of the brain [6].

Feedback could also cause suboptimal networks to exhibit choice correlations that seem to resemble the optimal prediction. If the feedback is noisy and projects into the same direction that encodes the stimulus, such as from a dynamic bias [60], then this could appear as information-limiting correlations, enhancing the match with Eq 7. This situation could be disambiguated by measuring the internal noise source providing the feedback, and of course this would require more simultaneous measurements.

3.4 Comparing choice correlations from internal and external noise

Since many stimulus-dependent response correlations are induced by external nuisance variation, not internal noise, we might not find informative stimulus-dependent noise correlations upon repeated presentations of a fixed stimulus. Those correlations may only be informative about a stimulus in the presence of natural nuisance variation. For example, if a picture of a face is shown repeatedly without changing its pose, then small expression changes can readily be identified by linear operations; if the pose can vary then the stimulus is only reflected in higher-order correlations [5].

In contrast, we *should* see some nonlinear choice correlations even when nuisance variables are fixed. This is because neural circuitry must combine responses nonlinearly to eliminate natural nuisance variation, and any internal noise passing through those same channels will thereby influence the choice (although they may be smaller and more difficult to detect than the fluctuations caused by the nuisance variation). This influence will manifest as nonlinear choice correlations. In other words, stimulus-dependent noise correlations need not predict a fixed stimulus, but they may predict the choice (Supplementary Information S8).

For optimal decoding, the choice correlations measured using fixed nuisance variables will differ from Eq 7, which should strictly hold only when there is natural nuisance variation. This is implicit in Eq 7, since the relevant quantities are conditioned only on the relevant stimulus s while averaging over the nuisance variations $\boldsymbol{n}$. However, under some conditions, a related prediction for nonlinear choice correlations holds even without averaging over nuisance variables (Supplementary Information S8).

3.5 Conclusion

Despite the clear importance of computation that is both nonlinear and distributed, and evidence for nonlinear coding in the cortex [31, 33–35], most neuroscience applications of population coding concepts have assumed linear codes and linear readouts [6, 28, 39, 61, 62]. The few that directly address nonlinear population codes either have an impossibly large amount of encoded information [16, 30], or investigate abstract properties unrelated to structured tasks [63]. Some experimental studies have been able to extract additional information from recorded populations using nonlinear decoders [31, 64], but the inferred properties of such decoders are based on recordings being a representative sample that can be extrapolated to larger populations. Unknown correlations and redundancy prevents that from being a reliable method [23, 65].

Our method to understand nonlinear neural decoding requires neural recordings in a behaving animal. The task must be hard enough that it makes some errors, so that there are behavioral fluctuations to explain. Finally, there should be a modest number of nonlinearly entangled nuisance variables. Unfortunately, many neuroscience experiments are designed without explicit use of nuisance variables. Although this simplifies the analysis, this simplification comes at a great cost, which is that the neural circuits are being engaged far from their natural operating point, and far from their purpose: there is little hope of understanding neural computation without challenging the neural systems with nonlinear tasks for which they are required.

Our statistical perspective on feedforward nonlinear coding in the presence of nuisance variables provides a useful framework for thinking about neural computation. Furthermore, choice-related activity provides guidance for designing interesting experiments to measure not only how information is encoded in the brain, but how it is decoded to generate behavior. In future work we aim to apply this theory to experimental data to test whether real brains decode neural information optimally in any nonlinear tasks.

4 Methods

4.1 Encoding models

4.1.1 Orientation estimation with varying spatial phase

Figure 1 illustrates how nuisance variation can eliminate a neuron's mean tuning to relevant stimulus variables, relegating the neural tuning to higher-order statistics like covariances. In this example, the subject estimates the orientation of a Gabor image, $G(\boldsymbol{x}|s,n)$, where $\boldsymbol{x}$ is spatial position in the image, and s and n are the orientation and spatial phase of the image, respectively (Supplemental Material S2). The model visual neurons are linear Gabor filters like idealized simple cells in primary visual cortex, corrupted by additive white Gaussian noise. Their responses are thus distributed as $\boldsymbol{r} \sim P(\boldsymbol{r}|s,n) = N(\boldsymbol{r}|\boldsymbol{f}(s,n),\epsilon I)$, where ϵ is the noise variance and the mean $\boldsymbol{f}(s,n) = \langle \boldsymbol{r}|s,n\rangle = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s,n)$ is determined by the overlap between the image and the receptive field.

When the spatial phase n is known, the mean neural response contains all the information about orientation s. The brain can decode responses linearly to estimate orientation near a reference s_0.

When the spatial phase varies, however, the each mean response to a fixed orientation will be combine across different phases: $\boldsymbol{f}(s) = \langle \boldsymbol{r}|s\rangle = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s) = \int dn \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s,n)p(n)$. Since each spatial phase can be paired with another phase π radians away that inverts the linear response, the phase-averaged mean is $\boldsymbol{f}(s) = 0$. Thus the brain cannot estimate orientation by decoding these neurons linearly; nonlinear computation is necessary.

The covariance provides one such tuned statistic. We define $\mathrm{Cov}_{ij}(\boldsymbol{r}|s,n)$ as the neural covariance for a fixed input image (noise correlations), and $\mathrm{Cov}_{ij}(\boldsymbol{r}|s)$ as the neural covariance when the nuisance varies (nuisance correlations). According to the law of total covariance,

$$\mathrm{Cov}_{ij}(\boldsymbol{r}|s) = \int dn\, (\mathrm{Cov}_{ij}(\boldsymbol{r}|s,n) + \delta f_i(s,n)\delta f_j(s,n))p(n) \quad (10)$$

where $\delta f_i(s,n) = f_i(s,n) - \langle f_i(s,n)\rangle_n$. Supplementary Information S2 shows in detail how $\mathrm{Cov}_{ij}(\boldsymbol{r}|s)$ is tuned to s.

4.1.2 Exponential family distribution and sufficient statistics

We assume the response distribution conditioned on the relevant stimulus (but not on nuisance variables) is approximately a member of the exponential family with nonlinear sufficient statistics,

$$p(\boldsymbol{r}|s) = b(\boldsymbol{r})\exp(\boldsymbol{H}(s)\cdot\boldsymbol{R}(\boldsymbol{r}) - A(s)) \quad (11)$$

where $\boldsymbol{R}(\boldsymbol{r})$ is a vector of sufficient statistics for the natural parameter $\boldsymbol{H}(s)$, $b(\boldsymbol{r})$ is the base measure, and $A(s)$ is the log-partition function. The sufficient statistics contain all of the information in the population response, and all other tuned statistics may be derived from them.

Estimation and inference are closely connected in the exponential family. In Supplemental Material S1.2, we show that the optimal local estimation can be achieved by linearly decoding the nonlinear sufficient statistics, $\hat{s} = \boldsymbol{w}^T\boldsymbol{R}(\boldsymbol{r}) + c$. The decoding weights minimize the variance of an unbiased decoder,

$$\boldsymbol{w}_{\text{opt}} = \frac{\boldsymbol{H}'(s)}{J} = \frac{\Gamma^{-1}\boldsymbol{F}'}{\boldsymbol{F}'^{\top}\Gamma^{-1}\boldsymbol{F}'} \quad (12)$$

where $\boldsymbol{F}' = \partial\langle\boldsymbol{R}(\boldsymbol{r})|s\rangle/\partial s$ is the sensitivity of the statistics to changing inputs, and $\Gamma = \mathrm{Cov}(\boldsymbol{R}|s)$ is the stimulus-conditioned response covariance which generally includes nuisance correlations (Section 2.2).

The variance of this unbiased local estimator from the neural responses is lower-bounded by the inverse Fisher information. For exponential family distributions with nonlinear sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$, the Fisher information is [23] (Supplemental Material S1.1)

$$J = \boldsymbol{F}'^{\top}\Gamma^{-1}\boldsymbol{F}' \quad (13)$$

4.1.3 Quadratic encoding

In a quadratic coding model, the distribution of neural responses is described by the exponential family with up to quadratic sufficient statistics, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_ir_j\}$ for $i,j \in \{1,\ldots,N\}$. A familiar example is the Gaussian distribution with stimulus-dependent covariance $\Sigma(s)$. In order to demonstrate the coding properties of a purely nonlinear neural code, here we assume that the mean tuning curve $f(s)$ and the stimulus-conditional covariances $\Sigma_{ij}(s)$ depend smoothly on the stimulus. We can quantify the information content of the neural population using Equation 13.

4.1.4 Cubic encoding

In our cubic coding model, the distribution of neural responses is described by the exponential family with up to cubic sufficient statistics, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_ir_j, r_ir_jr_k\}$ for $i,j,k \in \{1,\ldots,N\}$.

We approximate a three-neuron cubic code first using purely cubic components, and we then apply a stimulus-dependent affine transformation to include linear and quadratic statistics. The pure cubic code is used for a vector $\boldsymbol{z}$ with sufficient statistics $z_iz_jz_k$ (and a base measure $e^{-\|\boldsymbol{z}\|^4}$ to ensure the distribution is bounded and normalizable).

$$p(\boldsymbol{z}|s) = \tfrac{1}{Z}\exp\left(-\|\boldsymbol{z}\|^4 + \gamma s\, z_iz_jz_k\right) \quad (14)$$

We approximate this distribution by a mixture of four Gaussians. The mixture is chosen to reproduce the tetrahedral symmetry of the cubic distribution (Supplementary Figure 6), which allows the cubic statistics of responses to be stimulus dependent, leaving stimulus-independent quadratic and linear statistics.

