Revealing nonlinear neural decoding by analyzing choices

Qianli Yang[1], Edgar Walker[2,3], R. James Cotton[2,3], Andreas S. Tolias[1,2,3], 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 entangled with task-irrelevant nuisance variables. The neurons that encode these relevant signals typically constitute a nonlinear population code. Here we present a theoretical framework for quantifying how the brain uses or decodes its nonlinear information. Our theory obeys fundamental mathematical limitations on information content inherited from the sensory periphery, identifying redundant codes when there are many more cortical neurons than primary sensory neurons. The theory predicts that if the brain uses its nonlinear population codes optimally, then more informative patterns should be more correlated with choices. More specifically, the theory predicts a simple, easily computed quantitative relationship between fluctuating neural activity and behavioral choices that reveals the decoding efficiency. We analyze recordings from primary visual cortex of monkeys discriminating the distribution from which oriented stimuli were drawn, and find these data are consistent with the hypothesis of near-optimal nonlinear decoding.

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 in complex ways. As we show below, when these nuisance variations have nonlinear effects on responses, they can dilute or even abolish the mean tuning to the relevant stimulus. Then the brain cannot simply use linear computation, nor can we understand neural processing using linear models.

A quantitative account of nonlinear neural decoding of sensory stimuli must first express how populations of neurons encode or represent information. Past theories of nonlinear population codes made unsupported assumptions about the variability of these population responses [1, 2], substantially underestimating the redundancy of large cortical populations that are driven by a smaller population of sensory inputs. Here we correct this problem by generalizing information-limiting correlations [3] to nonlinear population codes, providing a more realistic theory of how much sensory information is encoded in the brain.

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, whereas *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. These are distinct processes, so the brain could encode a stimulus well while decoding it poorly, or vice-versa.

We provides a simple way of testing the hypothesis that the brain's decoding strategy is efficient, using a simple, novel statistic to assess whether neural response patterns that are informative about the task-relevant sensory input are also informative about the animal's behavior in the task. This extends past results on linear decoding [4] by relaxing multiple assumptions that are violated when neuronal response statistics beyond

*Correspondence: xaq@rice.edu

the mean are tuned.

These key theoretical advances are developed in general, and applied to simple models to illustrate the approach. Finally, we apply this test to analyze V1 data from macaque monkeys, finding direct experimental evidence for optimal nonlinear decoding.

2 Results

2.1 Framework for nonlinear computation

2.1.1 A simple example of a nonlinear code

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.

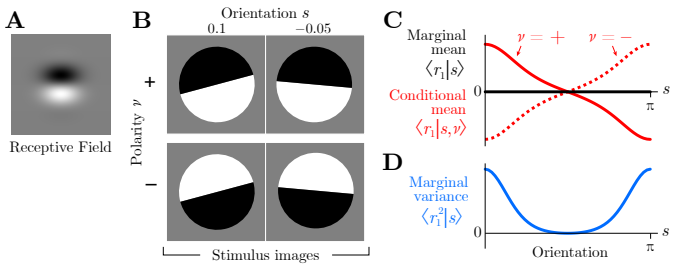

Figure 1: Simple nonlinear code for orientation induced by two polarities. (**A**) Receptive field for a linear neuron. (**B**) Four example images, each with an orientation $s \in [0, \pi)$ and a polarity $\nu \in \{-1, +1\}$. (**C**) The mean response of the linear neuron is tuned to orientation if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

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 evoking the largest response to one polarity will also evoke 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 [5].

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. Generically, nuisance variables make it hard to extract the task-relevant variables from sense data, which is the central task of perception [6–9]. For example, cells in early visual cortex are not tuned to object identity, since the object could appear at any location and V1 has not yet extracted the complex combinations of features that reveal object type independent of the nuisance variable of position. (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 extracts the variables of interest [6, 7, 10]. But can we understand *which* nonlinear function is useful, beyond simply stating what works? A statistical perspective helps here: the brain learns from its history of sensory inputs which statistics of its many sense data are tuned to the task-relevant variable. Good nonlinear computations then compute those statistics. In the orientation estimation task above, the relevant statistic was not the mean, but the variance.

The next sections formalize and develop these ideas to describe redundant nonlinear population codes.

2.1.2 Task, stimuli, neural responses, actions

Our mathematical framework describes a perceptual task, a stimulus with both relevant and irrelevant variables, neural responses, and behavioral choices.

In our task, an agent observes a multidimensional stimulus (s, ν) and must act upon one particular relevant aspect of that stimulus, s, while ignoring the rest, ν. The irrelevant stimulus aspects serve as nuisance variables for the task (ν is the Greek letter 'nu' and here stands for *nu*isance). 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 $p(\boldsymbol{r}|s, \nu)$.

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 $\hat{s}$:

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

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. 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 a subset of $\boldsymbol{r}$. 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 $\boldsymbol{r}$, s, and $\hat{s}$ to estimate the quality of nonlinear decoding.

We first develop the theory for local or fine-scale estimation tasks: the subject must directly report its estimate $\hat{s}$ for the relevant stimuli near a reference s_0, and we measure performance by the variance of this estimate, $\sigma^2_{\hat{s}}$. This fine-scale continuous estimation provides the simplest mathematical framing of the problem. In later sections we then generalize the problem to allow for binary discrimination as well as coarse tasks. These binary and coarse discriminations are not conceptually different from fine estimation, but the mathematical details are more complicated.

2.1.3 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 $p(\boldsymbol{r}|s, \nu)$ when everything about the stimulus is fixed. This could also include uncontrolled variation in internal states [11–14], like attention, motivation, or wandering thoughts. External noise is variability generated by the external world, or nuisance variables. Whether this should count as 'noise' is somewhat contentious. In some instances, most people readily describe external variation as noise, as for a 'white noise stimulus' or a random dot kinematogram. In other cases people might be more reticent to label this variability as noise, as for the uncontrolled polarity of an edge (Figure 1) or the lighting of a three-dimensional scene. Regardless of the name, external variability 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 [2, 15–22]. Signal correlations measure shared variation in mean responses $\boldsymbol{f}(s)$ averaged over the set of stimuli s: $\rho_{\text{signal}} = \text{Corr}(\boldsymbol{f}(s))$. (Internal) noise correlations measure shared variation that persists even when the stimulus is completely identical, nuisance variables and all: $\rho_{\text{noise}}(s, \nu) = \text{Corr}(\boldsymbol{r}|s, \nu)$.

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 ν: $\rho_{\text{nuisance}}(s) = \text{Corr}(\boldsymbol{f}(\nu)|s)$. Including both internal and external (nuisance) noise correlations gives $\text{Corr}\,(\boldsymbol{r}|s)$.

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). Whether from internal or external noise, stimulus-dependent correlations lead naturally to nonlinear population codes, as we explain below.

2.1.4 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 [3,24–27]. 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 S.1.2.2) is a linear function of the response. Linear codes for different variables may arise early in sensory processing, like orientation in V1, or after many stages of computation [6,9], like for objects in inferotemporal cortex.

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 [1,2]. These functional and statistical views are naturally linked because estimating higher-order statistics requires nonlinear operations. For instance, information from a stimulus-dependent covariance $Q(s) = \langle \boldsymbol{r}\boldsymbol{r}^\top | s \rangle$ can be decoded by quadratic operations $\boldsymbol{R} = \boldsymbol{r}\boldsymbol{r}^\top$ [28–30]. Table 1 compares the relevant neural response properties for linear and nonlinear codes.

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

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 (such as the quadratic example in the last column). 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).

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 $s = \mathrm{XOR}(r_1, r_2)$ (Figure 2A). We don't care about the specific value of r_1 by itself, and in fact r_1 alone tells us nothing about s. The same is true for r_2. The usual view on nonlinear computation is that the desired signal can be extracted by applying an XOR or product nonlinearity. However, there is an underlying statistical reason this works: 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 [31–35]. Other experiments have turned up evidence for nonlinear population codes by characterizing the nonlinear selectivity directly [10,36, 37].

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 polarity 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 $\langle \boldsymbol{r}|s \rangle$ either, 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 second moments, $\boldsymbol{r}\boldsymbol{r}^\top$ (Table 1). In this simple example, the nuisance variable of spatial phase ensures that quadratic statistics contains relevant information about the orientation, just like complex cells in V1 [5].

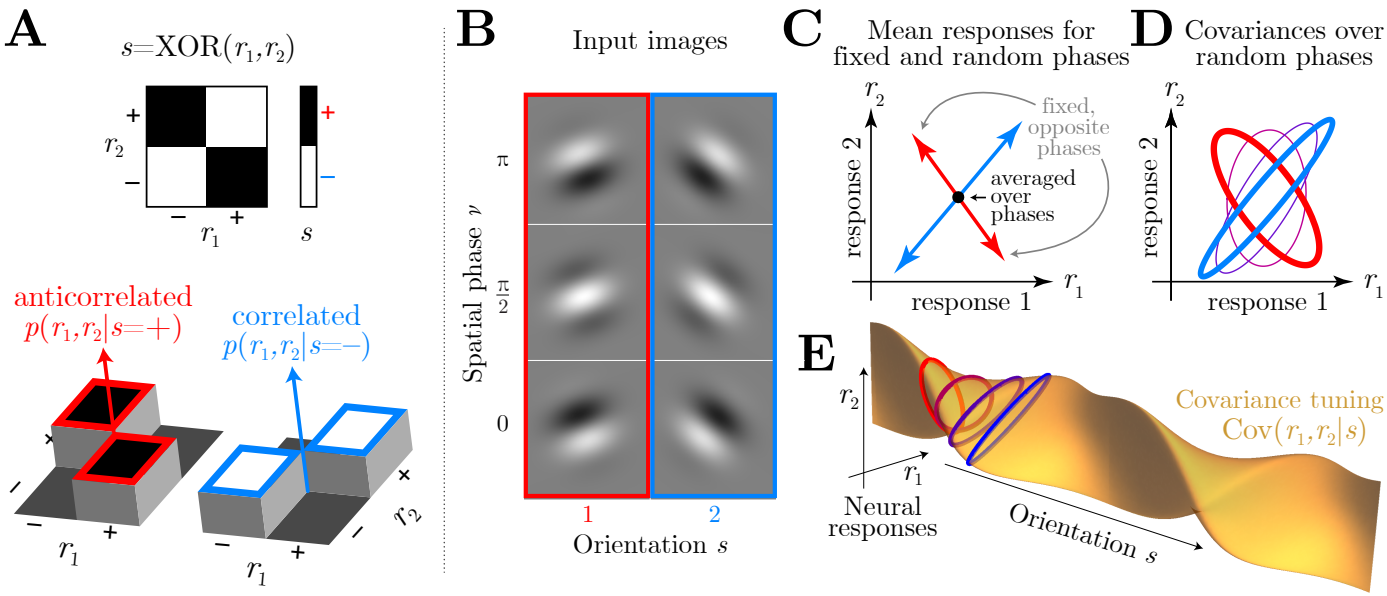

Figure 2: Nonlinear codes. (**A**) Simple example in which a stimulus s is the XOR of two neural responses (top). Conditional probabilities $p(r_1, r_2|s)$ of those responses (bottom) show they are anti-correlated when $s = +1$ (red) and positively correlated when $s = -1$ (blue). This stimulus-dependent 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 $\mathrm{Cov}(r_1, r_2|s)$ 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.

2.2 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 [38–46]. 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 [47].

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, we would struggle to acquire enough trials to fully disambiguate 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 [27]. Next we review these predictions, and then generalize them to nonlinear codes.

2.2.1 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:

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

This choice correlation is a conceptually simpler and more convenient measure than the more conventional statistic, 'choice probability' [48], but it has almost identical properties (Methods 4.2) [27, 47].

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 [27]:

$$C^{\mathrm{opt}}_{r_k} = \frac{d'_{r_k}}{d'} \tag{3}$$

where d' and d'_{r_k} are, respectively, the stimulus discriminability [49] 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,

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

for any stimulus-independent noise correlations, regardless of their structure.

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 non-simultaneous single neuron responses, which is vastly easier. Neural recordings in the vestibular system are consistent with near-optimal decoding according to this prediction [27].

2.2.2 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 $\boldsymbol{R}(\boldsymbol{r})$:

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

(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

$$C^{\mathrm{opt}}_{R_k(\boldsymbol{r})} = \frac{d'_{R_k(\boldsymbol{r})}}{d'} \tag{7}$$

where $d'_{R_k(\boldsymbol{r})}$ is the stimulus discriminability provided by $R_k(\boldsymbol{r})$ (Methods 4.2.1).

As an example of this relationship, we return to the orientation example. Here the response covariance $\Sigma(s) = \mathrm{Cov}(\boldsymbol{r}|s)$ 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, $\boldsymbol{R}(\boldsymbol{r}) = \mathrm{vec}(\boldsymbol{r}\boldsymbol{r}^\top)$, where $\mathrm{vec}(\cdot)$ 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 [5]. Figure 3 shows linear and nonlinear choice correlations for pairs of neurons, defined as $C_{r_i r_j} = \mathrm{Corr}(r_i r_j, \hat{s}|s)$. 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).

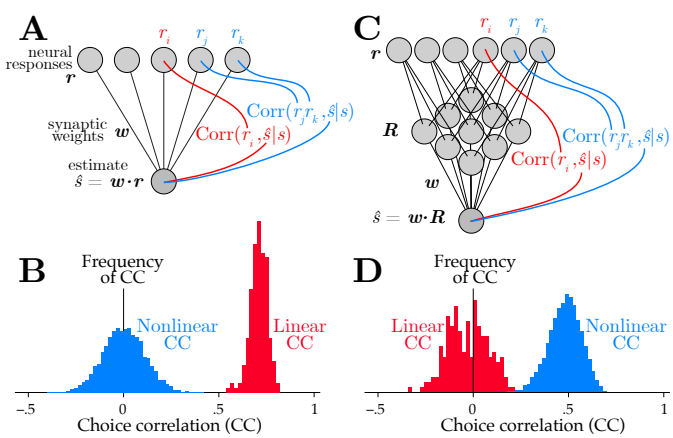

Figure 3: Linear and nonlinear choice correlations successfully distinguish network structure. A linearly decoded population (**A**) produces nonzero linear choice correlations (**B**), while the nonlinear choice correlations are randomly distributed around zero. The situation is reverse for a nonlinear 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, $C_{jk} = \mathrm{Corr}(r_j r_k, \hat{s}|s)$.

