nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-26793-9 OPEN

Revealing nonlinear neural decoding by analyzing choices

Qianli Yang [1,6,7], Edgar Walker[2,3], R. James Cotton [4,5], Andreas S. Tolias [1,2,3] & Xaq Pitkow [1,2,3]✉

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, describing 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. This relationship holds for optimal feedforward networks of modest complexity, when experiments are performed under natural nuisance variation. 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] Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. [2] Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA. [3] Baylor College of Medicine, Center for Neuroscience and Artificial Intelligence, Houston, TX, USA. [4] Shirley Ryan Ability Lab, Chicago, IL, USA. [5] Department of Physical Medicine and Rehabilitation, Northwestern University, Evanston, IL, USA. [6] Changzhou University, Aliyun School of Big Data, Changzhou, China. [7] Institute of Neuroscience, Key Laboratory of Primate Neurobiology, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing, China. ✉email: xaq@rice.edu

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 can be 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 covariability of this population responses[1,2], leading to substantially underestimated redundancy of large cortical populations. 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.

This paper makes four main contributions. First, it weaves together important concepts about tuning curves and nuisance variables, nonlinear computation, and redundant population codes, forming a general, unified description of feedforward encoding and decoding processes in the brain. This description is supported by intuitive explanations and concrete examples to illustrate how these concepts relate to each other and enrich familiar views of neural computation. Second, this paper provides a simple way of testing the hypothesis that the brain's decoding strategy is efficient, using a simple 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. Third, it establishes the technical details needed to apply this test in practical neuroscience experiments. Fourth, we apply this test to analyze V1 data from macaque monkeys, finding direct experimental evidence for optimal nonlinear decoding.

The "Results" section describes the main concepts, their formal connections, and applications. More specifically, the first sections introduce a framework for understanding nonlinear computation, including basic notation, internal and external (nuisance) noise and their effects on information content and formatting, and how this information can be isolated by nonlinear computation of the right statistics. Subsequent sections introduce a formalism for decoding, including notions of linear and nonlinear choice correlations, fine and coarse estimation tasks, and predictions about those correlations under optimal decoding. This section continues by describing how redundancy in the population responses appears as special high-order response statistics, and how they affect the predictions. The last sections present an experimental application of these ideas. A sketch of the details of our general predictions are presented in the "Methods" section, and are derived in full in the Supplement along with details of their application to specific models and our experimental data.

Results

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 (Fig. 1a). If an edge is presented to this model neuron, different rotation angles will change the overlap, producing a different mean. This neuron is then tuned to orientation.

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

Notice that stimuli aligned with the neuron's preferred orientation will generally elicit the highest or lowest response magnitude, depending on polarity. Edges 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 the primary visual cortex, which use squaring nonlinearities to achieve invariance to the polarity of an edge[4].

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

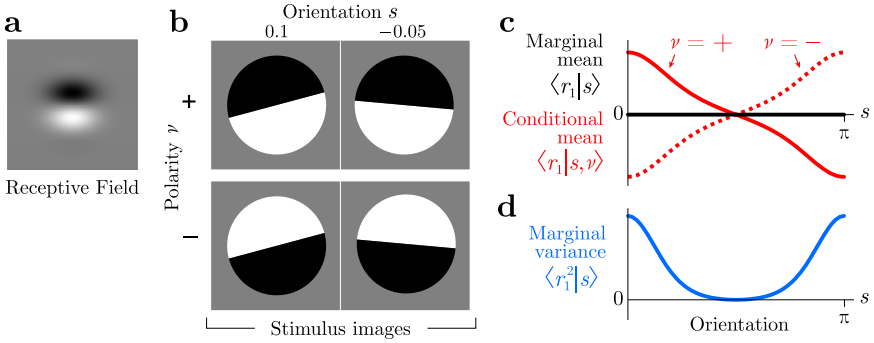

Fig. 1 Simple nonlinear code for orientation induced by two polarities. a Receptive field for a linear neuron. **b** Four example images, each with an orientation $s \in [0, \pi)$ and a polarity $\nu \in \{-1, +1\}$. **c** The mean response of the linear neuron is tuned to orientation if polarity were 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).

variables include the illuminant for guessing a surface color, position for object recognition, the 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[5–10]. For example, cells in the 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. 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.