To generate larger multivariate cubic codes for Figure (6), for simplicity we assume the pure cubic terms only couple disjoint triplets of variables, and sample independently from an approximately cubic distribution for each triplet. To convert this purely cubic distribution to a distribution with linear and quadratic information, we shift and scale these cubic samples $\boldsymbol{z}$ in a manner dependent on s:

$$\boldsymbol{r} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\boldsymbol{z} \tag{15}$$

where $\boldsymbol{f}(s)$ and $\Sigma(s)$ describes the desired signal-dependent mean and covariance (see Supplemental Material S4).

4.2 Nonlinear choice correlations

4.2.1 Estimating choice correlation

The nonlinear choice correlation between the stimulus estimate $\hat{s} = \boldsymbol{w}^\top \boldsymbol{R} + c$ and one nonlinear function R_k (the kth element of the vector $\boldsymbol{R}$) of recorded neural activity $\boldsymbol{r}$ is

$$C_{R_k} = \mathrm{Corr}(R_k(\boldsymbol{r}), \hat{s}|s) = \frac{(\Gamma \boldsymbol{w})_k}{\sqrt{\Gamma_{kk}\boldsymbol{w}^\top \Gamma \boldsymbol{w}}} \tag{16}$$

where $\boldsymbol{w}^\top \Gamma \boldsymbol{w} = \sigma^2_{\hat{s}}$ is the estimator variance.

To compute this quantity from neural responses to stimuli, we need to condition neural responses and behavior data on the same signal s, or on the same total input $(s, \boldsymbol{n})$ if we want to isolate the contribution of purely internal noise rather than nuisance variation (Supplemental Material S8). We combine choice correlations calculated under different stimulus conditions by balanced z-scoring [66].

4.2.2 Optimality test

Locally optimal linear estimator weights for decoding statistics $\boldsymbol{R}$ are given by linear regression as $\boldsymbol{w} \propto \Gamma^{-1}\boldsymbol{F}'$. Substituting these weights into (16), the optimal nonlinear choice correlation becomes

$$C^{\mathrm{opt}}_{R_k(\boldsymbol{r})} = \frac{\left(\Gamma\Gamma^{-1}\boldsymbol{F}'\right)_k}{\sqrt{\Gamma_{kk}\boldsymbol{F}'^\top \Gamma^{-1}\boldsymbol{F}'}} = \frac{F'_k}{\sqrt{\Gamma_{kk}}}\sigma_{\hat{s}} = \sqrt{\frac{J_{R_k(\boldsymbol{r})}}{J}} \tag{17}$$

where $J_{R_k(\boldsymbol{r})} = F'_k/\sqrt{\Gamma_{kk}}$ is the linear Fisher Information in $R_k(\boldsymbol{r})$.

For fine-scale discriminations, optimal choice correlations can be written in many equivalent ways:

$$C^{\mathrm{opt}}_{R_k} = \frac{d'_{R_k}}{d'} = \frac{\theta}{\theta_{R_k}} = \sqrt{\frac{\sigma^2_{\hat{s}}}{\sigma^2_{\hat{s},R_k}}} = \sqrt{\frac{J_{R_k}}{J}} \tag{18}$$

where $d' = \frac{\Delta \boldsymbol{F}}{\sigma}$ is the discriminability. These ways reflect the simple relationships between four quantities often used to represent information: d-prime is proportional to the square root of the Fisher information $d' = \Delta s\sqrt{J}$ [67]; estimator standard deviation is bounded by the inverse square root of the Fisher information, $\sigma_{\hat{s}} \geq \frac{1}{\sqrt{J}}$; discrimination threshold is proportional to the estimator standard deviation, $\theta = \sqrt{\sigma^2_{\hat{s}}}$. In different experiments (binary discrimination, continuous estimation), it can be most natural to express this relationship in different measured quantities.

In our simulations with binary choices for fine discrimination, we calculate the optimal nonlinear choice correlation using d-prime [68]. d'_{R_k} is estimated from neural responses generated by stimuli $s_\pm = s_0 \pm \Delta s/2$ near a reference stimulus s_0:

$$d'_{R_k} = \frac{\Delta F_k}{\sigma_{R_k}} = \frac{F_k(s_+) - F_k(s_-)}{\sqrt{\frac{1}{2}\left(\sigma^2_{R_k|s_+} + \sigma^2_{R_k|s_-}\right)}} \tag{19}$$

The discriminability for an decoded neural population is estimated from the unbiased decoder output's standard deviation, $d' = 1/\sigma_{\hat{s}_{\mathrm{ref}}}$.

4.2.3 Nonlinear choice correlation to analyze an unknown nonlinearity

In Figure 4, we generated neural responses given sufficient statistics that are polynomials up to third order, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j, r_i r_j r_k\}$ (Methods 4.1.4). In contrast, our model brain decodes the stimulus using a cascade of linear-nonlinear transformations, with Rectified Linear Units ($\mathrm{ReLU}(x) = \max(0, x)$) for the nonlinear activation functions. We used a fully-connected ReLU network with two hidden layers and 30 units per hidden layer,

$$\boldsymbol{r}^{(1)} = \boldsymbol{r} \tag{20}$$

$$\boldsymbol{r}^{(2)} = \mathrm{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{r}^{(1)} + \boldsymbol{b}^{(1)}) \tag{21}$$

$$\boldsymbol{r}^{(3)} = \mathrm{ReLU}(\boldsymbol{W}^{(2)}\boldsymbol{r}^{(2)} + \boldsymbol{b}^{(2)}) \tag{22}$$

$$\hat{s} = \boldsymbol{v} \cdot \boldsymbol{r}^{(3)} + \boldsymbol{b}^{(3)} \tag{23}$$

We trained the network weights and biases with backpropagation to estimate stimuli near a reference s_0 based on 20000 training pairs $(\boldsymbol{r}, s)$ generated by the cubic encoding model. This trained neural network extracted 91% of the information available to an optimal decoder.

4.3 Information-limiting correlations

Only specific correlated fluctuations limit the information content of large neural populations [27]. These fluctuations can ultimately be referred back to the stimulus as $\boldsymbol{r} \sim p(\boldsymbol{r}|s + ds)$, where ds is zero mean noise, whose variance $1/J_\infty$ determines the asymptotic variance of any stimulus estimator. These information-limiting correlations for nonlinear computation can be characterized by the covariance of the sufficient statistics, $\Gamma = \mathrm{Cov}(\boldsymbol{R}|s)$ conditioned on s; the information-limiting component arises specifically from the signal covariance $\mathrm{Cov}(\boldsymbol{F}(s)|s)$. Since the signal for local estimation of stimuli near a reference s_0 is $\boldsymbol{F}'(s) = \frac{d}{ds}\langle \boldsymbol{R}(\boldsymbol{r})|s\rangle$, the information-limiting component of the covariance is proportional to $\boldsymbol{F}'\boldsymbol{F}'^\top$:

$$\Gamma = \Gamma_0 + 1/J_\infty \boldsymbol{F}(s)'\boldsymbol{F}(s)'^\top \tag{24}$$

Here Γ_0 is any covariance of $\boldsymbol{R}$ that does *not* limit information in large populations. Substituting this expression into (13) for the nonlinear Fisher Information, we obtain

$$J = \boldsymbol{F}'\Gamma^{-1}\boldsymbol{F}' = \frac{1}{1/J_\infty + 1/J_0} \tag{25}$$

where $J_0 = \boldsymbol{F}'\Gamma_0^{-1}\boldsymbol{F}'$ is the nonlinear Fisher Information allowed by Γ_0. When the population size grows, the extensive information term J_0 grows proportionally, so the output information will asymptote to J_∞.

Author contributions

XP conceived the theoretical framework. XP and QY performed the theoretical analyses. QY performed the simulations. QY and XP wrote the manuscript.

Acknowledgements

The authors thank Jeff Beck, Valentin Dragoi, Arun Parajuli, Andreas S. Tolias, Edgar Walker, R. James Cotton and Alex Pouget for helpful conversations. This work was supported by NSF CAREER grant 1552868 and NeuroNex grant 1707400 to XP.

References

[1] Adelson EH, Bergen JR (1985) Spatiotemporal energy models for the perception of motion. Josa a 2: 284–299.

[2] DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends in cognitive sciences 11: 333–341.

[3] Rust NC, DiCarlo JJ (2010) Selectivity and tolerance (invariance) both increase as visual information propagates from cortical area v4 to it. Journal of Neuroscience 30: 12978–12995.

[4] Pagan M, Urban LS, Wohl MP, Rust NC (2013) Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. Nature neuroscience 16: 1132.

[5] Meyers EM, Borzello M, Freiwald WA, Tsao D (2015) Intelligent information loss: the coding of facial identity, head pose, and non-face information in the macaque face patch system. Journal of Neuroscience 35: 7069–7081.