2.2.3 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 4.2.2). The prediction is valid as long as the brain's nonlinearity can be expressed as a linear combination of the tested nonlinearities (Methods 4.2.2). Figure 4 shows a situation where information is encoded by linear, 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 details of the simulated brain's true nonlinear transformations.

2.2.4 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 [2, 28]. This would indeed imply immeasurably small choice correlations for large, optimally decoded populations.

A resolution to these concerns is information-limiting correlations [3]. 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 [2, 28]. Yet 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 [50]. Noise in the input is processed by the same pathway as the signal, and this generates noise correlations that can never be averaged away [3].

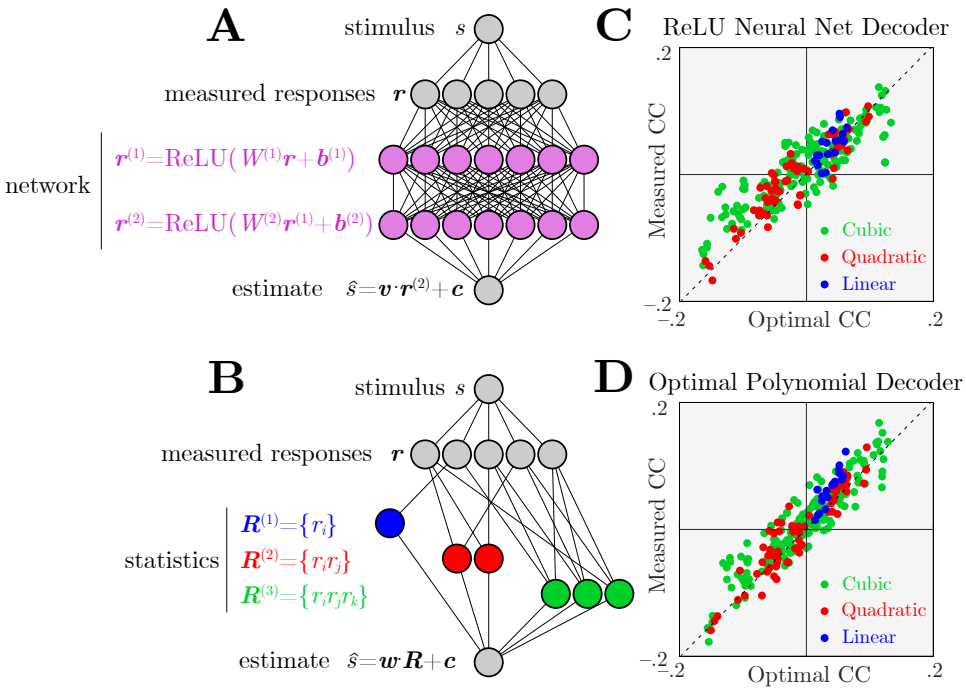

Figure 4: Identifying optimal nonlinear decoding by a generic neural network using nonlinear choice correlations. Neural responses $\boldsymbol{r}$ are constructed to encode stimulus information in polynomial sufficient statistics up to cubic order (Methods Eq. 13). These responses are decoded by an artificial nonlinear neural network or polynomial nonlinearities, and we evaluate the quality of the decoding using polynomial nonlinearities for both cases. (**A**) Architecture of a network that uses ReLU nonlinearities trained to extract the relevant information. (**B**) Architecture of a second network that instead uses polynomial nonlinearities to extract the relevant information. (**C**, **D**) Choice correlations based on polynomial statistics show that both networks' computations are consistent with optimal nonlinear decoding (Methods 4.2.2), even though the simulated networks used different implementations to extract the stimulus information. Horizontal axis shows optimal choice correlations (Eq. 7); vertical axis shows measured choice correlations (Eq. 5).

Previous work [3] 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, $\Gamma = \mathrm{Cov}(\boldsymbol{R}(\boldsymbol{r})|s)$ (Section 2.1.3). 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 cortical 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 choice correlations C_{R_k} can be substantial even for optimal nonlinear decoding: the discriminabilities d'_{R_k} of redundant statistics can be comparable to the discriminability d' of the whole population, producing ratios d'_{R_k}/d' that are not small (Figure 5).

2.2.5 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 the noise Γ_0 is smaller than the information-limiting variance 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

$$C_R^{\mathrm{sub}} \approx \alpha C_R^{\mathrm{opt}} + \chi_R \tag{9}$$

where χ_R depends on the particular type of suboptimal decoding (Supporting Information S.3.2). 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 one decoder that would be suboptimal without redundancy, but is nonetheless close to optimal when information limits are imposed. This rescue of optimality does not happen for all decoders, however. The figure also shows another decoder that is so suboptimal that it throws away most of the available information even when there is substantial redundancy. The patterns of choice correlations reflect this.

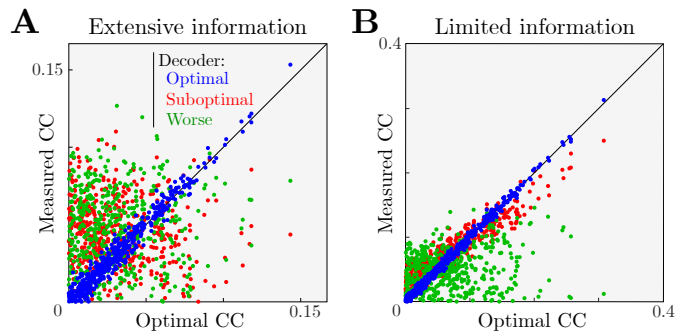

Figure 5: Information-limiting noise makes a network more robust to suboptimal decoding. (**A**) A simulated optimal decoder produces measured choice correlations that match our optimal predictions (blue, on diagonal). In contrast, when a noise covariance Γ_0 permits the population to have extensive information, then a suboptimal decoder can exhibits a pattern of choice correlations that does not match the prediction of optimal decoding. Here we show two suboptimal decoders, one that is blind to higher-order correlations ($\boldsymbol{w} \propto \boldsymbol{F}'$, red), and another 'worse' decoder that has the same weights but with 40% random sign flips (green). As in Figure 4, horizontal axis shows optimal choice correlations (Eq. 7) and vertical axis shows measured choice correlations (Eq. 5). (**B**) When information is limited, the same decoding weights may be less detrimental, and thus exhibit a similar pattern of choice correlations as an optimal decoder (red), or if they are sufficiently bad they may retain a suboptimal pattern of choice correlations (green).

In realistically redundant models with more cortical neurons than sensory inputs, many decoders could be

near-optimal, as we recently discovered in experimental data for a linear population code [27]. However, even in redundant codes there may be substantial inefficiencies and information loss [51], especially for unnatural tasks [52], so it is scientifically interesting to discover near-optimal nonlinear computation even in a redundant code.

2.3 Application to experimental data

2.3.1 Variance discrimination task

We applied our optimality test to data recorded with Utah arrays from primate visual cortex (V1) during a nonlinear decoding task (Method 4.4). Monkeys faced a Two-Alternative Forced Choice task (2AFC) in which they categorized an oriented grating based on whether it came from a wide or narrow distribution of orientations [53] (Figure 6A,B). The categorical target variable s is therefore the variance of the orientation distribution.

2.3.2 Evidence for optimal nonlinear computation in macaque brains

This is a coarse discrimination task, since the response statistics change significantly with the stimulus. As for fine discrimination, we again find that when decoding is optimal, random fluctuations in choices are correlated with neural responses to the same degree that those responses can discriminate between stimuli. However, as we describe in the Methods (4.4) and Supplement (S.6.1), this relationship is slightly more complicated for coarse discrimination, since the response statistics change significantly with the stimulus. For this reason we need to use a slightly more complicated measure of choice correlation that we call Normalized Average Conditional Choice Correlation (NACCC, Eq. 17). However, the end result is the same: choice correlations for optimal decoding are equal to the ratio of discriminabilities (Eq. 7; proof in Supplemental Information S.5, Eq. 138). Again there is a correction factor of order 1 for binary choices instead of continuous estimation (Methods 4.2.1, Supplemental Information S.6.4, Eq. 185).

V1 responses contain information about orientation [54]. Here we found that V1 responses also contain some linear information about the orientation *variance* (Figure 6C, blue). Since these neurons have linear information, they have already performed some useful nonlinear transformations of the input within their receptive field.

However, because neural responses in this brain area can be linearly decoded to compute orientation, a good decoder for the orientation variance would naturally be quadratic in those responses. Indeed, we found information in the quadratic statistics of neural responses, δr_i^2 and $\delta r_i \delta r_j$ (Figure 6C, red and green), suggesting that downstream nonlinear computations could extract additional information from the neural responses. Here we first eliminate the linear information when we compute these neural nonlinear statistics by using $\delta r_i = r_i - \langle r_i | \hat{s}_1 \rangle$, where $\hat{s}_1 = \boldsymbol{w}_{\text{opt}} \cdot \boldsymbol{r} + c$ is the optimal estimate decoded only from a linear combination of available neural responses.

We also found that these quadratic statistics contained substantial nonlinear information about the behavioral choice (Figure 6D). In general, there is no guarantee that the particular nonlinear statistics that are informative about the stimulus are also informative about the choice. Our theory of optimal decoding predicts specifically that these quantities should be directly proportional to each other.

Indeed, in two monkeys, we found that nonlinear choice correlations were highly correlated with nonlinear information (Figure 6E).

Remarkably, when we compare the measured nonlinear choice correlations to the ratio of discriminabilities after adjusting for the binary data (Methods 4.4), the slopes of this relationship for the two animals were near the value of 1 that Eq. 7 predicts for optimal decoding (Figure 6E).

2.3.3 Controls to find the origins of choice correlations

To evaluate whether internal noise correlations contribute nonlinear information or choice correlations, we created a shuffled data set that removed internal noise correlations while preserving external nuisance correlations. That is, we independently selected responses to trials with matched target stimulus (variance), nuisance (orientation), and choice, and repeated our analysis on these shuffled data (Figure 6F). The observed relationship between predicted and observed choice correlations was the same as in the original test, indicating that nuisance variations were sufficient to drive the nonlinear information and decoding.

We then shuffled the external nuisance correlations by randomly selecting responses to trials with matched

target stimulus and choice, but now using *unmatched* nuisance variables, and again repeated the analysis (Figure 6G). In other words, we picked responses from different trials that came from the same signal category (wide or narrow) and elicited the same choice but had different orientations, and we picked these trials (and thus their stimulus orientations) *independently* for neurons i and j. The strong statistical relationship observed between predicted and measured nonlinear choice correlations vanished with this shuffling, indicating that the nuisance variation was necessary for the nonlinear information and nonlinear decoding.

These shuffle controls removed noise correlations and nuisance correlations, respectively. Combining the conclusions from these controls, we find no evidence that the brain optimally decodes any stimulus-dependent internal noise correlations in this task. Recent analyses of these same data found that internal noise did in fact influence the monkeys' behavioral choices [55], but this effect was subtle and only apparent when examining the entire neural population simultaneously. In our analysis this effect is buried in the noise, so our method is not sensitive enough to tell if these large-scale patterns induced by internal noise are used optimally or suboptimally. However, we can detect that the brain contains information that is encoded nonlinearly due to external nuisance variation, and that this information is indeed decoded near-optimally by the brain.

One monkey performed slightly worse than an ideal observer, with a probability correct of 0.76, compared to the ideal of 0.82 (Methods 4.4) — even while its decoding was near-optimal, with an efficiency of 0.96±0.04 (mean ± 95% confidence intervals). This suggests that information is lost in the encoding stage somewhere between the stimulus and the recorded neurons, and not downstream of those neurons. The other monkey had similar overall performance (probability correct of 0.74) but worse decoding efficiency (0.75±0.08). This suggests the second monkey's task performance has limitations arising downstream of the recorded neurons.

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 the efficiency of the brain's decoding. Unlike previous theories [2,28], ours remains consistent with biological constraints due to the large cortical expansion of sensory representations by incorporating redundancy through nonlinear 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 decode 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 task.

On the negative side, this simple test does not offer a complete description of neural transformations. It instead tests just one important hypothesis about their functional role — that the brain performs optimal decoding. However, the theory does provide a practical way to estimate 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 which suboptimalities the brain tolerates for satisfactory performance [56].