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, v) and must act upon one particular relevant aspect of that stimulus, s, while ignoring the rest, v. The irrelevant stimulus aspects serve as nuisance variables for the task (v is the Greek letter 'nu' and here stands for nuisance). Together, these stimulus properties determine a complete sensory input that drives some responses $\mathbf{r}$ in a population of N neurons according to the distribution $p(\mathbf{r}|s, v)$.

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

$$(s, v) \rightarrow \mathbf{r} \rightarrow \mathbf{R} \rightarrow \hat{s} \quad (1)$$

We model the brain's estimate as a linear function of the downstream responses $\mathbf{R}$. 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 do not have direct access to $\mathbf{R}$, only a subset of $\mathbf{r}$. 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 fluctuations in $\mathbf{r}$, 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, $\sigma^2_{\hat{s}}$. In later sections, we then generalize the problem to allow for binary discrimination as well as coarse tasks, which are more complicated mathematically but not conceptually different.

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 the 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 $\mathbf{f}(s) = \sum_{\mathbf{r}} \mathbf{r}\, p(\mathbf{r}|s) = \langle \mathbf{r}|s \rangle$. (Angle brackets denote an average overall 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 $p(\mathbf{r}|s, v)$ 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—nuisance variables—leading to a neural response distribution $p(\mathbf{r}|s)$ 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 $\mathbf{f}(s)$ averaged over the set of stimuli s: $\rho_{\text{signal}} = \text{Corr}(\mathbf{f}(s))$ where again all averages are taken over all variables not fixed by a condition to the right of the vertical bar. (Internal) noise correlations measure shared variation that persists even when the stimulus is completely identical, nuisance variables and all: $\rho_{\text{noise}}(s, v) = \text{Corr}(\mathbf{r}|s, v)$.

For multidimensional stimuli, however, these correlations 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 v: $\rho_{\text{nuisance}}(s) = \text{Corr}(\mathbf{f}(s, v)|s)$. Including both internal and external (nuisance) noise correlations gives $\text{Corr}(\mathbf{r}|s)$.

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 (Fig. 2d, Table 1). Whether from internal or external noise, stimulus-dependent correlations lead naturally to nonlinear population codes, as we explain below.

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 signals 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[5,9], like for objects in the inferotemporal cortex.

If any of the relevant signals can only be extracted from nonlinear statistics of the neural responses[1,2], then we say that the population code is nonlinear (Table 1). One straightforward example is a stimulus-dependent covariance $Q(s) = \langle \mathbf{r}\mathbf{r}^\top|s \rangle$; its information can be decoded by quadratic operations $\mathbf{R} = \mathbf{r}\mathbf{r}^\top$ [28–30].