[6] Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111: 8619–8624.

[7] Ecker AS, Berens P, Keliris GA, Bethge M, Logothetis NK, Tolias AS (2010) Decorrelated neuronal firing in cortical microcircuits. science 327: 584–587.

[8] Ecker AS, Berens P, Cotton RJ, Subramaniyan M, Denfield GH, Cadwell CR, Smirnakis SM, et al. (2014) State dependence of noise correlations in macaque primary visual cortex. Neuron 82: 235–248.

[9] Denfield GH, Ecker AS, Shinn TJ, Bethge M, Tolias AS (2017) Attentional fluctuations induce shared variability in macaque primary visual cortex. bioRxiv : 189282.

[10] Ecker AS, Denfield GH, Bethge M, Tolias AS (2016) On the structure of neuronal population activity under fluctuations in attentional state. Journal of Neuroscience 36: 1775–1789.

[11] Jazayeri M, Movshon JA (2006) Optimal representation of sensory information by neural populations. Nature neuroscience 9: 690–696.

[12] Bondy AG, Cumming BG (2016) Feedback dynamics determine the structure of spike-count correlation in visual cortex. bioRxiv : 086256.

[13] Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nature reviews neuroscience 7: 358.

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

[15] Kohn A, Coen-Cagli R, Kanitscheider I, Pouget A (2016) Correlations and neuronal population information. Annual review of neuroscience 39.

[16] 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.

[17] Abbott LF, Dayan P (1999) The effect of correlated variability on the accuracy of a population code. Neural computation 11: 91–101.

[18] Cohen MR, Maunsell JH (2009) Attention improves performance primarily by reducing interneuronal correlations. Nature neuroscience 12: 1594.

[19] 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.

[20] Gawne TJ, Richmond BJ (1993) How independent are the messages carried by adjacent inferior temporal cortical neurons? Journal of Neuroscience 13: 2758–2771.

[21] Ralf H, Bethge M (2010) Evaluating neuronal codes for inference using fisher information. In: Advances in neural information processing systems. pp. 1993–2001.

[22] Burge J, Jaini P (2017) Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. PLoS computational biology 13: e1005281.

[23] Beck J, Bejjanki VR, Pouget A (2011) Insights from a simple expression for linear fisher information in a recurrently connected population of spiking neurons. Neural computation 23: 1484–1502.

[24] Paradiso M (1988) A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. Biological cybernetics 58: 35–49.

[25] Zohary E, Shadlen MN, Newsome WT (1994) Correlated neuronal discharge rate and its implications for psychophysical performance. Nature 370: 140–143.

[26] Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. Physical Review E 64: 051904.

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

[28] Pitkow X, Liu S, Angelaki DE, DeAngelis GC, Pouget A (2015) How can single sensory neurons predict behavior? Neuron 87: 411–423.

[29] Shamir M, Sompolinsky H (2004) Nonlinear population codes. Neural computation 16: 1105–1136.

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

[31] Pagan M, Simoncelli EP, Rust NC (2016) Neural quadratic discriminant analysis: Nonlinear decoding with v1-like computation. Neural computation 28: 2291–2319.

[32] Gutnisky DA, Dragoi V (2008) Adaptive coding of visual information in neural populations. Nature 452: 220.

[33] Kohn A, Smith MA (2005) Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. The Journal of neuroscience 25: 3661–3673.

[34] Averbeck BB, Lee D (2006) Effects of noise correlations on information encoding and decoding. Journal of Neurophysiology 95: 3633–3644.

[35] Ohiorhenuan IE, Mechler F, Purpura KP, Schmid AM, Hu Q, Victor JD (2010) Sparse coding and high-order correlations in fine-scale cortical networks. Nature 466: 617.

[36] Ponce-Alvarez A, Thiele A, Albright TD, Stoner GR, Deco G (2013) Stimulus-dependent variability and noise correlations in cortical mt neurons. Proceedings of the National Academy of Sciences 110: 13162–13167.

[37] Rigotti M, Barak O, Warden MR, Wang XJ, Daw ND, Miller EK, Fusi S (2013) The importance of mixed selectivity in complex cognitive tasks. Nature 497: 585–590.

[38] Pagan M, Rust NC (2014) Dynamic target match signals in perirhinal cortex can be explained by instantaneous computations that act on dynamic input from inferotemporal cortex. The Journal of Neuroscience 34: 11067–11084.

[39] Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA (1996) A relationship between behavioral choice and the visual responses of neurons in macaque mt. Visual neuroscience 13: 87–100.

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

[41] Dodd JV, Krug K, Cumming BG, Parker AJ (2001) Perceptually bistable three-dimensional

figures evoke high choice probabilities in cortical area mt. Journal of Neuroscience 21: 4809–4821.

[42] Krajbich I, Armel C, Rangel A (2010) Visual fixations and the computation and comparison of value in simple choice. Nature neuroscience 13: 1292.

[43] de Lafuente V, Romo R (2005) Neuronal correlates of subjective sensory experience. Nature neuroscience 8: 1698.

[44] Treue S, Trujillo JCM (1999) Feature-based attention influences motion processing gain in macaque visual cortex. Nature 399: 575.

[45] Roitman JD, Shadlen MN (2002) Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. Journal of neuroscience 22: 9475–9489.

[46] Gu Y, Angelaki DE, DeAngelis GC (2008) Neural correlates of multisensory cue integration in macaque mstd. Nature neuroscience 11: 1201.

[47] Purushothaman G, Bradley DC (2005) Neural population code for fine perceptual decisions in area mt. Nature neuroscience 8: 99.

[48] 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.

[49] Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA (1996) A relationship between behavioral choice and the visual responses of neurons in macaque mt. Visual Neuroscience 13: 87-100.

[50] Nienborg H, Cumming BG (2007) Psychophysically measured task strategy for disparity discrimination is reflected in v2 neurons. Nature neuroscience 10: 1608.

[51] Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural networks 4: 251–257.

[52] Poggio T, Koch C (1987) Synapses that compute motion. Scientific American 256: 46–53.

[53] Ma WJ, Navalpakkam V, Beck JM, Van Den Berg R, Pouget A (2011) Behavior and neural basis of near-optimal visual search. Nature neuroscience 14: 783.

[54] Davis KA, Ramachandran R, May BJ (2003) Auditory processing of spectral cues for sound localization in the inferior colliculus. Journal of the Association for Research in Otolaryngology 4: 148–163.

[55] Saez A, Rigotti M, Ostojic S, Fusi S, Salzman C (2015) Abstract context representations in primate amygdala and prefrontal cortex. Neuron 87: 869–881.

[56] Cohen MR, Newsome WT (2008) Context-dependent changes in functional circuitry in visual area mt. Neuron 60: 162–173.

[57] Lange RD, Haefner RM (2016) Inferring the brain9s internal model from sensory responses in a probabilistic inference framework. bioRxiv : 081661.

[58] Mante V, Sussillo D, Shenoy KV, Newsome WT (2013) Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503: 78-84.

[59] Yang Q, Pitkow X (2015) Robust nonlinear neural codes. Cosyne abstract .

[60] Haefner RM, Berkes P, Fiser J (2016) Perceptual decision-making as probabilistic inference by neural sampling. Neuron 90: 649–660.

[61] Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nature neuroscience 9: 1432.

[62] Graf AB, Kohn A, Jazayeri M, Movshon JA (2011) Decoding the activity of neuronal populations in macaque primary visual cortex. Nature neuroscience 14: 239.

[63] Babadi B, Sompolinsky H (2014) Sparseness and expansion in sensory representations. Neuron 83: 1213–1226.

[64] Maynard E, Hatsopoulos N, Ojakangas C, Acuna B, Sanes J, Normann R, Donoghue J (1999) Neuronal interactions improve cortical population coding of movement direction. Journal of Neuroscience 19: 8083–8093.

[65] Kanitscheider I, Coen-Cagli R, Kohn A, Pouget A (2015) Measuring fisher information accurately in correlated neural populations. PLoS computational biology 11: e1004218.

[66] 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.

[67] Berens P, Ecker AS, Gerwinn S, Tolias AS, Bethge M (2011) Reassessing optimal neural population codes with neurometric functions. Proceedings of the National Academy of Sciences 108: 4423–4428.

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

[69] Bethge M, Rotermund D, Pawelzik K (2002) Optimal short-term population coding: when fisher information fails. Neural computation 14: 2317–2351.

Supplemental material

S1 Exponential family distributions

For a stimulus s and a response $\boldsymbol{r}$, the conditional probability is a member of the exponential family when

$$p(\boldsymbol{r}|s) = b(\boldsymbol{r}) \exp\Big(\boldsymbol{H}(s)^\top \boldsymbol{R}(\boldsymbol{r}) - A(s)\Big) \tag{26}$$

where $\boldsymbol{H}(s)$ are the natural parameters, $\boldsymbol{R}(\boldsymbol{r})$ are the sufficient statistics, $A(s)$ and $b(\boldsymbol{r})$ are the log normalizer and base measure. The statistics $\boldsymbol{R}(\boldsymbol{r})$ are called sufficient because they contain all the information needed to estimate the stimulus s.