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 (Figure 4). 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: 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 nonlin-

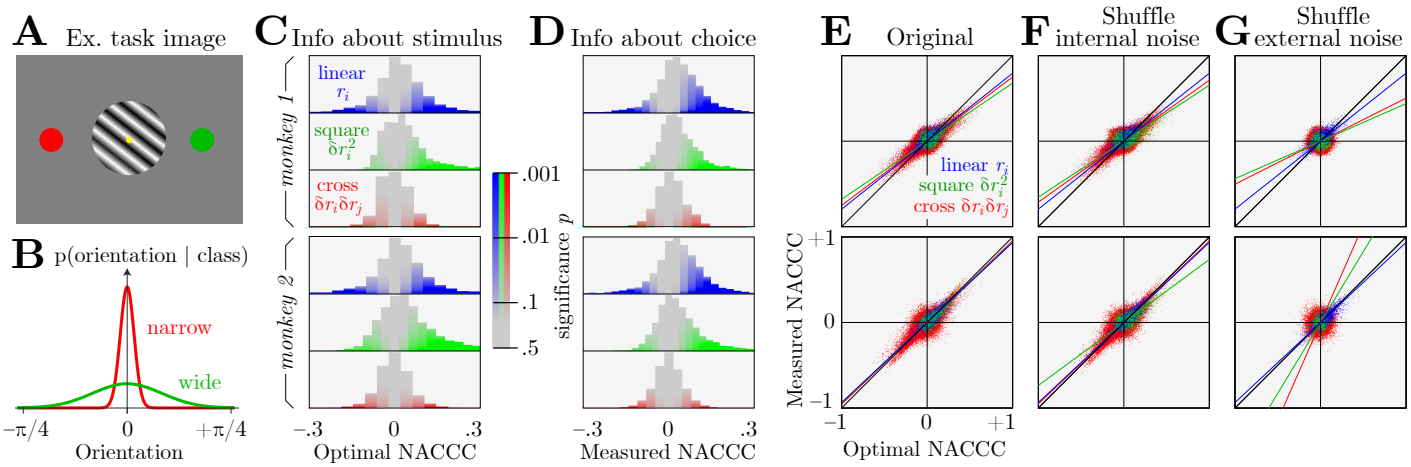

Figure 6: Nonlinear information and choice correlations in a variance discrimination task, for neural data from two monkeys. (**A**) Example oriented grating and saccade targets. (**B**) The orientations of the gratings were drawn from a narrow or wide distribution, and the monkey had to guess which by saccading to the appropriate target. (**C**) Neurons contain linear and nonlinear information about the task variable. This is revealed by the Normalized Average Conditional Choice Correlations (NACCC, Eq. 17) predicted for *optimal* decoding, which are proportional to the measured signal-to-noise ratios (Eq. 7) for each neural response pattern (blue r_i, green δr_i^2, red $\delta r_i \delta r_j$). Color saturation indicates statistical significance (Methods 4.4). (**D**) These neurons also contain significant information about the animal's choice, as computed by the *measured* NACCC. (**E**) The measured and optimal NACCCs are highly correlated, with a proportionality near 1 (lines). Each point represents one response pattern (*e.g.* $\delta r_i \delta r_j$) in one session. Top and bottom panels are data from two different monkeys. These two plotted quantities are strongly correlated (0.76, 0.65, 0.53 for linear, square and cross terms for monkey 1; 0.80, 0.83, 0.72 for monkey 2). (**F**) Shuffling internal noise correlations while preserving nuisance correlations maintains the relationship between prediction and nonlinear choice correlations, implying that internal noise is not responsible for the correlations. (**G**) Shuffling nuisance correlations across trials (Methods 4.4) nearly eliminates the relationship between measured and predicted nonlinear choice correlations (0.76, 0.05, 0.04 for monkey 1; 0.80, 0.10, 0.11 for monkey 2), implying that nuisance variation creates the nonlinear code.

earities 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 [57], that is not a compact way to express the complex nonlinearities of deep networks. Relatedly, it may be difficult to see statistically significant information or choice correlations for nonlinear statistics that provide many small contributions to the behavioral output. Alternatively, with 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 complement studies that demonstrate a match between the encodings of brains and artificial neural networks [10, 58].

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 [59], visual search with distractors (XOR-type tasks) [30, 60], sound localization in early auditory processing before the inferior colliculus [61], or context switching in higher-level cortex [62].

3.2 Limitations

For efficient decoding in a learned task, the optimality test (Eq. 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. Future work will demonstrate how we can use identify properties of suboptimal decoders [56] with nonlinear choice correlations.

Our approach is currently limited to feedforward processing, which unquestionably oversimplifies cortical processing. The approach can be generalized to recurrent networks by considering spatiotemporal statistics [56]. Nonetheless, feedforward models do a fair job of capturing the representational structure of the brain [10].

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 [63], 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, though of course this would require more simultaneous measurements.

3.3 Choice correlations from internal versus 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. Indeed, our analysis found no evidence of internal noise generating nonlinear choice correlations (Figure 6). 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 [9].

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, nonlinear noise correlations need not predict a fixed stimulus, but they may predict the choice (Supplementary Information S.4).

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 ν and internal noise. However, under some conditions, a related prediction for nonlinear choice correlations holds even without averaging over nuisance variables (Supplementary Information S.4).

3.4 Conclusion

Despite the clear importance of computation that is both nonlinear and distributed, and evidence for nonlinear coding in the cortex [30, 32–34], most neuroscience applications of population coding concepts have assumed linear codes and linear readouts [10, 27, 38, 64, 65]. The few that directly address nonlinear population codes either have an impossibly large amount of encoded information [2, 28], or investigate abstract properties unrelated to structured tasks [66]. Some experimental studies have been able to extract additional information from recorded populations using nonlinear decoders [30,67], 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, 68].

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. 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 nuisance variation in the experimental task may not match the natural tasks the brain has learned. For this reason it is important for neuroscience to use natural tasks, or at least naturalistic ones, when aiming to understand computational function [69].

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.

Author contributions

XP conceived the theoretical framework. XP and QY designed and performed the mathematical analyses. QY performed the simulations. EW, RJC and AT designed the experiments for a study with Wei Ji Ma; EW, RJC and AT performed the experiments; EW preprocessed the neural data; QY and XP analyzed the neural data. QY and XP wrote the manuscript; all authors discussed the results and commented on the manuscript.

Acknowledgements

The authors thank Jeff Beck, Valentin Dragoi, Arun Parajuli, Alex Pouget, and Haim Sompolinsky for helpful conversations. This work was supported by NSF CAREER grant 1552868 to XP, by NeuroNex grant 1707400 to XP and AT, and by NSF Grant No. PHY-1748958, NIH Grant No. R25GM067110, and the Gordon and Betty Moore Foundation Grant No. 2919.01.

Data and code availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. All custom code used for electrophysiology data collection and data processing are made publicly available at `github.com/atlab`. Experimental data for Figure 6 and code used for analysis and figure generation are available for download from `github.com/xaqlab/nonlinear_choice_correlation`.

References

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

[2] 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.

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

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

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

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

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

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

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

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

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

[21] 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.

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

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

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

[29] 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.

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

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

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

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

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

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

[44] 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.

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

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

[47] 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.

[48] 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.

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

[50] Kanitscheider I, Coen-Cagli R, Pouget A (2015) Origin of information-limiting noise correlations. Proceedings of the National Academy of Sciences 112: E6973–E6982.

[51] Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A (2012) Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron 74: 30-9.

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

[53] Qamar AT, Cotton RJ, George RG, Beck JM, Prezhdo E, Laudano A, Tolias AS, et al. (2013) Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. Proceedings of the National Academy of Sciences 110: 20332–20337.

[54] Hubel DH, Wiesel TN (1962) Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology 160: 106–154.

[55] Walker EY, Cotton RJ, Ma WJ, Tolias AS (2018) A neural basis of probabilistic computation in visual cortex. bioRxiv : 365973.

[56] Lakshminarasimhan K, Pouget A, DeAngelis G, Angelaki D, Pitkow X (2018) Inferring decoding strategies for multiple correlated neural populations. PLoS Computational Biology 14: e1006371.

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

[58] Kietzmann TC, Spoerer CJ, Sörensen L, Cichy RM, Hauk O, Kriegeskorte N (2019) Recurrence required to capture the dynamic computations of the human ventral visual stream. arXiv preprint arXiv:190305946 .

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

[60] 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.

[61] 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.

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

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

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

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

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

[67] 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.

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

[69] Pitkow X, Angelaki DE (2017) Inference in the brain: statistics flowing in redundant population codes. Neuron 94: 943–953.

[70] 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.

[71] Tolias AS, Ecker AS, Siapas AG, Hoenselaar A, Keliris GA, Logothetis NK (2007) Recording chronically from the same neurons in awake, behaving primates. Journal of neurophysiology 98: 3780–3790.

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

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

4 Online 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, $G(\boldsymbol{x}|s,\nu)$, where $\boldsymbol{x}$ is spatial position in the image, and s and ν are the orientation and spatial phase (nuisance) of the image, respectively (Supplemental Material S.1.1). 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 $\boldsymbol{r} \sim P(\boldsymbol{r}|s,\nu) = N(\boldsymbol{r}|\boldsymbol{f}(s,\nu),\epsilon I)$, where ϵ is the noise variance and the mean $\boldsymbol{f}(s,\nu) = \langle \boldsymbol{r}|s,\nu\rangle = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s,\nu)$ is determined by the overlap between the image and the receptive field.

When the spatial phase ν 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, each mean response to a fixed orientation will be combined across different phases: $\boldsymbol{f}(s) = \langle \boldsymbol{r}|s\rangle = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s) = \int d\nu \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s,\nu)p(\nu)$. 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 $\mathrm{Cov}_{ij}(\boldsymbol{r}|s,\nu)$ 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,

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

where $\delta f_i(s,\nu) = f_i(s,\nu) - \langle f_i(s,\nu)\rangle_\nu$. Supplementary Information S.1.1 shows in detail how $\mathrm{Cov}_{ij}(\boldsymbol{r}|s)$ is tuned to s.

4.1.2 Exponential family distribution and sufficient statistics

It is illuminating to 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. In this case, a finite number of sufficient statistics contains all of the information about the stimulus 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 S.1.2.2, we show that the optimal local estimation can be achieved by linearly decoding the nonlinear sufficient statistics, $\hat{s} = \boldsymbol{w}^\top \boldsymbol{R}(\boldsymbol{r}) + c$. The decoding weights minimize the variance of an unbiased decoder,

$$\boldsymbol{w}_{\mathrm{opt}} \propto \boldsymbol{H}'(s) \propto \Gamma^{-1}\boldsymbol{F}' \quad (12)$$

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 2.1.3).

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, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j\}$ 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 $f(s)$ is constant, while the stimulus-conditional covariances $\Sigma_{ij}(s)$ depend smoothly on the stimulus. We can quantify the information content of the neural population using Equation 61.

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, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j, r_i r_j r_k\}$ 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 $z_i z_j z_k$ (and a base measure $e^{-\|\boldsymbol{z}\|^4}$ to ensure the distribution is bounded and normalizable).

$$p(\boldsymbol{z}|s) = \tfrac{1}{Z}\exp\left(-\|\boldsymbol{z}\|^4 + \gamma s\, z_i z_j z_k\right) \quad (13)$$

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 S1), 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 (S1), 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:

$$\boldsymbol{r} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\boldsymbol{z} \quad (14)$$

where $\boldsymbol{f}(s)$ and $\Sigma(s)$ describes the desired signal-dependent mean and covariance (see Supplemental Material S.1.4).

4.2 Nonlinear choice correlations

For fine discrimination tasks, 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

$$C_{R_k} = \mathrm{Corr}(R_k(\boldsymbol{r}), \hat{s}|s) = \frac{(\Gamma\boldsymbol{w})_k}{\sqrt{\Gamma_{kk}\boldsymbol{w}^\top\Gamma\boldsymbol{w}}} \quad (15)$$

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

When the relevant response statistics change appreciably over the stimulus range used in the task, such as for the coarse variance discrimination task in Section 2.3), the relevant quantities change slightly. The optimal linear decoder of nonlinear statistics, $\hat{s} = \boldsymbol{w} \cdot \boldsymbol{R} + c$, has weights obtained through linear regression:

$$\boldsymbol{w} \propto \bar{\Gamma}^{-1} \Delta \boldsymbol{F} \tag{16}$$

where $\bar{\Gamma} = \langle \mathrm{Cov}\left(\boldsymbol{R}|s\right)\rangle_s$ is the average conditional covariance between $\boldsymbol{R}$ given the stimulus s. The differences from Eq 12 are $\Gamma \to \bar{\Gamma}$ and $\boldsymbol{F}' = d\boldsymbol{F}/ds \to \Delta \boldsymbol{F}/\Delta s$.

These differences are reflected in a slightly modified measure of correlation that we call Normalized Average Conditional Choice Correlations (NACCC),

$$B_{R_k} = \frac{\langle \mathrm{Cov}(R_k, \hat{s}|s)\rangle_s}{\sqrt{\langle \mathrm{Var}(R|s)\rangle_s \langle \mathrm{Var}(\hat{s}|s)\rangle_s}} = \frac{(\bar{\Gamma}\boldsymbol{w})_k}{\sqrt{\bar{\Gamma}_{kk} \boldsymbol{w}^\top \bar{\Gamma} \boldsymbol{w}}} \tag{17}$$

Note that Eq 17 is not actually a correlation coefficient, and may exceed the interval $[-1, 1]$ if $\mathrm{Var}(\hat{s}|s)$ and $\mathrm{Var}(R_k|s)$ are anticorrelated over s. As the stimulus range in a coarse task decreases, and the noise distribution $p(\boldsymbol{R}|s)$ becomes independent of the stimulus, then Eq 17 converges toward Eq 15.

The choice correlation for binary choices differs slightly from that for continuous estimation, for both fine and coarse discrimination tasks, by a factor ζ that is typically of order 1 (Supplemental Materials S.6.1).

4.2.1 Optimality test

Substituting the optimal weights (Eq 12) into (15), the optimal nonlinear choice correlation becomes

$$C^{\mathrm{opt}}_{R_k(\boldsymbol{r})} = \frac{\left(\Gamma\Gamma^{-1}\boldsymbol{F}'\right)_k}{\sqrt{\Gamma_{kk}\boldsymbol{F}'^\top \Gamma^{-1}\boldsymbol{F}'}} = \frac{F'_k}{\sqrt{\Gamma_{kk}}}\sigma_{\hat{s}} = \frac{d'_{R_k(\boldsymbol{r})}}{d'} \tag{18}$$

where $d'_{R_k(\boldsymbol{r})} = F'_k \Delta s/\sqrt{\Gamma_{kk}}$ is the fine discriminability provided by $R_k(\boldsymbol{r})$ for a stimulus difference of Δs. The same argument holds for coarse discrimination, where $\bar{\Gamma}$ in Eq 17 is canceled by $\bar{\Gamma}^{-1}$ in the optimal weights (Eq 16), yielding $B^{\mathrm{opt}}_{R_k(\boldsymbol{r})} = d'_{R_k}/d'$.

For fine-scale discrimination, optimal choice correlations can be written in many equivalent ways that reflect the simple relationships between four quantities often used to represent information: discriminability d-prime is proportional to the square root of the Fisher information $d' = \Delta s\sqrt{J}$ [70]; estimator variance is bounded by the inverse of the Fisher information, $\sigma_{\hat{s}}^2 \geq 1/J$; discrimination threshold is proportional to the estimator standard deviation, $\theta = \sqrt{\sigma_{\hat{s}}^2}$ with proportionality given by the threshold condition.