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

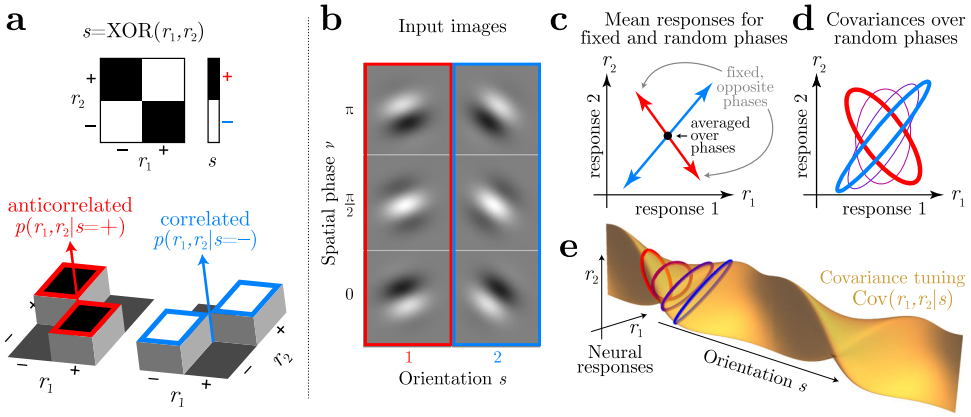

Fig. 2 Nonlinear codes. a Simple example in which a stimulus s is the XOR of two neural responses (top). Conditional probabilities $p(r_1, r_2|s)$ of those responses (bottom) show they are anti-correlated when $s = +1$ (red) and positively correlated when $s = -1$ (blue). This stimulus-dependent correlation between responses creates a nonlinear code. The remaining panels show that a similar stimulus-dependent correlation emerges in orientation discrimination with an 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 the 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.

Table 1 Neural response properties relevant for linear and nonlinear codes.

	Linear	Nonlinear	Quadratic
Trial data	$\mathbf{r}$	$\mathbf{R}(\mathbf{r})$	$\mathbf{r}\mathbf{r}^\top$
Signal	Mean($\mathbf{r}$|s)	Mean($\mathbf{R}$|s)	Mean($\mathbf{r}\mathbf{r}^\top$|s)
Noise	Cov($\mathbf{r}$|s)	Cov($\mathbf{R}$|s)	Cov($\mathbf{r}\mathbf{r}^\top$|s)

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).

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[36–38].

More typically, however, stimulus-dependent correlations arise from external noise, leading to what we call nuisance correlations. In the introduction (Fig. 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 (or any other statistic of the encoding model) on a single trial because it does not experience all possible nuisance variables to average over. However, in linear codes, the subject does not have access to the tuned mean response $\langle \mathbf{r}|s \rangle$ either, just a noisy single-trial version of the mean, namely $\mathbf{r}$. Analogously, the subject does not need access to the tuned covariance, just a noisy single-trial version of the second moments, $\mathbf{r}\mathbf{r}^\top$ (Table 1). In this simple example, the nuisance variable of the spatial phase ensures that quadratic statistics contain relevant information about the orientation, just like complex cells in V1[4].

Choice correlations predicted for optimal linear decoding. To study how neural information is used or decoded, past studies have examined whether neurons that are sensitive to sensory inputs also reflect an animal's behavioral outputs or choices[39–47]. This choice-related activity is hard to interpret, because it may reflect decoding of the recorded neurons, or merely correlations between them and other neurons that are decoded instead[48]. However, testable predictions about the choice-related activity can reveal the brain's decoding efficiency for linear codes[27]. Next, we discuss these predictions, and then generalize them to nonlinear codes.

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'[49], but it has almost identical properties (see the "Methods" subsection "Nonlinear choice correlations")[27,48].

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. There are several closely connected versions of this relationship that quantify information by estimator variance or Fisher information for continuous estimates, or by threshold or discriminability for binary estimates[27]—but the simplest is based on discriminability:

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

where d' and d'_{r_k} are, respectively, the stimulus discriminability[50] based on the behavior or on neuron k's response r_k (see the "Methods" subsection "Nonlinear choice correlations"). This relationship holds for binary classification derived from a locally

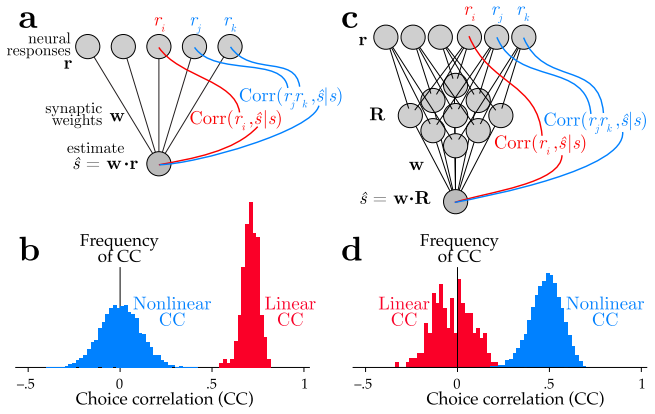

Fig. 3 Linear and nonlinear choice correlations successfully distinguish network structure. A linearly decoded population (**a**) produces nonzero linear choice correlations (**b**), while the nonlinear choice correlations are randomly distributed around zero. The situation is reversed for a nonlinear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

optimal linear estimator

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

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, with enough trials to reliably estimate their joint information content. 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].