S1.1 Fisher information

One measure of information content that a population response contains about a stimulus is the Fisher information $J(s)$ [16, 24–27, 29]. The Fisher information is given by

$$\begin{aligned} J &= -\left\langle \frac{\partial^2}{\partial s^2} \log p(\boldsymbol{r}|s) \right\rangle_{p(\boldsymbol{r}|s)} && (27)\\ &= \left\langle \left(\frac{\partial}{\partial s} \log p(\boldsymbol{r}|s)\right)^2 \right\rangle_{p(\boldsymbol{r}|s)} && (28) \end{aligned}$$

For distributions $p(\boldsymbol{r}|s)$ in the exponential family with sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$, we can compute these quantities analytically. We denote the mean of the sufficient statistics as $\boldsymbol{F}(s) = \langle \boldsymbol{R}(\boldsymbol{r})|s\rangle$. This mean $\langle \boldsymbol{R}|s\rangle$ can be obtained by differentiating $A(s)$ by the natural parameters $\boldsymbol{H}(s)$,

$$\boldsymbol{F} = \frac{\partial A(s)}{\partial \boldsymbol{H}(s)} \tag{29}$$

Equation 29 can give us the first and second derivatives of $A(s)$ over s.

$$\begin{aligned} A' &= \sum_i \frac{\partial A}{\partial H_i}\frac{dH_i}{ds} = \boldsymbol{H}'^\top \boldsymbol{F} && (30)\\ A'' &= \boldsymbol{H}''^\top \boldsymbol{F} + \boldsymbol{H}'^\top \boldsymbol{F}' && (31) \end{aligned}$$

Thus we can compute two definitions of Fisher information.

$$\begin{aligned} J &= -\left\langle \frac{\partial^2}{\partial s^2} \log P(\boldsymbol{r}|s) \right\rangle_{P(\boldsymbol{r}|s)} && (32)\\ &= A'' - \boldsymbol{H}''^\top \boldsymbol{F} && (33)\\ &= \boldsymbol{H}'^\top \boldsymbol{F}' && (34) \end{aligned}$$

and

$$\begin{aligned} J &= \left\langle \left(\frac{\partial}{\partial s} \log P(\boldsymbol{r}|s)\right)^2 \right\rangle_{P(\boldsymbol{r}|s)} && (35)\\ &= \boldsymbol{H}'^\top (\langle \boldsymbol{R}\boldsymbol{R}^\top\rangle - \boldsymbol{F}\boldsymbol{F}^\top)\boldsymbol{H}' && (36)\\ &= \boldsymbol{H}'^\top \Gamma \boldsymbol{H}' && (37) \end{aligned}$$

where $\Gamma = \mathrm{Cov}[\boldsymbol{R}(\boldsymbol{r})|s]$.

Since the two definition are equivalent, we have

$$\boldsymbol{H}' = \Gamma^{-1}\boldsymbol{F}' \tag{38}$$

Substituting Equation 38 into Equation 37, we find the Fisher Information for the exponential family [23]

$$J = \boldsymbol{F}'^\top \Gamma^{-1} \boldsymbol{F}' \tag{39}$$

S1.2 Estimation in the exponential family

Again assuming responses come from this distribution, we want to compute the maximum likelihood stimulus, $\hat{s}$, near a reference stimulus s_0:

$$\begin{aligned} \hat{s} &= \underset{s}{\mathrm{argmax}}\ p(\boldsymbol{r}|s) && (40)\\ &= \underset{s}{\mathrm{argmax}}\ \log p(\boldsymbol{r}|s) && (41)\\ &= \underset{s}{\mathrm{argmax}}\ \boldsymbol{H}(s)^\top \boldsymbol{R}(\boldsymbol{r}) - A(s) && (42) \end{aligned}$$

A Taylor expansion around the reference yields

$$\begin{aligned} &\boldsymbol{H}(s)^\top \boldsymbol{R}(\boldsymbol{r}) - A(s)\\ &\quad\approx [\boldsymbol{H}^\top \boldsymbol{R} - A]\\ &\quad\ +[\boldsymbol{H}'^\top \boldsymbol{R} - A'](s - s_0)\\ &\quad\ +\tfrac{1}{2}(s-s_0)^\top[\boldsymbol{H}''^\top \boldsymbol{R} - A''](s - s_0) + \cdots \end{aligned} \tag{43}$$

where all functions and derivatives are evaluated at s_0. We find the maximum by differentiating with respect to s and setting the result equal to zero:

$$0 = [\boldsymbol{H}'^\top \boldsymbol{R} - A'] + (s - s_0)[\boldsymbol{H}''^\top \boldsymbol{R} - A''] \tag{44}$$

The solution is

$$s = s_0 - \frac{\boldsymbol{H}'^\top \boldsymbol{R} - A'}{\boldsymbol{H}''^\top \boldsymbol{R} - A''} \tag{45}$$

Since $\boldsymbol{r}$ is a random quantity, we can express $\boldsymbol{R}$ as a mean and a deviation away from that mean: $\boldsymbol{R} = \langle \boldsymbol{R}|s_0\rangle + \delta\boldsymbol{R} = \boldsymbol{F} + \delta\boldsymbol{R}$. In this case, $\boldsymbol{H}''^\top \boldsymbol{R} - A'' = \boldsymbol{H}''^\top \boldsymbol{F} - A'' + \boldsymbol{H}''^\top \delta\boldsymbol{R}$, where the mean term is precisely the negative Fisher Information $-J(s_0)$. If the trial-to-trial fluctuations in the uncertainty are small relative to the average uncertainty then this Fisher term will dominate. Then we have

$$s = \boldsymbol{w}^\top \boldsymbol{R} + \boldsymbol{c} \tag{46}$$

where

$$\boldsymbol{w} = \frac{\boldsymbol{H}'}{J} = \frac{\Gamma^{-1}\boldsymbol{F}'}{\boldsymbol{F}'^\top \Gamma^{-1}\boldsymbol{F}'} \tag{47}$$

and where we used the results from Equations 13 and 38, with $\Gamma = \mathrm{Cov}(\boldsymbol{R}|s_0)$ and $\boldsymbol{F} = \langle \boldsymbol{R}|s_0\rangle$. Thus, in this limit, the optimal estimator for s is a linear decoding of the sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$.

S2 Orientation estimation task with varying spatial phase

In Figure 2B, the subject's task is to estimate orientation s near a reference s_0, based on images G of Gabor patterns given by

$$G(\boldsymbol{x}|s,n) = e^{-\|\boldsymbol{x}\|^2}\cos\left(\boldsymbol{k}\cdot\boldsymbol{x} + n\right) \tag{48}$$

where $\boldsymbol{k} = \kappa(\cos s, \sin s)$. Here the target s is the orientation of the pattern, $\boldsymbol{n}$ is a nuisance variable reflecting the spatial phase, $\boldsymbol{x}$ is the pixel location in the image, and $\boldsymbol{k}$ is a spatial frequency vector with amplitude $\kappa = \|\boldsymbol{k}\|$. We assume the spatial receptive field of simple cell j in primary visual cortex is also described by a Gabor function

$$\mathrm{RF}_j(\boldsymbol{x}, s_j, n_j) = e^{-\|\boldsymbol{x}\|^2}\cos\left(\boldsymbol{k}_j\cdot\boldsymbol{x} + n_j\right) \tag{49}$$

$$\boldsymbol{k}_j = \kappa(\cos s_j, \sin s_j) \tag{50}$$

where each neuron has a preferred orientation s_j, spatial phase n_j, and spatial frequency $\boldsymbol{k}_j$. Here for simplicity we assume that all neurons' preferred spatial frequencies have the same amplitude κ that matches the input image.

We model the mean neuronal responses by the overlap between the image and their linear receptive field. This overlap determines the tuning curve of each neuron:

$$\begin{aligned} f_j(s,n) &= \int d\boldsymbol{x}\, G(\boldsymbol{x}|s,n)\mathrm{RF}_j(\boldsymbol{x}, s_j, n_j) \\ &= \Big[e^{-\frac{1}{4}\kappa^2\cos(s-s_j)}\cos\left(n+n_j\right) \\ &\quad + e^{+\frac{1}{4}\kappa^2\cos(s-s_j)}\cos\left(n-n_j\right)\Big]\,\frac{\pi}{4}e^{-\frac{1}{4}\kappa^2} \end{aligned} \tag{51}$$

This expression can be written in the form:

$$f_j(s,n) = A_j(s)\cos\left(n + \psi_j(s)\right) \tag{52}$$

using the stimulus-dependent response amplitude

$$A_j(s) = C\sqrt{2\cosh 2\beta_j(s) + 2\cos 2n_j} \tag{53}$$

and phase

$$\psi_j(s) = n_j - \alpha_j(s) \tag{54}$$

where we define the constants

$$C = \frac{\pi}{4}\exp\left(-\frac{1}{4}\kappa^2\right) \tag{55}$$

$$\beta_j(s) = \frac{1}{4}\kappa^2\cos(s - s_j) \tag{56}$$

$$\alpha_j(s) = \tan^{-1}\frac{\exp\left(\beta_j(s)\right)\sin 2n_j}{\exp\left(-\beta_j(s)\right) + \exp\left(\beta_j(s)\right)\cos 2n_j} \tag{57}$$

Equation 52 reveals that the mean response of each neuron traces out a sinusoidal oscillation in n, where the amplitude and phase depend on s and the specific neuron j. The mean tuning for each pair of neurons therefore traces out an ellipse as a function of the nuisance variable, the input's spatial phase. When we *average* over the ellipse generated by the nuisance variable n, the mean tuning to s is abolished — but the response *covariances* (nuisance correlations) remain tuned to s.