In different experiments (binary discrimination, continuous estimation), it can be most natural to express this optimal decoding prediction as ratios of different measured quantities:

$$C^{\mathrm{opt}}_{R_k} = \frac{d'_{R_k}}{d'} = \frac{\theta}{\theta_{R_k}} = \sqrt{\frac{\sigma_{\hat{s}}^2}{\sigma^2_{\hat{s},R_k}}} = \sqrt{\frac{J_{R_k}}{J}} \tag{19}$$

These quantities reflect information between the stimulus and the neural or behavioral responses. Supplemental material S.5 shows how this can be computed easily for general binary discrimination using the total correlation between the responses and the stimuli, $D_{R_k} = \mathrm{Corr}\left(R_k, s\right)$, or a continuously varying behavioral choice $\hat{s}$ and the stimuli, $D_{\hat{s}} = \mathrm{Corr}\left(\hat{s}, s\right)$:

$$d' = \frac{2}{\sqrt{D^{-2} - 1}} \approx 2D \tag{20}$$

and likewise for d'_{R_k}. When the behavioral choice is binary rather than continuous, the correlations are modified by a factor δ near 1 (Supplemental Information S.6.3, Eq 182). For our experimental conditions, $\delta \approx 1.2 \pm 0.2$.

4.2.2 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). 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. 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 [3]. 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 $\boldsymbol{F}'\boldsymbol{F}'^\top$:

$$\Gamma = \Gamma_0 + \frac{1}{J_\infty}\boldsymbol{F}(s)'\boldsymbol{F}(s)'^\top \tag{21}$$

Here Γ_0 is any covariance of $\boldsymbol{R}$ that does *not* limit information in large populations. Substituting this expression into (61) for the nonlinear Fisher Information, we obtain

$$J = \boldsymbol{F}'\Gamma^{-1}\boldsymbol{F}' = \frac{1}{1/J_\infty + 1/J_0} \tag{22}$$

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_∞.

4.4 Application to neural data

All behavioral and electrophysiological data were obtained from two healthy, male rhesus macaque (*Macaca mulatta*) monkeys (L and T) aged 10 and 7 years and weighting 9.5 and 15.1 kg,

respectively. All experimental procedures complied with guidelines of the NIH and were approved by the Baylor College of Medicine Institutional Animal Care and Use Committee (permit number: AN-4367). Animals were housed individually in a room located adjacent to the training facility on a 12h light/dark cycle, along with around ten other monkeys permitting rich visual, olfactory, and auditory social interactions. Regular veterinary care and monitoring, balanced nutrition and environmental enrichment were provided by the Center for Comparative Medicine of Baylor College of Medicine. Surgical procedures on monkeys were conducted under general anesthesia following standard aseptic techniques.

Monkeys faced a Two-Alternative Forced Choice (2AFC) to guess whether an oriented drifting grating stimulus came from a narrow or wide distribution of orientations, centered on zero with standard deviations $\sigma_+ = 15^\circ$ and $\sigma_- = 3^\circ$. Visual contrast was set to 64%. Each trial was initiated by a beeping sound and the appearance of a fixation target (0.15° visual angle) in the center of the screen. The monkey fixated on a fixation target for 300ms within 0.5°–1° visual angle. The stimulus appeared at the center of the screen. After 500ms, colored targets appeared randomly on the left and right, and the monkey then saccades to one of these targets to indicate its choice (red and green targets correspond to narrow and wide distributions).

After the monkey was fully trained, we implanted a 96-electrode microelectrode array (Utah array, Blackrock Microsystems, Salt Lake City, UT, USA) with a shaft length of 1 mm over parafoveal area V1 on the right hemisphere. The neural signals were pre-amplified at the head stage by unity gain preamplifiers (HS-27, Neuralynx, Bozeman MT, USA). These signals were then digitized by 24-bit analog data acquisition cards with 30 dB onboard gain (PXI-4498, National Instruments, Austin, TX) and sampled at 32 kHz. The spike detection was performed offline according to a previously described method [12, 71]. Code for spike detection is available online at `github.com/atlab/spikedetection`. For each behavioral session and in both monkeys, 95 multiunit neural responses r_k were measured by spike counts in the 500 ms preceding the saccade target onset.

The animals did not perform well on all days, so for further analysis we selected sessions where the performance exceeded 0.7 for monkey 1 (85% of all sessions) and 0.75 for monkey 2 (68% of all sessions).

The task-relevant stimulus s is the large or small variance $s_\pm = \sigma_\pm^2$ of the distribution over orientations. Orientation is a nuisance variable ν, drawn from $p(\nu|s) = \mathcal{N}(\nu|0, s)$, which has sufficient statistics that are quadratic in ν. If the orientation itself can be estimated locally from linear functions of the neural responses, then the stimulus can be decoded quadratically from those neural responses, $\hat{s} = \hat{\nu}^2$. A binary guess about the variance is given by $\hat{s}_\pm = \operatorname{sgn}(\hat{\nu}^2 - \theta^2)$ where θ is the animal's orientation threshold. This threshold is optimal where the two stimuli are equally probable: $p(\nu|s_+) = p(\nu|s_-)$, implying that $\theta^2_{\text{opt}} = (\log s_+ - \log s_-) / \left(s_-^{-1} - s_+^{-1}\right)$. The probability of correctly guessing the orientation variance is $\frac{1}{2}\left(p(\hat{s}_\pm = +|s_+) + p(\hat{s}_\pm = -|s_-)\right)$, where these probabilities can be computed from the cumulative normal distribution on the correct side of the optimal orientation threshold, $p(\hat{s}_\pm = +|s_+) = 2\int_{\theta_{\text{opt}}}^\infty d\nu\, p(\nu|s_+) = \operatorname{erfc}\left(\theta_{\text{opt}}/\sqrt{2s_+}\right)$; similarly, $p(\hat{s}_\pm = -|s_-) = 1 - \operatorname{erfc}(\theta_{\text{opt}}/\sqrt{2s_-})$. Using values of $s_\pm$ for our task, this gives an optimal fraction correct of 0.82.

We computed choice correlations using NACCC (Eq 17), and discriminability based on total correlations between stimulus and response (Eq 20). We adjusted the optimal prediction by constant factors ζ and δ to account for binary choices using the equations in Supplement S.6.4, with thresholds estimated by logistic regression between choice and the absolute value of the stimulus orientation. We estimated the slopes of the relationship between measured and predicted choice correlation using the angle of the principal component of the bivariate data. We computed standard deviations for these quantities by bootstrapping 100 times.

For our two shuffle controls testing whether correlations between neurons were informative about the stimulus or choice, we selected responses independently from $r_i \sim p(r_i|s, \nu, \hat{s})$ (Figure 6F) or $r_i \sim p(r_i|s, \hat{s})$ (Figure 6G).

We evaluate statistical significance of the measured and predicted optimal choice correlations using p-values for null distributions based on 100 shuffled choices and 100 shuffled stimuli, while preserving correlations between neural responses. Both null distributions are approximately Gaussian with zero means, so we compute the p-value of the choice correlations with respect to the corresponding Gaussian, $p = 1 - \frac{1}{2}\operatorname{erfc}(-|x|/\sqrt{2\sigma_x})$ where x is the quantity of interest and σ_x is its standard deviation (Figure 6C,D).

Supplemental material

S.0 Overview

This supplemental material contains mathematical details and proofs of the central ideas presented in the main text.

S.1 Encoding models

- **S.1.1** Orientation estimation task with phase as nuisance
- **S.1.2** Exponential family distributions
- **S.1.3** Quadratic codes
- **S.1.4** Cubic codes

S.2 Information-limiting correlations
S.3 Analyzing decoding quality
S.4 Choice correlations from internal and external sources
S.5 Coarse discrimination and choice correlations
S.6 Orientation variance discrimination task

S.1 Encoding models

S.1.1 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

$$G(\boldsymbol{x}|s,\nu) = e^{-\|\boldsymbol{x}\|^2}\cos\left(\boldsymbol{k}\cdot\boldsymbol{x}+\nu\right) \tag{23}$$

where $\boldsymbol{k} = \kappa(\cos s, \sin s)$. Here the target s is the orientation of the pattern, ν 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

$$\mathrm{RF}_j(\boldsymbol{x}, s_j, \nu_j) = e^{-\|\boldsymbol{x}\|^2}\cos\left(\boldsymbol{k}_j\cdot\boldsymbol{x}+\nu_j\right) \tag{24}$$

$$\boldsymbol{k}_j = \kappa(\cos s_j, \sin s_j) \tag{25}$$

where each neuron has a preferred orientation s_j, spatial phase ν_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:

$$\begin{aligned} f_j(s,\nu) &= \int d\boldsymbol{x}\, G(\boldsymbol{x}|s,\nu)\mathrm{RF}_j(\boldsymbol{x}, s_j, \nu_j) \\ &= \Big[e^{-\frac{1}{4}\kappa^2\cos(s-s_j)}\cos(\nu+\nu_j) \\ &\quad + e^{+\frac{1}{4}\kappa^2\cos(s-s_j)}\cos(\nu-\nu_j)\Big]\,\frac{\pi}{4}e^{-\frac{1}{4}\kappa^2} \end{aligned} \tag{26}$$

This expression can be written in the form:

$$f_j(s,\nu) = A_j(s)\cos\left(\nu+\psi_j(s)\right) \tag{27}$$

using the stimulus-dependent response amplitude

$$A_j(s) = C\sqrt{2\cosh 2\beta_j(s) + 2\cos 2\nu_j} \tag{28}$$

and phase

$$\psi_j(s) = \nu_j - \alpha_j(s) \tag{29}$$

where we define the quantities

$$C = \frac{\pi}{4}\exp\left(-\frac{1}{4}\kappa^2\right) \tag{30}$$

$$\beta_j(s) = \frac{1}{4}\kappa^2\cos(s-s_j) \tag{31}$$

$$\alpha_j(s) = \tan^{-1}\frac{\exp\left(\beta_j(s)\right)\sin 2\nu_j}{\exp\left(-\beta_j(s)\right)+\exp\left(\beta_j(s)\right)\cos 2\nu_j} \tag{32}$$

Equation 27 reveals that the mean response of each neuron traces out a sinusoidal oscillation in ν, 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 ν, 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

$$P(\boldsymbol{r}|\nu, s) = \mathcal{N}(\boldsymbol{f}(s,\nu), \boldsymbol{I}) \tag{33}$$

If the spatial phase ν 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

$$\begin{aligned} f_j(s) = \langle r_j|s\rangle &= \int r_j\, p(r_j|s)dr_j && (34)\\ &= \iint r_j\, p(r_j|s,\nu)\, p(\nu)\, dr_j\, d\nu && (35)\\ &= \int f_j(s,\nu)p(\nu)\, d\nu && (36)\\ &= \frac{1}{2\pi}\int f_j(s,\nu)\, d\nu && (37)\\ &= \frac{A_j(s)}{2\pi}\int_0^{2\pi}\cos\left(\nu+\psi_j(s)\right)d\nu && (38)\\ &= 0 && (39) \end{aligned}$$

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 $\mathrm{Cov}_{ij}[\boldsymbol{r}|s,\nu]$ 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

$$\begin{aligned}
\mathrm{Cov}_{ij}[\boldsymbol{r}|s] &= \langle (r_i - f_i(s))(r_j - f_j(s))|s\rangle && (40)\\
&= \langle r_i r_j|s\rangle = \iint r_i r_j\, p(\boldsymbol{r}|s) dr_i dr_j && (41)\\
&= \int d\nu \iint r_i r_j\, p(\boldsymbol{r}|s,\nu) p(\nu) dr_i dr_j && (42)\\
&= \int d\nu\, p(\nu) \langle r_i r_j|s,\nu\rangle && (43)\\
&= \int d\nu\, p(\nu) \left(\mathrm{Cov}_{ij}[\boldsymbol{r}|s,\nu] + f_i(s,\nu) f_j(s,\nu)\right) && (44)\\
&= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi}\int d\nu f_i(s,\nu) f_j(s,\nu) && (45)\\
&= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi} D_{ij}(s) && (46)
\end{aligned}$$

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

$$\begin{aligned}
D_{ij}(s) &= \int d\nu f_i(s,\nu) f_j(s,\nu)\\
&= \int d\nu\, A_i(s)\cos(\nu + \psi_i(s)) A_j(s)\cos(\nu + \psi_j(s)) && (47)\\
&= \pi\cos(\psi_i(s) - \psi_j(s)) A_i(s) A_j(s)
\end{aligned}$$

Here when we compute Equation 47, we used the trigonometric identity: $2\cos(x)\cos(y) = \cos(x+y) + \cos(x-y)$, and $\int \cos(2\nu + \psi_i + \psi_j) d\nu = 0$.

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 (34), quadratic (or higher-order) nonlinear computations can be used to estimate the orientation.

S.1.2 Exponential family distributions

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

$$p(\boldsymbol{r}|s) = b(\boldsymbol{r}) \exp\left(\boldsymbol{\Theta}(s)^\top \boldsymbol{R}(\boldsymbol{r}) - A(s)\right) \tag{48}$$

where $\boldsymbol{\Theta}(s)$ 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.