Nonlinear choice correlations for optimal decoding. 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 Eq. (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 $\mathbf{R}(\mathbf{r})$:

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

(see the "Methods" subsection "Nonlinear choice correlations"), where $R_k(\mathbf{r})$ 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),

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

then the nonlinear choice correlation satisfies the equation

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

where $d'_{R_k(\mathbf{r})}$ is the stimulus discriminability provided by $R_k(\mathbf{r})$ (see the "Methods" subsection "Optimality test"). *This simple Equation (7) is the most important in this paper, and it is the basis of most predictions and intuitions we present in subsequent sections.*

Equation (7) predicts that choice correlations of individual statistics will be stronger for more informative statistics, those with higher discriminability d'_{R_k}. This reflects either stronger stimulus tuning of the statistic and/or lower variability of that statistic. This effect does not depend on whether the variability is shared. The variability that is not decoded dilutes the choice correlation; variability that is decoded increases it. Finally, choice correlations for optimal decoding can never be negative: if a statistic is tuned to increase with the stimulus, its fluctuations should correlate with choices that increase as well.

As an example of this relationship, we return to the orientation task. Here the response covariance $\Sigma(s) = \mathrm{Cov}(\mathbf{r}|s)$ depends on the stimulus, but the mean $\mathbf{f} = \langle \mathbf{r}|s \rangle = \langle \mathbf{r} \rangle$ 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, $\mathbf{R}(\mathbf{r}) = \mathrm{vec}(\mathbf{r}\mathbf{r}^\top)$, 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[4]. 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 (a suboptimal strategy for this example), linear choice correlations are strong while nonlinear choice correlations are near zero (Fig. 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 (Fig. 3c, d).

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[51]. 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].

The previous work[3] characterized linear information-limiting correlations for fine discrimination tasks by decomposing the noise covariance into $\Sigma = \Sigma_0 + \epsilon \mathbf{f}'\mathbf{f}'^\top$, 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 responses that encode the signal, $\mathbf{f}(s) = \langle \mathbf{r}|s \rangle$, but rather the nonlinear statistics $\mathbf{F}(s) = \langle \mathbf{R}(\mathbf{r})|s \rangle$. Likewise, the noise does not

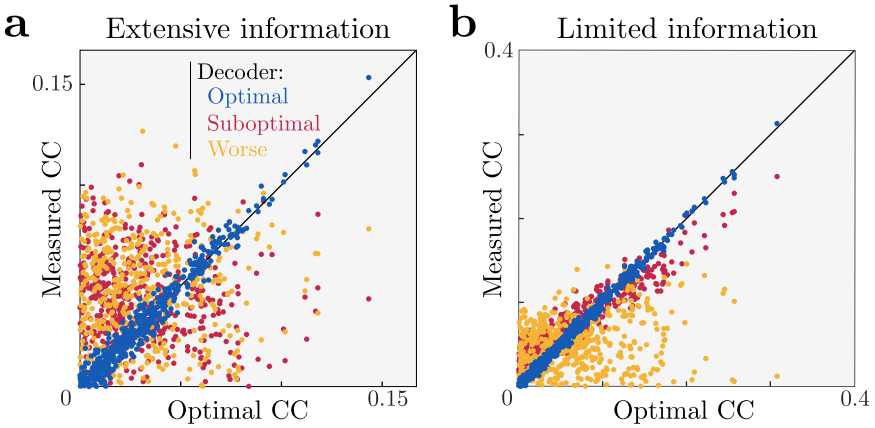

Fig. 4 Information-limiting noise makes a network more robust to suboptimal decoding. a A simulated optimal decoder produces measured choice correlations that match our optimal predictions (blue, on diagonal). In contrast, when a noise covariance Γ_0 permits the population to have extensive information, 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 ($\mathbf{w} \propto \mathbf{F}'$, red), and another 'worse' decoder that has the same weights but with 40% random sign flips (green). As in Fig. 5, 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).