Assuming each neuron's response variability is drawn independently from a standard Gaussian $\mathcal{N}(0,1)$, we can write the response distribution as

$$P(\boldsymbol{r}|n,s) = \mathcal{N}(\boldsymbol{f}(s,n), \boldsymbol{I}) \tag{58}$$

If the spatial phase n were fixed and known, the brain could estimate the orientation just from the mean tuning of the neural responses. However, if the spatial phase is unknown and varies between stimulus presentations uniformly from 0 to 2π, the mean tuning $\boldsymbol{f}(s)$ can be expressed as

$$\begin{aligned} f_j(s) = \langle r_j|s\rangle &= \int r_j\, p(r_j|s)dr_j && (59)\\ &= \iint r_j\, p(r_j|s,n)\,p(n)\,dr_j\,dn && (60)\\ &= \int f_j(s,n)p(n)\,dn && (61)\\ &= \frac{1}{2\pi}\int f_j(s,n)\,dn && (62)\\ &= \frac{A_j(s)}{2\pi}\int_0^{2\pi}\cos\left(n+\psi_j(s)\right)dn = 0 && (63) \end{aligned}$$

This shows that there is no signal in the mean responses.

However, the brain can perform quadratic computations to eliminate the nuisance variable. We can define $\mathrm{Cov}_{ij}[\boldsymbol{r}|s,n]$ as the neural covariance (noise correlations) when everything in the image is fixed, and $\mathrm{Cov}_{ij}[\boldsymbol{r}|s]$ as the neural covariance when the nuisance is unknown and free to vary (nuisance correlations). Then $\mathrm{Cov}_{ij}[\boldsymbol{r}|s]$ is

$$\begin{aligned}
\mathrm{Cov}_{ij}[\boldsymbol{r}|s] &= \langle (r_i - f_i(s))(r_j - f_j(s))|s\rangle && (64)\\
&= \langle r_i r_j|s\rangle = \iint r_i r_j\, p(\boldsymbol{r}|s) dr_i dr_j && (65)\\
&= \int dn \iint r_i r_j\, p(\boldsymbol{r}|s,n)p(n) dr_i dr_j && (66)\\
&= \int dn\, p(n)\langle r_i r_j|s,n\rangle && (67)\\
&= \int dn\, p(n)\left(\mathrm{Cov}_{ij}[\boldsymbol{r}|s,n] + f_i(s,n)f_j(s,n)\right) && (68)\\
&= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi}\int dn f_i(s,n)f_j(s,n) && (69)\\
&= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi}D_{ij}(s) && (70)
\end{aligned}$$

where $D_{ij}(s)$ is given by

$$\begin{aligned}
D_{ij}(s) &= \int dn f_i(s,n)f_j(s,n)\\
&= \int dn\, A_i(s)\cos(n+\psi_i(s))A_j(s)\cos(n+\psi_j(s))\\
&= \pi\cos(\psi_i(s)-\psi_j(s))A_i(s)A_j(s)
\end{aligned} \qquad (71)$$

Here when we compute Equation 71, we used the trigonometric identity: $2\cos(x)\cos(y) = \cos(x+y) + \cos(x-y)$, and $\int\cos(2n+\psi_i+\psi_j)dn = 0$.

This demonstrates that the neural covariance $\mathrm{Cov}_{ij}[\boldsymbol{r}|s]$ depends on the orientation s. While linear computation is useless for estimating orientation since the mean responses are untuned (59), quadratic (or higher-order) nonlinear computations can be used to estimate the orientation.

S3 Quadratic coding model

In a purely quadratic coding model (no linear information), the distribution of neural responses is described by the exponential family with quadratic sufficient statistics, $p(\boldsymbol{r}|s) \sim \exp\left[\boldsymbol{H}(s)^\top\boldsymbol{R}(\boldsymbol{r})\right]$ where $\boldsymbol{R}(\boldsymbol{r}) = (\ldots, r_i r_j, \ldots)$. A familiar example is a Gaussian distribution with stimulus-dependent covariance: $p(\boldsymbol{r}|s) = N(\boldsymbol{f}, \Sigma(s))$.

As a concrete example we construct a covariance that rotates with stimulus s. Any covariance matrix needs to be positive semidefinite. We build $\Sigma(s)$ by setting the eigenvalues to be positive and s-independent and eigenvectors to form an orthogonal basis that rotates with s:

$$\Sigma(s) = V(s)\Lambda V(s)^\top \qquad (72)$$

where $V(s) = \exp As$ is a rotation matrix in which $A = -A^\top$ is a real antisymmetric matrix with pure imaginary eigenvalues, and Λ is a diagonal matrix composed of all positive eigenvalues.

To calculate the Fisher Information (Equation 13), we need to first calculate the derivative of the mean $\boldsymbol{F}' = \frac{\partial}{\partial s}\langle\boldsymbol{R}(\boldsymbol{r})|s\rangle$ and covariance $\Gamma = \mathrm{Cov}[\boldsymbol{R}(\boldsymbol{r})|s]$ of the quadratic sufficient statistics.

Because the mean of $\boldsymbol{r}$ is not dependent on the stimulus in this example, we can compute $F'_{ij} = \langle r_i r_j|s\rangle' = \Sigma'_{ij}(s)$, where $\Sigma'_{ij}(s)$ is the derivative of the covariance of $\boldsymbol{r}$,

$$\Sigma'(s) = Ue^{\Omega s}(\Omega X - X\Omega)e^{-\Omega s}U^\top \qquad (73)$$

Here Ω is a diagonal matrix of eigenvalues for A, U is an orthogonal matrix of the eigenvectors of A, and $X = U^\top\Lambda U$.

The elements in Γ can be expressed as $\Gamma_{ij,kn} = \langle r_i r_j r_k r_n|s\rangle - \langle r_i r_j|s\rangle\langle r_k r_n|s\rangle$. We can use the following identity for a Gaussian to compute this fourth-order quantity:

$$\begin{aligned}
\langle r_i r_j r_k r_n|s\rangle = &\langle r_i r_j|s\rangle\langle r_k r_n|s\rangle + \langle r_j r_n|s\rangle\langle r_i r_n|s\rangle\\
&+ \langle r_i r_n|s\rangle\langle r_j r_k|s\rangle
\end{aligned} \qquad (74)$$

where

$$\langle r_i r_j|s\rangle = \Sigma_{ij} + f_i f_j \qquad (75)$$

Substitution of the response covariance (Equation 72) into Equation 74 allows us to calculate the covariance Γ of the quadratic sufficient statistics, and thereby to estimate the stimulus and Fisher information for this quadratic code.

S4 Cubic codes

In Figure 6 we assume the brain encodes the stimulus using a cubic code. A simple cubic code in

$\boldsymbol{z} = (z_i, z_j, z_k) \in \mathbb{R}^3$ can be written as

$$p(\boldsymbol{z}|s) = \tfrac{1}{Z} \exp\left(\gamma(s) z_i z_j z_k - |\boldsymbol{z}|^4\right) \tag{76}$$

where we include the base measure $e^{-|\boldsymbol{z}|^4}$ to ensure normalizability (Figure 6A).

For mathematical convenience, we approximate this code by a mixture of Gaussians.

$$p(\boldsymbol{z}|s) \approx \sum_{a=1}^{4} p(a) p(\boldsymbol{z}|a) \tag{77}$$

$$= \sum_a \frac{1}{4} \mathcal{N}(\boldsymbol{z}|\mu_a, \Sigma_a) \tag{78}$$

where

$$\mu_a = \frac{s}{\sqrt{1+s^2}} \boldsymbol{v}_a \tag{79}$$

and

$$\Sigma_a = \frac{1}{(1+s^2)^2} (I + s^2 \boldsymbol{v}_a \boldsymbol{v}_a^\top) \tag{80}$$

The vectors $\boldsymbol{v}_a$ reflect the four corners of the tetrahedron, $v_{a,i} = \pm 1$, to match the tetrahedral symmetry of the pure cubic code (Equation 76, Figure 6). To sample from this distribution, we randomly choose a component a and then sample from the gaussian $\mathcal{N}(\boldsymbol{z}|\mu_a, \Sigma_a)$ conditioned on that component.

This distribution has zero mean and identity covariance but a nontrivial skewness tensor, and qualitatively matches the corresponding distribution for the true exponential family distribution with cubic sufficient statistics (Figure 6).