S.1.2.1 Fisher information

One measure of information content that a population response contains about a stimulus is the Fisher information $J(s)$ [1–3,24–26]. The Fisher information is given by

$$\begin{aligned}
J &= -\left\langle \frac{\partial^2}{\partial s^2} \log p(\boldsymbol{r}|s)\right\rangle_{\boldsymbol{r}|s} && (49)\\
&= \left\langle \left(\frac{\partial}{\partial s} \log p(\boldsymbol{r}|s)\right)^2\right\rangle_{\boldsymbol{r}|s} && (50)
\end{aligned}$$

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 $\boldsymbol{\Theta}(s)$,

$$\boldsymbol{F} = \frac{\partial A(s)}{\partial \boldsymbol{\Theta}(s)} \tag{51}$$

Equation 51 can give us the first and second derivatives of $A(s)$ over s.

$$\begin{aligned}
A' &= \sum_i \frac{\partial A}{\partial \Theta_i}\frac{d\Theta_i}{ds} = \boldsymbol{\Theta}'^\top \boldsymbol{F} && (52)\\
A'' &= \boldsymbol{\Theta}''^\top \boldsymbol{F} + \boldsymbol{\Theta}'^\top \boldsymbol{F}' && (53)
\end{aligned}$$

Thus we can compute two definitions of Fisher information.

$$\begin{aligned}
J &= -\left\langle \frac{\partial^2}{\partial s^2} \log P(\boldsymbol{r}|s)\right\rangle_{P(\boldsymbol{r}|s)} && (54)\\
&= A'' - \boldsymbol{\Theta}''^\top \boldsymbol{F} && (55)\\
&= \boldsymbol{\Theta}'^\top \boldsymbol{F}' && (56)
\end{aligned}$$

and

$$\begin{aligned}
J &= \left\langle \left(\frac{\partial}{\partial s} \log P(\boldsymbol{r}|s)\right)^2\right\rangle_{P(\boldsymbol{r}|s)} && (57)\\
&= \boldsymbol{\Theta}'^\top (\langle \boldsymbol{R}\boldsymbol{R}^\top\rangle - \boldsymbol{F}\boldsymbol{F}^\top)\boldsymbol{\Theta}' && (58)\\
&= \boldsymbol{\Theta}'^\top \Gamma \boldsymbol{\Theta}' && (59)
\end{aligned}$$

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

Since the two definition are equivalent, we have

$$\boldsymbol{\Theta}' = \Gamma^{-1}\boldsymbol{F}' \tag{60}$$

Substituting Equation 60 into Equation 59, we find the Fisher Information for the exponential family [23]

$$J = \boldsymbol{F}'^\top \Gamma^{-1} \boldsymbol{F}' \tag{61}$$

S.1.2.2 Optimal 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:

$$\begin{aligned}
\hat{s} &= \underset{s}{\mathrm{argmax}}\; p(\boldsymbol{r}|s) && (62)\\
&= \underset{s}{\mathrm{argmax}}\; \log p(\boldsymbol{r}|s) && (63)\\
&= \underset{s}{\mathrm{argmax}}\; \boldsymbol{\Theta}(s)^\top \boldsymbol{R}(\boldsymbol{r}) - A(s) && (64)
\end{aligned}$$

A Taylor expansion around the reference yields

$$\begin{aligned}
&\boldsymbol{\Theta}(s)^\top \boldsymbol{R}(\boldsymbol{r}) - A(s)\\
&\approx [\boldsymbol{\Theta}^\top \boldsymbol{R} - A]\\
&+ [\boldsymbol{\Theta}'^\top \boldsymbol{R} - A'](s - s_0)\\
&+ \tfrac{1}{2}(s - s_0)^\top [\boldsymbol{\Theta}''^\top \boldsymbol{R} - A''](s - s_0) + \cdots
\end{aligned} \tag{65}$$

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

$$0 = [\boldsymbol{\Theta}'^\top \boldsymbol{R} - A'] + (\hat{s} - s_0)[\boldsymbol{\Theta}''^\top \boldsymbol{R} - A''] \tag{66}$$

The solution is

$$\hat{s} = s_0 - \frac{\boldsymbol{\Theta}'^\top \boldsymbol{R} - A'}{\boldsymbol{\Theta}''^\top \boldsymbol{R} - A''} \tag{67}$$

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, $\boldsymbol{\Theta}''^\top \boldsymbol{R} - A'' = \boldsymbol{\Theta}''^\top \boldsymbol{F} - A'' + \boldsymbol{\Theta}''^\top \delta\boldsymbol{R}$, 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

$$\hat{s} = \boldsymbol{w}^\top \boldsymbol{R} + \boldsymbol{c} \tag{68}$$

where

$$\boldsymbol{w} = \frac{\boldsymbol{\Theta}'}{J} = \frac{\Gamma^{-1}\boldsymbol{F}'}{\boldsymbol{F}'^\top \Gamma^{-1} \boldsymbol{F}'} \tag{69}$$

and where we used the results from Equations 61 and 60, 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 $\boldsymbol{R}(\boldsymbol{r})$.

S.1.3 Quadratic codes

In a purely quadratic coding model (no linear information), the distribution of neural responses is described by the exponential family with quadratic sufficient statistics, $p(\boldsymbol{r}|s) \sim \exp\left[\boldsymbol{\Theta}(s)^\top \boldsymbol{R}(\boldsymbol{r})\right]$ 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:

$$\Sigma(s) = V(s)\Lambda V(s)^\top \tag{70}$$

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 of $\Sigma(s)$.

To calculate the Fisher Information (Equation 61), 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 $\boldsymbol{r}$,

$$\Sigma'(s) = U e^{\Omega s}(\Omega X - X\Omega)e^{-\Omega s}U^\dagger \tag{71}$$

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

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:

$$\begin{aligned}\langle r_i r_j r_k r_n|s\rangle = &\langle r_i r_j|s\rangle \langle r_k r_n|s\rangle + \langle r_j r_n|s\rangle \langle r_i r_k|s\rangle \\ &+ \langle r_i r_n|s\rangle \langle r_j r_k|s\rangle\end{aligned} \tag{72}$$

where

$$\langle r_i r_j|s\rangle = \Sigma_{ij} + f_i f_j \tag{73}$$

Substitution of the response covariance (Equation 70) into Equation 72 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.

S.1.4 Cubic codes

In Figure S1 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

$$p(\boldsymbol{z}|s) = \tfrac{1}{Z}\exp\left(\gamma(s) z_i z_j z_k - \|\boldsymbol{z}\|^4\right) \tag{74}$$

where we include the base measure $e^{-\|\boldsymbol{z}\|^4}$ to ensure normalizability (Figure S1A).

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

$$p(\boldsymbol{z}|s) \approx \sum_{a=1}^{4} p(a)p(\boldsymbol{z}|a, s) \tag{75}$$

$$= \sum_a \frac{1}{4}\mathcal{N}(\boldsymbol{z}|\mu_a(s), M_a(s)) \tag{76}$$

where

$$\boldsymbol{m}_a(s) = \frac{s}{\sqrt{1+s^2}}\boldsymbol{v}_a \tag{77}$$

and

$$M_a(s) = \frac{I + s^2 \boldsymbol{v}_a \boldsymbol{v}_a^\top}{(1+s^2)^2} \tag{78}$$

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 74, Figure S1). To sample from this distribution, we randomly choose a component a and then sample from the gaussian $\mathcal{N}(\boldsymbol{z}|\boldsymbol{m}_a(s), M_a(s))$ 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 S1).

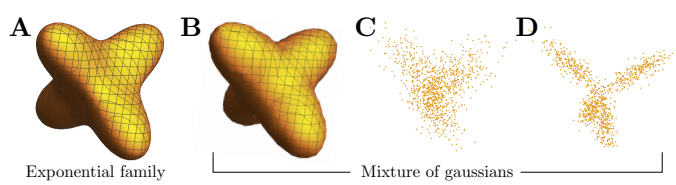

Figure S1: Multivariate skewed distributions. (**A**) Isoprobability contour of an exponential family distribution with cubic statistics in three dimensions, drawn from $p(\boldsymbol{z}|s) \propto \exp\left(s z_1 z_2 z_3 - \|\boldsymbol{z}\|^4\right)$. (**B**) Isoprobability contour for a mixture of four gaussians (Eq 76). (**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.

$$p(\boldsymbol{z}|s) = \prod_\alpha p(\boldsymbol{z}_\alpha|s) = \prod_\alpha p(z_{\alpha_1}, z_{\alpha_2}, z_{\alpha_3}|s) \tag{79}$$

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:

$$\boldsymbol{r} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\,\boldsymbol{z}\,\Sigma^{1/2}(s) \tag{80}$$

$$\boldsymbol{z} \sim \frac{1}{Z(s)}\exp\left[\sum_{ijk}\gamma_{ijk}(s) z_i z_j z_k - \|\boldsymbol{z}\|^4\right] \tag{81}$$

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

$$p(\boldsymbol{r}|a, s) = \mathcal{N}(\boldsymbol{r}|\boldsymbol{f}(s) + \boldsymbol{m}_a(s), \Sigma^{1/2}(s) M_a(s) \Sigma^{1/2}(s)) \tag{82}$$

Note that the linear and quadratic information terms are independent of the component a.

S.2 Information-limiting correlations

Information-limiting correlations [3] describe variability that cannot be averaged away because they are indistinguishable from changes in the stimulus. These fluctuations 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:

$$\begin{aligned}\Gamma =& \mathrm{Cov}_{\boldsymbol{r},ds}(\boldsymbol{R}(\boldsymbol{r})|s) \\ =& \left\langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s,ds)\right\rangle_{ds} + \mathrm{Cov}_{ds}\left\langle \boldsymbol{R}(\boldsymbol{r})|s,ds\right\rangle_{\boldsymbol{r}}\end{aligned} \tag{83}$$

where $\langle\cdot\rangle$ indicates an expectation value over the subscripted variable. The first term can be computed as follows,

$$\left\langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s,ds)\right\rangle_{ds} = \left\langle \Gamma(s+ds)\right\rangle_{ds} \tag{84}$$

$$\approx \left\langle \Gamma_0 + ds\,\Gamma'\right\rangle_{ds} \tag{85}$$

$$= \Gamma_0 \tag{86}$$

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 83 can be expressed as

$$\mathrm{Cov}_{ds}\left\langle \boldsymbol{R}(\boldsymbol{r})|s,ds\right\rangle_{\boldsymbol{r}} \tag{87}$$

$$= \mathrm{Cov}_{ds}(\boldsymbol{F}(s+ds)) \tag{88}$$

$$\approx \mathrm{Cov}_{ds}(\boldsymbol{F}(s) + ds\,\boldsymbol{F}'(s)) \tag{89}$$

$$= \frac{1}{J_\infty}\boldsymbol{F}'(s)\boldsymbol{F}'(s)^\top \tag{90}$$

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 83 can therefore be written as

$$\Gamma = \Gamma_0 + \frac{1}{J_\infty}\boldsymbol{F}(s)'\boldsymbol{F}(s)'^\top \tag{91}$$

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}:

$$\Gamma^{-1} = \Gamma_0^{-1} - \frac{\Gamma_0^{-1}\boldsymbol{F}'\boldsymbol{F}'^\top\Gamma_0^{-1}}{J_\infty + \boldsymbol{F}'\Gamma_0^{-1}\boldsymbol{F}'^\top} \tag{92}$$

Substituting these equations into the nonlinear Fisher Information (Equation 61) and simplifying, we obtain

$$J_{R(\boldsymbol{r})} = \frac{1}{1/J_\infty + 1/J_0} \tag{93}$$

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 [1, 2], so for large populations the output information saturates at J_∞.

S.3 Analyzing decoding quality

S.3.1 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, $\boldsymbol{\Psi}(\boldsymbol{r}) = (r_i, r_ir_j, r_ir_jr_k, ...)$.

The locally optimal decoder is a weighted sum of the sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$ (Equation 68):

$$\hat{s}_{\mathrm{opt}} = \boldsymbol{w}\cdot\boldsymbol{R}(\boldsymbol{r}). \tag{94}$$

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

$$\hat{s}_{\mathrm{brain}} = \boldsymbol{v}\cdot\boldsymbol{g}(\boldsymbol{r}). \tag{95}$$

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 $\hat{s}_{\mathrm{brain}}$ 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, $\boldsymbol{R}(\boldsymbol{r}) = (r_i, r_ir_j, r_ir_jr_k, ...)$. 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,

$$\hat{s}_{\mathrm{brain}} = \boldsymbol{v}\cdot\boldsymbol{r}^{(3)} + \boldsymbol{b}^{(3)} \tag{96}$$

$$\boldsymbol{r}^{(3)} = \mathrm{ReLU}(\boldsymbol{W}^{(2)}\boldsymbol{r}^{(2)} + \boldsymbol{b}^{(2)}) \tag{97}$$

$$\boldsymbol{r}^{(2)} = \mathrm{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{r}^{(1)} + \boldsymbol{b}^{(1)}) \tag{98}$$

$$\boldsymbol{r}^{(1)} = \boldsymbol{r} \tag{99}$$

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 applied the nonlinear choice correlation test (Equation 7) using monomial nonlinearities $\Psi(\boldsymbol{r})$ up to third order, *e.g.* r_i, r_ir_j, $r_i^2r_j$, r_k^3, etc. 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 $\hat{s}_{\mathrm{brain}}$. 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 $\sqrt{J} \approx 1/\sigma_{\hat{s}_{\mathrm{brain}}}$. 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 variance was $\sigma_\Psi^2 = \frac{1}{2}\mathrm{Var}(\Psi(\boldsymbol{r}_+)) + \frac{1}{2}\mathrm{Var}(\Psi(\boldsymbol{r}_-))$. The trained neural network's estimate $\hat{s}_{\mathrm{brain}}$ has a variance $\sigma^2_{\hat{s}_{\mathrm{brain}}}$ near the reference

stimulus s_0. Based on these quantities, Figure 4 shows that we can successfully identify that the brain is near-optimal.

S.3.2 Decoding efficiency

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 [27]:

$$C_{R_k} = \frac{(\Gamma \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} = \frac{(\Gamma_0 \boldsymbol{w} + \frac{1}{J_\infty} \boldsymbol{F}' \boldsymbol{F}'^\top \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} \tag{100}$$

For unbiased decoding, $\boldsymbol{w}^\top \boldsymbol{F}' = 1$. Some manipulation gives [27]

$$C_{R_k} = \frac{(\Gamma_0 \boldsymbol{w})_k}{\Gamma_{0k} \sigma_{0\hat{s}}} \frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} \frac{\Gamma_{0k}}{\Gamma_k} + \frac{F'_k}{\sigma_k} \sigma_{\hat{s}} \frac{1/J_\infty}{\sigma_{\hat{s}}^2} \tag{101}$$

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:

$$\sigma_{\hat{s}}^2 = \boldsymbol{w}^\top \Gamma \boldsymbol{w} = \boldsymbol{w}^\top \Gamma \boldsymbol{w} + \frac{1}{J_\infty} \boldsymbol{w}^\top \boldsymbol{F}' \boldsymbol{F}'^\top \boldsymbol{w} \tag{102}$$

$$= \sigma_{0\hat{s}}^2 + 1/J_\infty \tag{103}$$

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

$$\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} = \sqrt{1 - \frac{1/J_\infty}{\sigma_{\hat{s}}^2}} = \sqrt{1-\alpha} \tag{104}$$

with $\alpha = \frac{1/J_\infty}{\sigma_{\hat{s}}^2}$. Substituting these into (Eq 101) 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:

$$C_R^{\text{sub}} \approx \alpha C_R^{\text{opt}} + C_R^{\text{sub}} \sqrt{1-\alpha} \tag{105}$$

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.