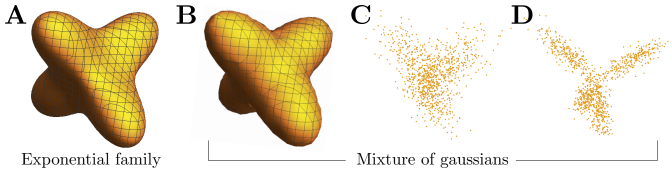

Figure 6: Multivariate skewed distributions. (**A**) Isoprobability contour of an exponential family distribution with cubic statistics in three dimensions, drawn from $p(\boldsymbol{z}|s) \propto \exp\left(s z_1 z_2 z_3 - \|\boldsymbol{z}\|^4\right)$. (**B**) Isoprobability contour for a mixture of four gaussians (Eq 78). (**C,D**) Samples drawn from the mixture form, with $s = 1, 2$.

For simplicity, we consider pure cubic codes with non-overlapping cliques of three variables.

$$p(\boldsymbol{z}|s) = \prod_\alpha p(\boldsymbol{z}_\alpha|s) = \prod_\alpha p(z_{\alpha_1}, z_{\alpha_2}, z_{\alpha_3}|s) \tag{81}$$

To convert this purely cubic distribution into a distribution with linear and quadratic information as well, we simply shift and scale the distribution in a manner dependent on s:

$$\boldsymbol{r} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\, \boldsymbol{z}\, \Sigma^{1/2}(s) \tag{82}$$

$$\boldsymbol{z} \sim \frac{1}{Z(s)} \exp\left[\sum_{ijk} \gamma_{ijk}(s) z_i z_j z_k - |\boldsymbol{z}|^4\right] \tag{83}$$

These affine transformations can be incorporated directly into each component of the mixture of gaussians,

$$p(\boldsymbol{r}|a) = \mathcal{N}(\boldsymbol{r}|\boldsymbol{f}(s) + \boldsymbol{m}_a(s), \Sigma^{1/2}(s) S_a(s) \Sigma^{1/2}(s) \tag{84}$$

Note that the linear and quadratic information terms are independent of the component $\boldsymbol{a}$.

S5 Using nonlinear choice correlation to analyze unknown nonlinearities

The true nonlinearity that the brain uses to estimate the stimulus is unknown. Thus a crucial question in our decoding analysis is, which nonlinearities to consider? One reasonable set is polynomials in $\boldsymbol{r}$, *i.e.* a Taylor series expansion of the neural nonlinearities, $\boldsymbol{\Psi}(\boldsymbol{r}) = (r_i, r_i r_j, r_i r_j r_k, \ldots)$.

The locally optimal decoder is a weighted sum of the sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$ (Equation 46):

$$\hat{s}_{\text{opt}} = \boldsymbol{w} \cdot \boldsymbol{R}(\boldsymbol{r}). \tag{85}$$

However, the brain might choose a different nonlinear basis $\boldsymbol{g}(\boldsymbol{r})$:

$$\hat{s}_{\text{brain}} = \boldsymbol{v} \cdot \boldsymbol{g}(\boldsymbol{r}). \tag{86}$$

As long as the brain's nonlinear function spans the same function basis as the sufficient statistics, we can still get all of the information about stimulus from neural population. This allows us to use choice correlation between brain's estimate $\hat{s}_{\text{brain}}$ and our analysis nonlinearity $\Psi(\boldsymbol{r})$ to check the optimality condition (Equation 7).

In Figure 4, we assumed that the optimal nonlinear basis function $\boldsymbol{R}$ is polynomial nonlinearity up to third order, $\boldsymbol{R}(\boldsymbol{r}) = (r_i, r_i r_j, r_i r_j r_k, \ldots)$. We used cubic codes described in Methods 4.1.4 to generate neural responses for which $\boldsymbol{R}(\boldsymbol{r})$ are sufficient statistics for the stimulus. In this simulation, 18 neuronal responses (six cliques of size 3) were generated using cubic codes.

Our model brain decodes the stimulus using a cascade of linear-nonlinear transformations, with Rectified

Linear Units ($\mathrm{ReLU}(x) = \max(0, x)$) for the nonlinear activation functions. We used a fully-connected ReLU network with two hidden layers and 30 units per hidden layer,

$$\hat{s}_{\text{brain}} = \boldsymbol{v} \cdot \boldsymbol{r}^{(3)} + \boldsymbol{b}^{(3)} \tag{87}$$

$$\boldsymbol{r}^{(3)} = \mathrm{ReLU}(\boldsymbol{W}^{(2)}\boldsymbol{r}^{(2)} + \boldsymbol{b}^{(2)}) \tag{88}$$

$$\boldsymbol{r}^{(2)} = \mathrm{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{r}^{(1)} + \boldsymbol{b}^{(1)}) \tag{89}$$

$$\boldsymbol{r}^{(1)} = \boldsymbol{r} \tag{90}$$

We trained the neural network with 20000 response samples generated from a cubic code driven by stimuli near the reference s_0. We optimized the estimation performance for the neural network using backpropagation to find weights $\{\boldsymbol{W}^{(\ell)}\}$, biases $\{\boldsymbol{b}^{(\ell)}\}$, and readout vector $\boldsymbol{v}$ that minimized the mean squared error. Our trained neural network performed near-optimally, extracting 91% of the Fisher information compared to optimal decoding based on the true sufficient statistics.

Feigning ignorance of our simulated brain's true decoder, we used mononomial nonlinearities $\Psi(\boldsymbol{r})$ in our the nonlinear choice correlation test (Equation 7). The simulated choice correlations were calculated by Equation 5, where $\boldsymbol{R}(\boldsymbol{r}) = \Psi(\boldsymbol{r})$ based on neural responses driven by the reference stimulus s_0, and the stimulus estimate was $\hat{s}_{\text{brain}}$. The optimal choice correlation is computed using Equation 7, where $\sqrt{J_{\Psi(\boldsymbol{r})}} = d'_\Psi/\Delta s = \frac{\Delta \boldsymbol{F}_\Psi}{\Delta s\,\sigma_\Psi}$, and $\sqrt{J} \approx 1/\sigma_{\hat{s}_{\text{brain}}}$. We computed $\Delta\boldsymbol{F}_\Psi$ based on neural population responses $\boldsymbol{r}_+$ and $\boldsymbol{r}_-$ driven by stimuli $s_+ = s_0 \pm \Delta s/2$. The change in mean was $\Delta\boldsymbol{F}_\Psi = \langle\Psi(\boldsymbol{r}_+)\rangle - \langle\Psi(\boldsymbol{r}_-)\rangle$, and the average standard deviation was $\sigma_\Psi = \sqrt{\frac{1}{2}\mathrm{Var}(\Psi(\boldsymbol{r}_+)) + \frac{1}{2}\mathrm{Var}(\Psi(\boldsymbol{r}_-))}$. $\sigma^2_{\hat{s}_{\text{brain}}}$ is the variance of estimate of reference stimulus s_0 using the trained neural network. Based on these quantities, Figure 4 shows that we can successfully identify that the brain is near-optimal.

S6 Information-limiting correlations

Information-limiting correlations can ultimately be referred back to the stimulus, to appear as $\boldsymbol{r} \sim p(\boldsymbol{r}|s + ds)$, where ds is zero mean noise with variance $1/J_\infty$ which determines the uncertainty of stimulus. Applying the law of total covariance, we can decompose the covariance of nonlinear statistics $\boldsymbol{R}(\boldsymbol{r})$ conditioned on the stimulus into two parts:

$$\begin{aligned}\Gamma =& \mathrm{Cov}_{\boldsymbol{r},ds}(\boldsymbol{R}(\boldsymbol{r})|s) \\ =& \left\langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s, ds)\right\rangle_{ds} + \mathrm{Cov}_{ds}\left\langle \boldsymbol{R}(\boldsymbol{r})|s, ds\right\rangle_{\boldsymbol{r}}\end{aligned} \tag{91}$$

where $\langle\cdot\rangle_p$ indicates an expectation value over the distribution p. The first term can be computed as follows,

$$\left\langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s, ds)\right\rangle_{ds} = \left\langle \Gamma(s + ds)\right\rangle_{ds} \tag{92}$$

$$\approx \left\langle \Gamma_0 + ds\,\Gamma'\right\rangle_{ds} \tag{93}$$

$$= \Gamma_0 \tag{94}$$

Here we denote the covariance of $\boldsymbol{R}(\boldsymbol{r})$ given s and ds as $\Gamma(s + ds)$. The second equality used a Taylor expansion of $\Gamma(s + ds)$ around s. The third equality used the fact that the mean of ds is zero. Γ_0 is the covariance of $\boldsymbol{R}$ in the absence of information-limiting correlations. The second term in Equation 91 can be expressed as

$$\mathrm{Cov}_{ds}\left\langle \boldsymbol{R}(\boldsymbol{r})|s, ds\right\rangle_{\boldsymbol{r}} \tag{95}$$

$$= \mathrm{Cov}_{ds}(\boldsymbol{F}(s + ds)) \tag{96}$$

$$\approx \mathrm{Cov}_{ds}(\boldsymbol{F}(s) + ds\,\boldsymbol{F}'(s)) \tag{97}$$

$$= \frac{1}{J_\infty}\boldsymbol{F}'(s)\boldsymbol{F}'(s)^\top \tag{98}$$

Here we have written the mean of $\boldsymbol{R}(\boldsymbol{r})$ given s and ds as $\boldsymbol{F}(s + ds)$. The second equality used a first-order expansion of $\boldsymbol{F}(s + ds)$ around s. The third equality used the fact that the variance of ds is $1/J_\infty$.