S.4 Choice correlations from internal versus 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 ν,

$$\boldsymbol{R}(\boldsymbol{r}) = \boldsymbol{F}' s + \boldsymbol{G}' \nu + \boldsymbol{\eta} \tag{106}$$

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

$$\Gamma = H + \xi \boldsymbol{G}' \boldsymbol{G}'^\top \tag{107}$$

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

$$\boldsymbol{w} = \frac{\Gamma^{-1} \boldsymbol{F}'}{J_1} \tag{108}$$

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 $p(\nu)$. For distributions in the exponential family, this information saturates the Cramer-Rao bound on an estimator's variance, so that $J_1 = 1/\sigma_{\hat{s}}^2$. [72] The normalization by J_1 ensures the decoding is locally unbiased. These weights are used to estimate the stimulus according to

$$\hat{s} = \boldsymbol{w}^\top \boldsymbol{R}(\boldsymbol{r}) + b \tag{109}$$

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

$$c_{R_k}^{\text{sub}} = \text{Corr}(R_k, \hat{s}|s, \nu) \tag{110}$$

We include the superscript c^{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:

$$c_{R_k}^{\text{sub}} = \frac{\text{Cov}(R_k, \hat{s}|s, \nu)}{\sigma_{R_k|s,n} \sigma_{\hat{s}|s,n}} \tag{111}$$

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

$$\text{Cov}(\boldsymbol{R}, \hat{s}|s, \nu) = \langle \boldsymbol{R}\hat{s}|s, n\rangle \tag{112}$$

$$= \langle \boldsymbol{R}\boldsymbol{R}^\top|s, n\rangle \boldsymbol{w} \tag{113}$$

$$= \frac{H\Gamma^{-1}\boldsymbol{F}'}{J_1} \tag{114}$$

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:

$$\Gamma^{-1} = (H + \xi \boldsymbol{G}'\boldsymbol{G}'^\top)^{-1} \tag{115}$$

$$= H^{-1} - \frac{H^{-1}\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}}{1/\xi + \boldsymbol{G}'^\top H^{-1}\boldsymbol{G}'} \tag{116}$$

Substituting this inverse covariance into Equation 112, we obtain

$$\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s,\nu) \tag{117}$$

$$= \frac{1}{J_1} H (H^{-1} - \frac{H^{-1}\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'^\top})\boldsymbol{F}' \tag{118}$$

$$= \frac{1}{J_1}(\boldsymbol{F}' - \frac{\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}\boldsymbol{F}'}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'^\top}) \tag{119}$$

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, $(\hat{s}, \hat{\nu})$:

$$\boldsymbol{J}(s,\nu) = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}'^\top H^{-1}\boldsymbol{F}' & \boldsymbol{F}'^\top H^{-1}\boldsymbol{G}' \\ \boldsymbol{G}'^\top H^{-1}\boldsymbol{F}' & \boldsymbol{G}'^\top H^{-1}\boldsymbol{G}' \end{bmatrix} \tag{120}$$

With these substitutions, we have

$$\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s,\nu) = \frac{1}{J_1}\left(\boldsymbol{F}' - \frac{J_{12}}{1/\xi + J_{22}}\boldsymbol{G}'\right) \tag{121}$$

The denominator of Equation 111 involves the variance of the sufficient statistics,

$$\sigma^2_{R_k|s,n} = H_{kk} \tag{122}$$

and the variance of the brain's decoder,

$$\sigma^2_{\hat{s}} = \boldsymbol{w}^\top H \boldsymbol{w}$$

$$= \boldsymbol{w}^\top(\Gamma - \xi\boldsymbol{G}'\boldsymbol{G}'^\top)\boldsymbol{w}$$

$$= \frac{1}{J_1} - \frac{J_{12}^2}{\xi J_1^2}\frac{1}{(1/\xi + J_{22})^2} \tag{123}$$

where we used the following results:

$$\boldsymbol{w}^\top \boldsymbol{G}'\boldsymbol{G}'^\top \boldsymbol{w} = \left(\frac{\boldsymbol{F}'\Gamma^{-1}}{J_1}\boldsymbol{G}'\right)^2$$

$$= \frac{1}{J_1^2}\left(\boldsymbol{F}' H^{-1}\boldsymbol{G}' - \frac{\boldsymbol{F}'\Gamma^{-1}\boldsymbol{G}'\boldsymbol{G}' H^{-1}\boldsymbol{G}'}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'}\right)^2$$

$$= \frac{1}{J_1^2}\left(J_{12} - \frac{J_{12}J_{22}}{1/\xi + J_{22}}\right)^2$$

$$= \frac{J_{12}^2}{\xi^2 J_1^2}\frac{1}{(1/\xi + J_{22})^2} \tag{124}$$

Combining the results from Equation 121, 123 and 122, we can compute Equation 111

$$c^{\mathrm{sub}}_{R_k} = \mathrm{Corr}(R_k, \hat{s}|s,\nu)$$

$$= \frac{\mathrm{Cov}(R_k, \hat{s}|s,\nu)}{\sigma_{R_k|s,n}\sigma_{\hat{s}|s,n}}$$

$$= \frac{\frac{1}{J_1}\left(F'_k - \frac{J_{12}}{1/\xi + J_{22}}G'_k\right)}{\sqrt{H_{kk}}\sigma_{\hat{s}|s,n}} \tag{125}$$

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

$$C^{\mathrm{opt}}_{R_k} = \sqrt{\frac{J_{1,R_k}}{J_1}} = \frac{F'_k}{\sigma_{R_k|s}\sqrt{J_1}} \tag{126}$$

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: $C^{\mathrm{opt}}_{R_k}$ and $c^{\mathrm{sub}}_{R_k}$. We find that the nuisance-conditioned choice correlations $c^{\mathrm{sub}}_{R_k}$ relate to the optimal nuisance-averaged choice correlations $C^{\mathrm{opt}}_{R_k}$ according to

$$c^{\mathrm{sub}}_{R_k} = \beta_k C^{\mathrm{opt}}_{R_k} - \gamma_k \tag{127}$$

where we have defined the following constants:

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

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

$$= \sqrt{\frac{H_{kk} + \xi G'^2_k}{H_{kk}}}\frac{1}{\sqrt{1 - \frac{J_{12}^2}{\xi J_1}\frac{1}{(1/\xi + J_{22})^2}}}, \tag{128}$$

and

$$\gamma_k = \frac{G'_k}{\sqrt{H_{kk}}}\frac{J_{12}}{(1/\xi + J_2)J_1\sigma_{\hat{s}|s,n}} \tag{129}$$

The slope β_k and offset γ_k of the relationship between these two types of choice correlations (Equation 127) 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 S2A). 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 S2B).

For the simulations in Figure S2, 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 $H = I$: $p(\boldsymbol{r}|s,\nu) = \mathcal{N}(\boldsymbol{F}'s + \boldsymbol{G}'\nu, I)$. In Figure S2A, $\boldsymbol{F}'$ 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 S2B, 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 108, and the stimulus is estimated using Equation 109. Choice correlations in this fixed-nuisance experiment are computed by Equation 110 (vertical axis in Figure S2). The predicted optimal choice correlation is computed by Equation 126 (horizontal axis in Figure S2). In this setting, $\beta_k \approx 1$ when $J_{12} = 0$.

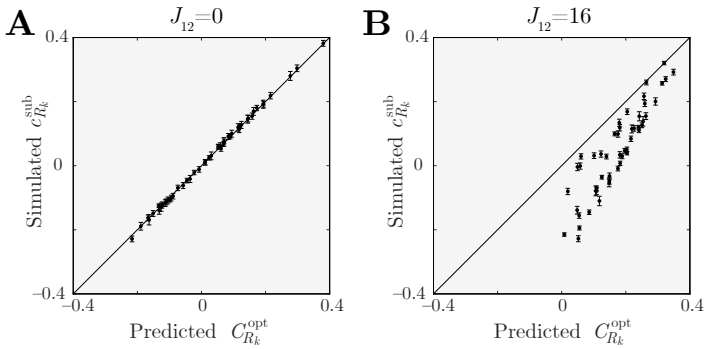

Figure S2: Comparing choice correlations caused by internal and external noise. (**A**) When estimates of nuisance variables are independent of estimates of task-relevant signals, the optimal choice correlations driven by internal noise, $c^{\text{sub}}_{R_k}$, match the optimal pattern $C^{\text{opt}}_{R_k}$ 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.

S.5 Coarse discrimination and choice correlations

We now derive a relationship between nonlinear neural thresholds and nonlinear choice correlations for *coarse* binary discrimination tasks, choosing between stimulus s_+ and s_-. The main ideas are the same as for fine discrimination, but there are a few more subtleties involved when the statistical structure of the response depends on the stimulus.

We assume the brain decodes neural activity $\boldsymbol{r}$ as a linear weighted sum of nonlinear statistics $\boldsymbol{R}(\boldsymbol{r})$, using weights given by linear regression as

$$\boldsymbol{w} \propto \mathrm{Cov}(\boldsymbol{R})^{-1}\mathrm{Cov}(\boldsymbol{R}, s) \tag{130}$$

The latter factor reflects the signal strength,

$$\mathrm{Cov}(\boldsymbol{R}, s) = \langle \boldsymbol{R}s\rangle - \langle \boldsymbol{R}\rangle\langle s\rangle \tag{131}$$
$$= \tfrac{1}{2}(\boldsymbol{F}_+ - \boldsymbol{F}_-)ds = \tfrac{1}{2}\Delta\boldsymbol{F}\, ds \tag{132}$$

We assume that the two values $s_\pm = s_0 \pm \Delta s$ are equally probable, and notate the mean responses as $\boldsymbol{F}_\pm = \boldsymbol{F}(s_\pm) = \langle \boldsymbol{R}|s_\pm\rangle$. The factor $\mathrm{Cov}\,\boldsymbol{R}$ includes covariability induced by both signal and noise. Using the law of total covariance, these contributions can be separated as

$$\mathrm{Cov}\,\boldsymbol{R} = \langle \mathrm{Cov}(\boldsymbol{R}|s)\rangle_s + \mathrm{Cov}_s\,\langle \boldsymbol{R}|s\rangle \tag{133}$$
$$= \bar{\Gamma} + \tfrac{1}{4}\Delta\boldsymbol{F}\Delta\boldsymbol{F}^\top \tag{134}$$

where the first term is the average noise covariance across the stimulus ensemble, $\bar{\Gamma} = \langle \mathrm{Cov}(\boldsymbol{R}|s)\rangle_s$, and the second term reflects variance along the signal direction. As for fine discrimination, noise variance along the signal direction has no influence on the optimal readout direction, since it cannot be removed. Using the Sherman-Morrison formula, we find that the decoder is

$$\begin{aligned}\boldsymbol{w} &\propto \mathrm{Cov}(\boldsymbol{R})^{-1}\,\mathrm{Cov}(\boldsymbol{R}, s)\\ &= \left(\bar{\Gamma} + \tfrac{1}{4}\Delta\boldsymbol{F}\Delta\boldsymbol{F}^\top\right)^{-1}\tfrac{1}{2}\Delta\boldsymbol{F}\,\Delta s\\ &\propto \left(\bar{\Gamma}^{-1} - \frac{\frac{1}{4}\bar{\Gamma}^{-1}\Delta\boldsymbol{F}\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}}{1+\frac{1}{4}\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}\Delta\boldsymbol{F}}\right)\tfrac{1}{2}\Delta\boldsymbol{F}\\ &\propto \bar{\Gamma}^{-1}\Delta\boldsymbol{F}\end{aligned} \tag{135}$$

For unbiased decoding, the proportionality is given by $1/\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}\Delta\boldsymbol{F}$.

S.5.1 Average conditional choice correlations

The core desideratum for a measure of choice correlations is to isolate the non-stimulus fluctuations that correlate with choices. The typical way to ensure this is to measure correlations between neural responses and choices only when the stimulus is completely ambiguous, *i.e.* at the decision boundary. Other studies have sought to expand the range of stimuli that can be used for these correlations [38, 73]. Mathematically, we examine the statistical relationship between neural responses and choices that remains after *conditioning* on the stimulus, via $p(\boldsymbol{R}, \hat{s}|s)$. Here we quantify this relationship through a conditional covariance, $\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s)$. For coarse discrimination, the strength (and pattern) of this correlation may depend on the particular stimulus used. To account for this, we compute an average over possible stimuli, $\langle \mathrm{Cov}(R_k, \hat{s}|s)\rangle_s$. If we normalize by root mean variances, we obtain

$$B_{R_k} = \frac{\langle \mathrm{Cov}(R_k, \hat{s}|s)\rangle_s}{\sqrt{\langle \mathrm{Var}(R_k|s)\rangle_s\,\langle \mathrm{Var}(\hat{s}|s)\rangle_s}} \tag{136}$$

This nonlinear choice correlation can be rewritten as

$$\begin{aligned}B_{R_k} &= \frac{(\boldsymbol{w}^\top\,\langle \mathrm{Cov}(\boldsymbol{R}|s)\rangle_s)_k}{\sqrt{\bar{\Gamma}_{kk}\;\boldsymbol{w}^\top\langle \mathrm{Cov}(\boldsymbol{R}|s)\rangle_s\,\boldsymbol{w}}}\\ &= \frac{(\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}\bar{\Gamma})_k}{\sqrt{\bar{\Gamma}_{kk}\;\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}\bar{\Gamma}\bar{\Gamma}^{-1}\Delta\boldsymbol{F}}}\\ &= \frac{\Delta\boldsymbol{F}_k}{\sqrt{\bar{\Gamma}_{kk}\;\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}\Delta\boldsymbol{F}}}\end{aligned} \tag{137}$$

We recognize that this expression contains the ratio of sensitivities for the neural statistic R_k and the entire population $\boldsymbol{r}$ in coarse discrimination, $d'_k = \Delta\boldsymbol{F}_k/\sqrt{\bar{\Gamma}_{kk}}$ and $d' = \sqrt{\Delta\boldsymbol{F}^\top\bar{\Gamma}^{-1}\Delta\boldsymbol{F}}$. We therefore find the same result as for optimal fine discrimination (Eq 18):

$$B^{\text{opt}}_{R_k} = \frac{d'_k}{d'} \tag{138}$$

S.5.2 Signal estimation from total correlations

It is useful to express the discriminability through the total correlation between the responses and the stimulus,

$$\begin{aligned} D_{R_k,s} &= \mathrm{Corr}(R_k, s) \qquad (139)\\ &= \frac{\mathrm{Cov}(R_k, s)}{\sigma_{R_k}\sigma_s} \\ &= \frac{\frac{1}{2}\Delta F_k\, ds}{\sqrt{(\bar{\Gamma}_{kk} + \frac{1}{4}\Delta F_k^2)\sigma_s^2}} \\ &= \frac{1}{\sqrt{\frac{4\bar{\Gamma}_{kk}}{\Delta F_k^2} + 1}} \\ &= \frac{1}{\sqrt{4d'^{-2}_k + 1}} \qquad (140) \end{aligned}$$

In these equations we used the fact that for binary discrimination, the standard deviation of the signal is related to the difference between the two possible signal values, $\sigma_s = \frac{1}{2}(s_+ - s_-) = ds$. We can invert Eq 140 to find

$$d'_k = \frac{2}{\sqrt{D^{-2}_{R_k,s} - 1}} \qquad (141)$$

This dependence is plotted in Figure S3.