Equation 91 can therefore be written as

$$\Gamma = \Gamma_0 + \frac{1}{J_\infty}\boldsymbol{F}(s)'\boldsymbol{F}(s)'^\top \tag{99}$$

which is a rank-one perturbation of the covariance Γ_0.

To compute the nonlinear Fisher Information, $J_{R(\boldsymbol{r})} = \boldsymbol{F}'^\top\Gamma^{-1}\boldsymbol{F}'$, we can use the Sherman-Morrison lemma to compute Γ^{-1}:

$$\Gamma^{-1} = \Gamma_0^{-1} - \frac{\Gamma_0^{-1}\boldsymbol{F}'\boldsymbol{F}'^\top\Gamma_0^{-1}}{J_\infty + \boldsymbol{F}'\Gamma_0^{-1}\boldsymbol{F}'^\top} \tag{100}$$

Substituting these equations into the nonlinear Fisher Information (Equation 13) and simplifying, we obtain

$$J_{R(\boldsymbol{r})} = \frac{1}{1/J_\infty + 1/J_0} \tag{101}$$

Here $J_0 = \boldsymbol{F}'^\top\Gamma_0^{-1}\boldsymbol{F}'$ is the nonlinear Fisher Information in the absence of information-limiting correlations. When the population size grows, the term J_0 grows proportionally [16,29], so for large populations the output information saturates at J_∞.

S7 Nonlinear choice correlation for suboptimal decoding

A decoder that would be suboptimal for one population code could be near-optimal in the presence of information-limiting noise. In this case, nonlinear choice correlations can be decomposed into a sum of two terms, one from the information-limiting component and the other from the rest of the noise [28]:

$$C_{R_k} = \frac{(\Gamma \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} = \frac{(\Gamma_0 \boldsymbol{w} + \frac{1}{J_\infty}\boldsymbol{F}'\boldsymbol{F}'^{\top}\boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} \tag{102}$$

For unbiased decoding, $\boldsymbol{w}^\top \boldsymbol{F}' = 1$. Some manipulation gives

$$C_{R_k} = \frac{(\Gamma_0 \boldsymbol{w})_k}{\Gamma_{0k}\sigma_{0\hat{s}}} \frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} \frac{\Gamma_{0k}}{\Gamma_k} + \frac{F'_k}{\sigma_k}\sigma_{\hat{s}} \frac{1/J_\infty}{\sigma^2_{\hat{s}}} \tag{103}$$

where $\Gamma_{0k} = (\Gamma_0)_{kk} \approx \Gamma_{kk}$ for small information-limiting noise variance $1/J_\infty \ll \Gamma_{0k}$ (which nonetheless can have 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 covariance of the sufficient statistics is Γ_0. The variance of $\hat{s}$ can itself be decomposed into two terms as well:

$$\begin{aligned}\sigma^2_{\hat{s}} &= \boldsymbol{w}^\top \Gamma \boldsymbol{w} = \boldsymbol{w}^\top \Gamma \boldsymbol{w} + \frac{1}{J_\infty}\boldsymbol{w}^\top \boldsymbol{F}'\boldsymbol{F}'^{\top}\boldsymbol{w} \\ &= \sigma^2_{0\hat{s}} + 1/J_\infty\end{aligned} \tag{104}$$

where we assume unbiased decoding, which implies $\boldsymbol{w}^\top \boldsymbol{F}' = 1$. This expression allows us to represent the ratio $\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}}$ as

$$\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} = \sqrt{1 - \frac{1/J_\infty}{\sigma^2_{\hat{s}}}} = \sqrt{1-\alpha} \tag{105}$$

with $\alpha = \frac{1/J_\infty}{\sigma^2_{\hat{s}}}$. Substituting these into (Eq 103) we find that the choice correlation for a suboptimal decoder in the presence of information-limiting correlations is a weighted sum of the choice correlations for optimal and suboptimal decoding:

$$C_R^{\text{sub}} \approx \alpha C_R^{\text{opt}} + C_R^{\text{sub}}\sqrt{1-\alpha} \tag{106}$$

Here C_R^{sub} and C_R^{opt} are, respectively, the choice correlations for suboptimal decoding without information-limiting noise (so $\Gamma = \Gamma_0$), and choice correlations for optimal decoding.

The slope α between choice correlations and those predicted from optimal decoding is equal to the fraction of estimator variance explained by information-limiting noise. This slope therefore provides an estimate of the efficiency of the brain's decoding.

S8 Comparing choice correlations from internal or external noise

The response covariance that drives fluctuations in choices could arise from internal or external (nuisance) variability, or both. Choice correlations predicted for optimal decoding differ depending on whether we condition on the nuisance variables or not. In the main text, we described optimal choice correlations under the distribution $p(\boldsymbol{r}|s)$. This includes variations caused by external nuisance variables, which is sensible since this is what the brain's decoder must handle. However, it is also potentially informative to examine how purely internal variability correlates with choice, as this is often how choice correlations are assessed. In this section, we derive the choice correlations driven by purely internal noise, for a decoder that learned to remove external nuisance variation as well.

For simplicity we assume that the nonlinear sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$ are linearly tuned to both the stimulus s and a scalar nuisance variable n,

$$\boldsymbol{R}(\boldsymbol{r}) = \boldsymbol{F}'s + \boldsymbol{G}'n + \boldsymbol{\eta} \tag{107}$$

where $\boldsymbol{F}'$ and $\boldsymbol{G}'$ characterize the sensitivity of $\boldsymbol{R}(\boldsymbol{r})$ to stimulus s and nuisance n, and an internal noise source $\boldsymbol{\eta}$ has zero mean with covariance H. We assume the brain has a prior over the nuisance variation, $p(n)$, with zero mean and variance ξ. The total covariance for internal and external fluctuations is then

$$\Gamma = H + \xi \boldsymbol{G}'\boldsymbol{G}'^{\top} \tag{108}$$

When we measure choice correlations while fixing the nuisance variables in the experiment, we assume the brain retains its decoding strategy accounting for both internal noise and unknown nuisance variation, and not the optimal decoding strategy when the nuisance is fixed and known. These decoding weights are

$$\boldsymbol{w} = \frac{\Gamma^{-1}\boldsymbol{F}'}{J_1} \tag{109}$$

where the denominator $J_1 = \boldsymbol{F}'^{\top}\Gamma^{-1}\boldsymbol{F}'$ is the Fisher information about s when there is natural nuisance variation following $p(n)$. For distributions in the exponential family, this information saturates the Cramer-Rao bound on an estimator's variance, so that $J_1 = 1/\sigma^2_{\hat{s}}$. [69] The normalization by J_1 ensures the decoding is locally unbiased. These weights are used to estimate the stimulus according to

$$\hat{s} = \boldsymbol{w}^\top \boldsymbol{R}(\boldsymbol{r}) + b \tag{110}$$

Choice correlations in this fixed-nuisance experiment will be denoted by a lowercase c:

$$c^{\rm sub}_{R_k} = \mathrm{Corr}(R_k, \hat{s}|s, n) \tag{111}$$

We include the superscript $c^{\rm sub}$ as a reminder that these choice correlations do not follow the optimal pattern when the decoder is not matched to only the purely internal variability, as here.

We can express these choice correlations as:

$$c^{\rm sub}_{R_k} = \frac{\mathrm{Cov}(R_k, \hat{s}|s, n)}{\sigma_{R_k|s,n}\sigma_{\hat{s}|s,n}} \tag{112}$$

The covariance between $\hat{s}$ and $\boldsymbol{R}$ is

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{R}, \hat{s}|s, n) &= \langle \boldsymbol{R}\hat{s}|s, n\rangle \qquad (113)\\ &= \langle \boldsymbol{R}\boldsymbol{R}^\top|s, n\rangle \boldsymbol{w} \qquad (114)\\ &= \frac{H\Gamma^{-1}\boldsymbol{F}'}{J_1} \qquad (115)\end{aligned}$$

For the scalar nuisance variable we assume here, we can use the Sherman-Morrison lemma to decompose the inverse of the total covariance into a rank-one perturbation of the internal noise inverse covariance:

$$\begin{aligned} \Gamma^{-1} &= (H + \xi\boldsymbol{G}'\boldsymbol{G}'^\top)^{-1} \qquad (116)\\ &= H^{-1} - \frac{H^{-1}\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}}{1/\xi + \boldsymbol{G}'^\top H^{-1}\boldsymbol{G}'} \qquad (117)\end{aligned}$$

Substituting this inverse covariance into Equation 113, we obtain

$$\begin{aligned} &\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s, n) \qquad (118)\\ &= \frac{1}{J_1} H(H^{-1} - \frac{H^{-1}\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'^\top})\boldsymbol{F}' \qquad (119)\\ &= \frac{1}{J_1}(\boldsymbol{F}' - \frac{\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}\boldsymbol{F}'}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'^\top}) \qquad (120)\end{aligned}$$