Similarly, we can express the behavioral discriminability d' in terms of the correlation between the estimate and the stimulus, $D_{\hat{s},s} = \mathrm{Corr}(\hat{s}, s)$:

$$d' = \frac{2}{\sqrt{D^{-2}_{\hat{s},s} - 1}} \qquad (142)$$

The relationship between discriminability d' and total correlation D is linear when D is relatively small. Thus we can approximate the optimal nonlinear choice correlation as:

$$B^{\mathrm{opt}}_{R_k} \approx \frac{D_{R,s}}{D_{\hat{s},s}} \qquad (143)$$

Here the total correlation is computed based on a continuous estimate $\hat{s}$. When the behavioral outcome is a binary choice, this relationship is more complicated. Section S.6.3 calculates the relationship between $D_{\hat{s}}$ and $D_{\hat{s}_\pm}$ for one particular task.

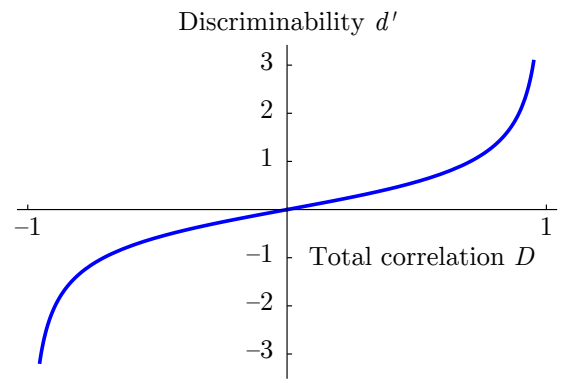

Figure S3: Stimulus discriminability d' for a response variable R, versus total correlation between that variable and the stimulus, $D = \mathrm{Corr}\,(R, s)$, according to Eq 141.

S.6 Orientation variance discrimination task

S.6.1 Coarse tasks: Continuous estimation versus binary discrimination

The experiment of Section 2.3 defines an orientation variance discrimination task in which the relevant statistics are quadratic functions of the orientation. The quadratic decoding model described in the main text could suffice for this problem. However, in our case the variances to be distinguished are quite different, such that the nuisance variation differs substantially between these two stimulus categories. As described in Methods 4.2.1, coarse tasks with stimulus-dependent variability generate a slightly different prediction compared to fine tasks (or coarse tasks with stimulus-independent variability).

Moreover, there are minor differences between the predictions for continuous estimation and binary discrimination, and these differences are more complicated for coarse tasks than fine ones. Here we describe in detail the somewhat lengthy computation of the ratio ζ between choice correlations for continuous quadratic estimation and binary quadratic decoding. For coarse discrimination, the ratio ζ will depend on the input statistics and threshold, but for fine discrimination ζ becomes a constant. Regardless, for our cases of interest these numbers are generally near 1.

We begin by assuming that the variance estimate is the square of the orientation estimate $\hat{s} = \hat{\nu}^2$, and a binary guess about the variance is given by $\hat{s}_\pm = \mathrm{sgn}\,(\hat{\nu}^2 - \theta^2)$ where θ is the animal's orientation threshold. We assume that $\hat{\nu}$ is an unbiased estimate of the orientation ν, so $\langle\hat{\nu}|\nu\rangle = \nu$. We denote one neuron's mean response to the orientation by $\langle r|\nu\rangle = \mu(\nu)$ which we approximate linearly as $\mu(\nu) \approx \bar{\mu} + \mu'\nu$ with $\bar{\mu} = \mu(0)$. The mean behavioral choice is $\langle\hat{s}_\pm|s\rangle = m_s$. Since the stimulus is binary, we will denote this mean with a subscript, $\langle\hat{s}_\pm|s_+\rangle = m_+$ or $\langle\hat{s}_\pm|s_-\rangle = m_-$.

The joint distribution $p(r, \hat{\nu}|s)$ arises from both internal noise and nuisance variation, $p(r, \hat{\nu}|s) = \int d\nu\, p(r, \hat{\nu}|\nu)p(\nu|s)$. For a given orientation ν, the neural response r and orientation estimate $\hat{\nu}$ follow a bivariate normal distribution,

$$p(r, \hat{\nu}|\nu) = \mathcal{N}\left(\begin{matrix} r \\ \hat{\nu} \end{matrix}\middle|\begin{matrix} \mu(\nu) \\ \nu \end{matrix}; \begin{bmatrix} H_{rr|\nu} & H_{r\hat{\nu}|\nu} \\ H_{r\hat{\nu}|\nu} & H_{\hat{\nu}\hat{\nu}|\nu} \end{bmatrix}\right) \qquad (144)$$

which summarizes all of the internal noise given the sensory input.

By design, the nuisance variable ν is normally distributed, $p(\nu|s) = \mathcal{N}(\nu|0, s)$, so we can write the marginal distribution $p(r, \hat{\nu}|s)$ as

$$\begin{aligned} p(r, \hat{\nu}|s) &= \mathcal{N}\left(\begin{matrix} r \\ \hat{\nu} \end{matrix}\middle|\begin{matrix} \bar{\mu} \\ 0 \end{matrix}; \begin{bmatrix} H_{rr|\nu} + \mu'^2 s & H_{r\hat{\nu}|\nu} + \mu' s \\ H_{r\hat{\nu}|\nu} + \mu' s & H_{\hat{\nu}\hat{\nu}|\nu} + s \end{bmatrix}\right) \\ &= \mathcal{N}(\boldsymbol{\mu}(s), \Sigma(s)) \qquad (145) \end{aligned}$$

For now we suppress the explicit dependence on s.

The conditional covariance between the nonlinear statistic R

and choice is

$$\begin{aligned}&\mathrm{Cov}(\hat{s}_\pm, R|s) \\ &\quad = \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2) r^2 \right\rangle_{r,\hat{\nu}} - \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\right\rangle_{\hat{\nu}} \left\langle r^2 \right\rangle_r\end{aligned} \tag{146}$$

where $R = r^2$ and we reiterate that we are suppressing the conditioning on s. The second moment is

$$\begin{aligned}&\left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2) r^2 \right\rangle_{r,\hat{\nu}} \\ &= \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2) \left\langle r^2|\hat{\nu}\right\rangle_{r|\hat{\nu}} \right\rangle_{\hat{\nu}} \\ &= \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)(\Sigma_{r|\hat{\nu}} + \mu^2_{r|\hat{\nu}})\right\rangle_{\hat{\nu}} \\ &= \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2) \left[\Sigma_{rr} - \frac{\Sigma^2_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}} + \left(\mu + \frac{\Sigma_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}}\hat{\nu}\right)^2\right]\right\rangle_{\hat{\nu}}\end{aligned} \tag{147}$$

where we used the conditional distribution

$$p(r|\hat{\nu}) = \mathcal{N}\left(r \middle| \mu + \frac{\Sigma_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}}\hat{\nu},\, \Sigma_{rr} - \frac{\Sigma^2_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}}\right) \tag{148}$$

This can be written as

$$\left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)(a\hat{\nu}^2 + b\hat{\nu} + c)\right\rangle_{\hat{\nu}} \tag{149}$$

for coefficients

$$a = \frac{\Sigma^2_{r\hat{\nu}}}{\Sigma^2_{\hat{\nu}\hat{\nu}}} \tag{150}$$

$$b = 2\frac{\Sigma_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}}\mu \tag{151}$$

$$c = \Sigma_{rr} - \frac{\Sigma^2_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}} + \mu^2 \tag{152}$$

Note that this is an expectation over $\hat{\nu}$ only. Such an expected value can be written as a sum of integrals:

$$\begin{aligned}&\left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\hat{\nu}^\alpha\right\rangle_{\hat{\nu}} \\ &= \left[\int_{-\infty}^{-\theta} - \int_{-\theta}^{\theta} + \int_{\theta}^{\infty}\right]\hat{\nu}^\alpha p(\hat{\nu})d\hat{\nu} \\ &= \left[\int_{-\infty}^{-\theta} - \left(\int_{-\infty}^{\theta} - \int_{-\infty}^{-\theta}\right) + \left(\int_{-\infty}^{\infty} - \int_{-\infty}^{\theta}\right)\right]\hat{\nu}^\alpha p(\hat{\nu})d\hat{\nu} \\ &= \left[2\int_{-\infty}^{-\theta} - 2\int_{-\infty}^{\theta} + \int_{-\infty}^{\infty}\right]\hat{\nu}^\alpha p(\hat{\nu})d\hat{\nu}\end{aligned} \tag{153}$$

These integrals can be expressed in terms of error functions, where $\sigma^2_{\hat{\nu}}$ is the marginal variance for $p(\hat{\nu}|s)$:

$$\int_{-\infty}^{\theta} d\hat{\nu}\, \hat{\nu}^0 \mathcal{N}(\hat{\nu}|0,\sigma^2_{\hat{\nu}}) = \tfrac{1}{2}\,\mathrm{erfc}\left(\tfrac{\theta}{\sqrt{2}\sigma_{\hat{\nu}}}\right) \tag{154}$$

$$\int_{-\infty}^{\theta} d\hat{\nu}\, \hat{\nu}^1 \mathcal{N}(\hat{\nu}|0,\sigma^2_{\hat{\nu}}) = -p_{\hat{\nu}}(\theta)\sigma^2_{\hat{\nu}} \tag{155}$$

$$\int_{-\infty}^{\theta} d\hat{\nu}\, \hat{\nu}^2 \mathcal{N}(\hat{\nu}|0,\sigma^2_{\hat{\nu}}) = \tfrac{1}{2}\,\mathrm{erfc}\left(\tfrac{\theta}{\sqrt{2}\sigma_{\hat{\nu}}}\right)\sigma^2_{\hat{\nu}} - p_{\hat{\nu}}(\theta)\sigma^2_{\hat{\nu}}\theta \tag{156}$$

Note that $p_{\hat{\nu}}(\theta)$ has units of $[\nu]^{-1}$, so units are consistent across these expressions.

Combining these with Eq 153 we obtain

$$m = \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\right\rangle = 2\,\mathrm{erfc}\left(\tfrac{\theta}{\sqrt{2}\sigma}\right) - 1 \tag{157}$$

$$\left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\hat{\nu}\right\rangle = 0 \tag{158}$$

$$\left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\hat{\nu}^2\right\rangle = \sigma^2_{\hat{\nu}} m + 4\theta\sigma^2_{\hat{\nu}} p_{\hat{\nu}}(\theta) \tag{159}$$

where we have used the identity $\mathrm{erfc}(-x) = 2 - \mathrm{erfc}(x)$ and the symmetry $p_{\hat{\nu}}(\theta) = p_{\hat{\nu}}(-\theta)$. The first term, m, is the mean of $\hat{s}_\pm$, and will appear several times in the equations below.

Returning to Eq 147, we have

$$\begin{aligned}&\mathrm{Cov}(\hat{s}_\pm, R|s) \\ &\quad = \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2) r^2 \right\rangle_{r,\hat{\nu}} - \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\right\rangle_{\hat{\nu}|s} \left\langle r^2 \right\rangle_r \\ &\quad = \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)(a\hat{\nu}^2 + b\hat{\nu} + c)\right\rangle_{\hat{\nu}} \\ &\qquad - \left\langle \mathrm{sgn}\,(\hat{\nu}^2-\theta^2)\right\rangle_{\hat{\nu}} \left\langle r^2 \right\rangle_r \\ &\quad = a\left[\sigma^2_{\hat{\nu}} m + 4\theta\sigma^2_{\hat{\nu}} p_{\hat{\nu}}(\theta)\right] m + cm - m(\Sigma_{rr} + \mu^2) \\ &\quad = 4\theta\frac{\Sigma^2_{r\hat{\nu}}}{\Sigma_{\hat{\nu}\hat{\nu}}} p_{\hat{\nu}}(\theta)\end{aligned} \tag{160}$$

Note that all of the erfc terms have canceled.

Compare that to the corresponding covariance for continuous estimation,

$$\mathrm{Cov}\,(\hat{s}, R|s) = 2\Sigma^2_{r\hat{\nu}} \tag{161}$$

The conditional variance of a binary output $\hat{s} = \pm 1$ is simply

$$\mathrm{Var}(\hat{s}_\pm|s) = 1 - \langle \hat{s}_\pm|s\rangle^2 \tag{162}$$

$$= 1 - m^2 \tag{163}$$

whereas, the variance for the continuous estimator is

$$\mathrm{Var}(\hat{s}|s) = \left\langle (\hat{\nu}^2-\theta^2)^2|s\right\rangle - \left\langle \hat{\nu}^2-\theta^2|s\right\rangle^2 \tag{164}$$

$$= 2\Sigma^2_{\hat{\nu}\hat{\nu}} \tag{165}$$

The variance of r^2, $\mathrm{Var}\,(r^2|s) = 2\Sigma^2_{rr} + 4\Sigma_{rr}\mu^2$, is the same whether the behavioral estimate is continuous or binary.

Our goal here is to compute the change in our measure of nonlinear choice correlation, namely,

$$\zeta = \frac{B^\pm_R}{B_R} = \frac{\frac{\left\langle \mathrm{Cov}(\hat{s}_\pm, R|s)\right\rangle_s}{\sqrt{\left\langle \mathrm{Var}(\hat{s}_\pm|s)\right\rangle_s \langle \mathrm{Var}(R|s)\rangle_s}}}{\frac{\langle \mathrm{Cov}(\hat{s}, R|s)\rangle_s}{\sqrt{\langle \mathrm{Var}(\hat{s}|s)\rangle_s \langle \mathrm{Var}(R|s)\rangle_s}}} \tag{166}$$

where the averages over $p(s) = 1/2$ include equal proportions of the binary stimuli s_+ and s_-. Substituting our calculations above, and reintroducing the dependencies on s, we find

$$\zeta = \frac{\langle \mathrm{Cov}\,(\hat{s}_\pm, R|s)\rangle_s}{\langle \mathrm{Cov}\,(\hat{s}, R|s)\rangle_s}\sqrt{\frac{\langle \mathrm{Var}\,(\hat{s}|s)\rangle_s}{\langle \mathrm{Var}\,(\hat{s}_\pm|s)\rangle_s}} \tag{167}$$

$$= \frac{\sum_s p(s) 4\theta \frac{\Sigma^2_{r\hat{\nu}|s}}{\Sigma_{\hat{\nu}\hat{\nu}|s}} p_{\hat{\nu}}(\theta|s)}{\sum_s p(s) 2\Sigma^2_{r\hat{\nu}|s}}\sqrt{\frac{\sum_s p(s) 2\Sigma^2_{\hat{\nu}\hat{\nu}|s}}{\sum_s p(s)(1-m^2_s)}} \tag{168}$$

For tasks where the variability is dominated by external nuisance variables rather than by internal noise, i.e. $H \ll s(\mu', 1)(\mu', 1)^\top$, we can approximate the covariances by $\Sigma_{rr} \approx \mu'^2 s$, $\Sigma_{r\hat{\nu}} \approx \mu' s$, and $\Sigma_{\hat{\nu}\hat{\nu}} \approx s$. Substituting these approximations into the expression above, we obtain

$$\zeta \approx \frac{\frac{1}{2}\sum_s 4\theta \frac{\mu'^2 s^2}{s} \frac{e^{-\theta^2/2s}}{\sqrt{2\pi s}}}{\frac{1}{2}\sum_s 2\mu'^2 s^2}\sqrt{\frac{\frac{1}{2}\sum_s 2s^2}{1 - \frac{1}{2}\sum_s m^2_s}} \tag{169}$$

In our task conditions, $3 = \sqrt{s_-} \ll \sqrt{s_+} = 15$, so some terms dominate in the sums. Moreover, we assume that the threshold θ

lies far enough between $\sqrt{s_-} < \theta < \sqrt{s_+}$ that $e^{-\theta^2/2s_-} \approx 0$ and $e^{-\theta^2/2s_+} \approx 1$. We then find

$$\zeta \approx \frac{\frac{1}{2}4\theta\sqrt{\frac{s_+}{2\pi}}}{\frac{1}{2}2s_+^2}\sqrt{\frac{\frac{1}{2}2s_+^2}{1-\frac{1}{2}\sum_s m_s^2}} \tag{170}$$

$$= \frac{2}{\sqrt{\pi}}\frac{\theta}{\sqrt{s_+}}\frac{1}{\sqrt{1-\frac{1}{2}\sum_s m_s^2}} \tag{171}$$

Empirically, we find that $1 - \langle m^2 \rangle_s \approx \frac{1}{2}$ (Figure S4). In that case, we obtain

$$\zeta \approx \frac{2\theta}{\sqrt{\pi s_+}} \tag{172}$$

This expression is independent of the statistics of r. Therefore the same correction factor holds for cross-terms like $r_j r_k$, which can be expressed as linear combination of squares, $R_{jk} = r_j r_k = \frac{1}{2}(r_j + r_k)^2 - \frac{1}{2}r_j^2 - \frac{1}{2}r_k^2$. We use this correction factor ζ to adjust our predicted quadratic choice correlations in Figure 6.

To find the behavioral threshold θ for Eq 172, we used logistic regression of choice $\hat{s}_\pm$ on the absolute value of the stimulus orientation, $|\nu|$, and assign the threshold θ to be the orientation where the probability of both choices was equal.

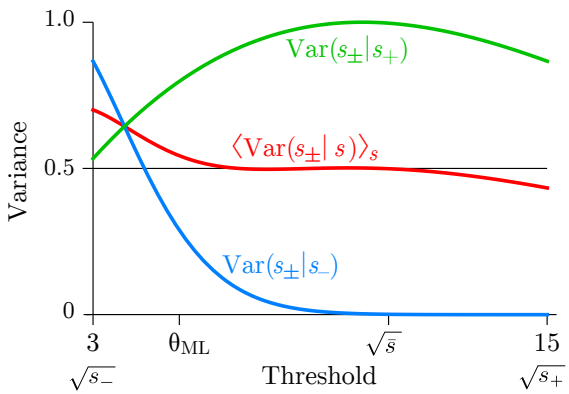

Figure S4: The average variance of $\hat{s}_\pm$ conditioned on the stimulus s (red) is approximately 1/2 over a wide range of thresholds.

S.6.2 Fine tasks: Continuous estimation versus binary discrimination

For fine discrimination, the stimulus s is effectively constant, so we need not take averages.

$$\zeta^{\text{fine}} = \frac{\mathrm{Cov}\,(\hat{s}_\pm, R|s)}{\mathrm{Cov}\,(\hat{s}, R|s)}\sqrt{\frac{\mathrm{Var}\,(\hat{s}|s)}{\mathrm{Var}\,(\hat{s}_\pm|s)}} \tag{173}$$

$$= \frac{4\theta\frac{\Sigma^2_{r\hat{\nu}|s}}{\Sigma_{\hat{\nu}\hat{\nu}|s}}p_{\hat{\nu}}(\theta|s)}{2\Sigma^2_{r\hat{\nu}|s}}\sqrt{\frac{2\Sigma^2_{\hat{\nu}\hat{\nu}|s}}{1-m_s^2}} \tag{174}$$

After several cancelations, and using the fact that for fine discrimination, $\theta = s \approx s_+ \approx s_-$, we find

$$\zeta^{\text{fine}} = \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}}\sqrt{\frac{8}{1-(2\,\mathrm{erfc}(\frac{1}{\sqrt{2}})-1)^2}} \tag{175}$$

$$\approx 0.735 \tag{176}$$

Observe that for fine discrimination, the ratio ζ is a constant, independent of the underlying statistics.

S.6.3 Total correlation for binary and continuous estimates

We showed in Methods 4.2.1 that the discriminability is related to the total correlation between signal and response. However, those relationships were based on continuous estimates of the binary stimulus. As above, when the behavioral choice is also binary, we can adjust the calculation slightly. Here we compare the total correlations for continuous and binary response, $D_{\hat{s},s}$ and $D_{\hat{s}_\pm,s}$.

$$\begin{aligned}
\mathrm{Cov}\,(\hat{s}, s) &= \langle\langle \hat{s}|s\rangle\, s\rangle_s - \langle\hat{s}\rangle\,\langle s\rangle \\
&= \left\langle\left\langle \hat{\nu}^2|s\right\rangle s\right\rangle_s - \left\langle\left\langle \hat{\nu}^2|s\right\rangle\right\rangle_s \bar{s} \\
&= \left\langle \Sigma_{\hat{\nu}\hat{\nu}|s} s\right\rangle_s - \langle\langle \Sigma_{\hat{\nu}\hat{\nu}}|s\rangle\rangle_s\,\bar{s} \\
&= \langle (H_{\hat{\nu}\hat{\nu}} + s)s\rangle_s - \langle (H_{\hat{\nu}\hat{\nu}} + s)\rangle_s\,\bar{s} \\
&= H_{\hat{\nu}\hat{\nu}}\bar{s} + \left\langle s^2\right\rangle_s - (H_{\hat{\nu}\hat{\nu}} + \bar{s})\bar{s} \\
&= \mathrm{Var}(s) \\
&= \tfrac{1}{2}(s_+^2 + s_-^2) - \tfrac{1}{4}(s_+ + s_-)^2 \\
&= \tfrac{1}{4}\Delta s^2
\end{aligned} \tag{177}$$

In contrast, the total covariance of $\hat{s}_\pm$ is

$$\begin{aligned}
\mathrm{Cov}\,(\hat{s}_\pm, s) &= \langle\langle \hat{s}_\pm|s\rangle\, s\rangle_s - \langle\hat{s}_\pm\rangle\,\langle s\rangle \\
&= \langle m_s s\rangle_s - \langle m_s\rangle_s\,\bar{s} \\
&= \tfrac{1}{2}\left(m(s_+)s_+ + m(s_-)s_-\right) \\
&\quad - \tfrac{1}{4}\left(m(s_+) + m(s_-)\right)\bar{s} \\
&= \tfrac{1}{4}\Delta m \Delta s
\end{aligned} \tag{178}$$

The total variance of $\hat{s}$ is

$$\begin{aligned}
\mathrm{Var}\,(\hat{s}) &= \left\langle \hat{\nu}^4\right\rangle - \left\langle \hat{\nu}^2\right\rangle^2 \\
&= \left\langle\left\langle \hat{\nu}^4|s\right\rangle\right\rangle_s - \left\langle\left\langle \hat{\nu}^2|s\right\rangle\right\rangle_s^2 \\
&= \left\langle 3\Sigma^2_{\hat{\nu}\hat{\nu}|s}\right\rangle_s - \left\langle \Sigma_{\hat{\nu}\hat{\nu}|s}\right\rangle_s^2 \\
&= \tfrac{3}{2}\left(\Sigma^2_{\hat{\nu}\hat{\nu}|+} + \Sigma^2_{\hat{\nu}\hat{\nu}|-}\right) - \left(\tfrac{1}{2}\left(\Sigma_{\hat{\nu}\hat{\nu}|+} + \Sigma_{\hat{\nu}\hat{\nu}|-}\right)\right)^2 \\
&= \tfrac{1}{4}\left(5\Sigma^2_{\hat{\nu}\hat{\nu}|+} - 2\Sigma_{\hat{\nu}\hat{\nu}|+}\Sigma_{\hat{\nu}\hat{\nu}|-} + 5\Sigma^2_{\hat{\nu}\hat{\nu}|-}\right) \\
&= \tfrac{1}{4}(5(H_{\hat{\nu}\hat{\nu}} + s_+)^2 - 2(H_{\hat{\nu}\hat{\nu}} + s_+)(H_{\hat{\nu}\hat{\nu}} + s_-) \\
&\quad + 5(H_{\hat{\nu}\hat{\nu}} + s_-)^2) \\
&= \tfrac{1}{4}(8H^2_{\hat{\nu}\hat{\nu}} + H_{\hat{\nu}\hat{\nu}}\,(10s_+ - 2s_+ - 2s_- + 10s_-) \\
&\quad + 5s_+^2 - 2s_+s_- + 5s_-^2) \\
&= 2H^2_{\hat{\nu}\hat{\nu}} + 2H_{\hat{\nu}\hat{\nu}}(s_+ + s_-) \\
&\quad + \tfrac{1}{4}\left(5s_+^2 - 2s_+s_- + 5s_-^2\right)
\end{aligned} \tag{179}$$

In the limit where the nuisance variability dominates the internal variability, and $s_+ \gg s_-$, this simplifies to

$$\mathrm{Var}\,(\hat{s}) \approx \tfrac{5}{4}s_+^2 \tag{180}$$

The total variance of $\hat{s}_\pm$ is

$$\mathrm{Var}\,(\hat{s}_\pm) = 1 - \langle\hat{s}_\pm\rangle^2 = 1 - \bar{m}^2 \tag{181}$$

where $\bar{m} = \frac{1}{2}(m_+ + m_-)$.

Combining these computations, we see that ratio of total cor-

relations for binary $\hat{s}_\pm$ and continuous $\hat{s}$ is

$$\begin{aligned}
\delta &= \frac{D_{\hat{s},s}}{D_{\hat{s}_\pm,s}} \\
&= \frac{\mathrm{Corr}(\hat{s}_\pm, s)}{\mathrm{Corr}\,(\hat{s}, s)} \\
&= \frac{\mathrm{Cov}(\hat{s}_\pm, s)}{\mathrm{Cov}\,(\hat{s}, s)} \sqrt{\frac{\mathrm{Var}\,(\hat{s})}{\mathrm{Var}\,(\hat{s}_\pm)}} \\
&\approx \sqrt{\tfrac{5}{4}} \frac{1}{1 - \frac{s_-}{s_+}} \frac{\Delta m}{\sqrt{1 - \tilde{m}^2}}
\end{aligned} \tag{182}$$

All of these quantities are measurable from data or are given by the task.

S.6.4 Optimal binary nonlinear coarse choice correlations

We can now combine our results above to create a prediction for optimal binary nonlinear coarse choice correlations. From Eq 143 and Eq 166, we have

$$B_R^{\mathrm{opt},\pm} = \zeta \frac{D_{R,s}}{D_{\hat{s},s}} \tag{183}$$

From Eq 182 we can adjust the

$$D_{\hat{s},s} = \delta D_{\hat{s}_\pm,s} \tag{184}$$

Combining these we have

$$B_R^{\mathrm{opt},\pm} = \frac{\zeta}{\delta} \frac{D_{R,s}}{D_{\hat{s}_\pm,s}} \tag{185}$$

where ζ and δ are determined by experimentally measurable quantities. Their precise values depends on the monkey and the session, but the ratio is typically $\zeta/\delta \approx 0.62 \pm 0.33$. When plotting the data in Figure 6, we apply these corrections to each session before combining different sessions together.