This last expression can be rewritten using elements of the Fisher information matrix, whose inverse bounds the covariance of any joint estimator of the signal and nuisance variables, $(\hat{s}, \hat{n})$:

$$\boldsymbol{J}(s, n) = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22}\end{bmatrix} = \begin{bmatrix} \boldsymbol{F}'^\top H^{-1}\boldsymbol{F}' & \boldsymbol{F}'^\top H^{-1}\boldsymbol{G}' \\ \boldsymbol{G}'^\top H^{-1}\boldsymbol{F}' & \boldsymbol{G}'^\top H^{-1}\boldsymbol{G}'\end{bmatrix} \tag{121}$$

With these substitutions, we have

$$\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s, n) = \frac{1}{J_1}\left(\boldsymbol{F}' - \frac{J_{12}}{1/\xi + J_{22}}\boldsymbol{G}'\right) \tag{122}$$

The denominator of Equation 112 involves the variance of the sufficient statistics,

$$\sigma^2_{R_k|s,n} = H_{kk} \tag{123}$$

and the variance of the brain's decoder,

$$\begin{aligned} \sigma^2_{\hat{s}} &= \boldsymbol{w}^\top H\boldsymbol{w} \qquad (124)\\ &= \boldsymbol{w}^\top(\Gamma - \xi\boldsymbol{G}'\boldsymbol{G}'^\top)\boldsymbol{w} \qquad (125)\\ &= \frac{1}{J_1} - \frac{J_{12}^2}{\xi J_1^2}\frac{1}{(1/\xi + J_{22})^2} \qquad (126)\end{aligned}$$

where we used the following results:

$$\begin{aligned} \boldsymbol{w}^\top\boldsymbol{G}'\boldsymbol{G}'^\top\boldsymbol{w} &= \left(\frac{\boldsymbol{F}'\Gamma^{-1}}{J_1}\boldsymbol{G}'\right)^2 \qquad (127)\\ &= \frac{1}{J_1^2}\left(\boldsymbol{F}' H^{-1}\boldsymbol{G}' - \frac{\boldsymbol{F}'\Gamma^{-1}\boldsymbol{G}'\boldsymbol{G}' H^{-1}\boldsymbol{G}'}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'}\right)^2 \qquad (128)\\ &= \frac{1}{J_1^2}\left(J_{12} - \frac{J_{12}J_{22}}{1/\xi + J_{22}}\right)^2 \qquad (129)\\ &= \frac{J_{12}^2}{\xi^2 J_1^2}\frac{1}{(1/\xi + J_{22})^2} \qquad (130)\end{aligned}$$

Combining the results from Equation 122, 126 and 123, we can compute Equation 112

$$\begin{aligned} c^{\rm sub}_{R_k} &= \mathrm{Corr}(R_k, \hat{s}|s, n) \qquad (131)\\ &= \frac{\mathrm{Cov}(R_k, \hat{s}|s, n)}{\sigma_{R_k|s,n}\sigma_{\hat{s}|s,n}} \qquad (132)\\ &= \frac{\frac{1}{J_1}\left(F'_k - \frac{J_{12}}{1/\xi + J_{22}}G'_k\right)}{\sqrt{H_{kk}}\sigma_{\hat{s}|s,n}} \qquad (133)\end{aligned}$$

The optimal choice correlation when there is natural nuisance variation (Eq 7) is given by

$$C^{\rm opt}_{R_k} = \sqrt{\frac{J_{1,R_k}}{J_1}} = \frac{F'_k}{\sigma_{R_k|s}\sqrt{J_1}} \tag{134}$$

where $J_{1,R_k} = F'_k/\sigma_{R_k|s}$ is the Fisher Information in R_k about s when there is natural nuisance variation, and $\sigma_{R_k|s} = \sqrt{H_{kk} + \xi G'^2_k}$ is the standard deviation of the statistic R_k, again when there is natural nuisance variation.

The choice correlations for the same decoder differ under experimental conditions with and without nuisance variation: $C^{\text{opt}}_{R_k}$ and $c^{\text{sub}}_{R_k}$. We find that the nuisance-conditioned choice correlations $c^{\text{sub}}_{R_k}$ relate to the optimal nuisance-averaged choice correlations $C^{\text{opt}}_{R_k}$ according to

$$c^{\text{sub}}_{R_k} = \beta_k C^{\text{opt}}_{R_k} - \gamma_k \tag{135}$$

where we have defined the following constants:

$$\beta_k = \frac{\sigma_{R_k|s}}{\sigma_{R_k|s,n}} \frac{1}{\sqrt{J_1}\sigma_{\hat{s}|s,n}} \tag{136}$$

$$= \sqrt{\frac{H_{kk} + \xi G'^2_k}{H_{kk}}} \frac{1}{\sqrt{J_1}\sigma_{\hat{s}|s,n}} \tag{137}$$

$$= \sqrt{\frac{H_{kk} + \xi G'^2_k}{H_{kk}}} \frac{1}{\sqrt{1 - \frac{J_{12}^2}{\xi J_1}\frac{1}{(1/\xi + J_{22})^2}}} \tag{138}$$

and

$$\gamma_k = \frac{G'_k}{\sqrt{H_{kk}}} \frac{J_{12}}{(1/\xi + J_2) J_1 \sigma_{\hat{s}|s,n}} \tag{139}$$

The slope β_k and offset γ_k of the relationship between these two types of choice correlations (Equation 135) depends on the amount of nuisance variation compared to internal noise and the suboptimality of the brain's decoding strategy. When the signal and nuisance can be disentangled, that is, estimated nearly independently using the statistics $\boldsymbol{R}(\boldsymbol{r})$, then J_{12} is small and the choice correlations driven purely by internal fluctuations closely match the optimal choice correlations in the presence of nuisance variation (Figure 7A). In contrast, when nuisance variations remain partialy confused with the signal, then J_{12} is large and the choice correlations for fixed nuisance variables may differ from the optimal pattern seen when allowing nuisance variables to change from trial to trial (Figure 7B).

For the simulations in Figure 7, we set the sufficient statistics to be linear $\boldsymbol{R}(\boldsymbol{r}) = \boldsymbol{r}$ for simplicity. Neural responses were generated from a Gaussian distribution with a stimulus-dependent mean and identity covariance $H = I$: $p(\boldsymbol{r}|s,n) = \mathcal{N}(\boldsymbol{F}'s + \boldsymbol{G}'n, I)$. In Figure 7A, $\boldsymbol{F}'$ and $\boldsymbol{G}'$ are set to be orthogonal to ensure $J_{12} = \boldsymbol{F}'^{\top} H^{-1} \boldsymbol{G}' = 0$. They are picked from the eigenvector of a symmetric matrix $A^{\top}A$, where A is a matrix whose elements are generated from uniform distribution bounded by 0 and 1. In Figure 7B, each element in $\boldsymbol{F}'$ and $\boldsymbol{G}'$ is drawn from a uniform distribution over the interval $[0, 1]$. We simulate 10000 responses of a population with $N = 50$ neurons. The stimulus is set to 0 and the nuisance is fixed to be 1. The brain's decoder assumes a Gaussian prior over the nuisance variation with zero mean and variance $\xi = 2$. The decoding weights follow Equation 109, and the stimulus is estimated using Equation 110. Choice correlations in this fixed-nuisance experiment are computed by Equation 111 (vertical axis in Figure 7). The predicted optimal choice correlation is computed by Equation 134 (horizontal axis in Figure 7). In this setting, $\beta_k \approx 1$ when $J_{12} = 0$.

In this context, it is especially noteworthy that a mismatch between choice correlations and the optimal pattern might not indicate that the brain is suboptimal, but instead that the experimental task may not match the natural tasks for which the brain could have been optimized.

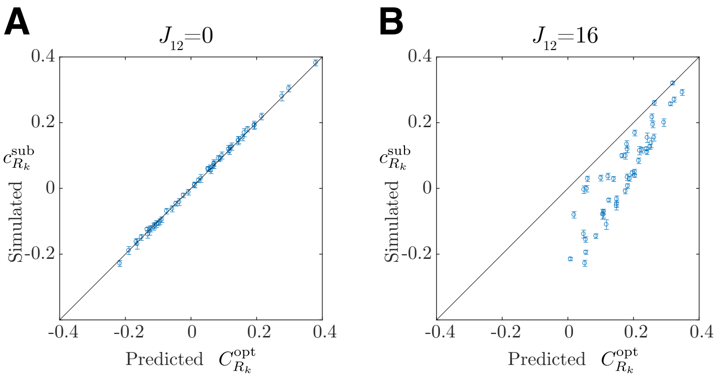

Figure 7: Comparing choice correlations caused by internal and external noise. (**A**) When estimates of nuisance variables are independent of estimates of task-relevant signals, the optimal choice correlations driven by internal noise, $c^{\text{sub}}_{R_k}$, match the optimal pattern $C^{\text{opt}}_{R_k}$ expected for optimal decoding under natural nuisance variation (Equation 7). (**B**) When the signal and nuisance variables remain confounded by an estimator and decoding is evaluated under different conditions than those for which it was optimized, then the choice correlations need not match this optimal prediction.