comprise only second-order covariance of $\mathbf{r}$, $\mathrm{Cov}(\mathbf{r}|s)$, but rather the second-order covariance of the relevant nonlinear statistics, $\Gamma = \mathrm{Cov}(\mathbf{R}(\mathbf{r})|s)$ (see the "Results" subsection "Signal and noise"). Analogous to the linear case, these correlations can be locally decomposed as

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

where ϵ is again the variance of the information-limiting component, and Γ_0 is any other covariance that can be averaged away in large populations, including internal noise and external nuisance variation. The information-limiting noise bounds the estimator variance σ_s^2 to no smaller than ϵ even with optimal decoding. Likewise, the Fisher information cannot exceed the value of $1/\epsilon$, and the discriminability d' cannot exceed $ds/\sqrt{\epsilon}$ for a stimulus change of ds[3]. Neither additional cortical neurons nor additional decoded statistics can improve performance beyond this bound.

Consequences of optimal decoding on choice correlations. The simple formula of Eq. (7) provides useful insights into the relationship between neural activity and choice when that activity is decoded optimally in a natural task. First, the choice correlations do not depend on the shape or magnitude of the internal noise or nuisance correlations, because these dimensions are deliberately avoided by optimal decoding, whose weights cancel those correlations (Eq. (18)). The only aspect of shared variability that matters for choice is the information-limiting component, i.e. that which is indistinguishable from a change in the task-relevant stimulus. This information-limiting variability increases choice correlations mostly by decreasing the overall behavioral discriminability d'; the shared variability is large only at the population level and typically only has a small contribution to any one statistic[3,27] and thus to its discriminability d'_k. With a smaller denominator and roughly unaffected numerator, the ratio in Eq. (7) rises with information-limiting correlations.

Likewise, the total number of the decoded statistics only affects the prediction of Eq. (7) insofar as it affects the available information. This is especially important when considering nonlinear statistics, as there are potentially so many of them. When optimally decoded, greater numbers of *independently* informative statistics will increase the total discriminability exhibited by the behavioral choices, while the discriminability for each statistic remains unchanged. According to Eq. (7), choice correlations for optimal decoding are the ratio of these two, so as the behavioral behavioral d' increases and the individual terms remain fixed, the choice correlations shrink. On the other hand, greater numbers of *redundant* statistics change neither the total information content nor the behavioral choice. These added statistics can have similar tuning and similar fluctuations as each other (which is what makes them redundant). Any single redundant statistic might or might not be decoded, but it is correlated with others that are. According to Eq. (7), the individual d'_{R_k} are unchanged when adding more redundant statistics; the total information d' is unchanged; and thus their ratio is fixed, consistent with the unchanged choice correlations.

When there are many fewer sensory inputs than cortical neurons, as seen in the brain, many distinct statistics $R_k(\mathbf{r})$ will carry redundant information. Under these conditions, many choice correlations C_{R_k} can be quite large 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 a significant fraction of 1 (Fig. 4). Supplementary Information S.7.1 illustrates this effect for a redundant nonlinear population code: the brain need not decode all functions of all neurons to extract essentially all of the information (Fig. S6A), and neuroscientists need not compute choice correlations for all possible statistics to establish decoding efficiency (Fig. S6B).

Which nonlinear statistics? If the brain's decoder optimally uses all available information, choice correlations will obey the prediction of Eq. (7) even if the specific nonlinear statistics extracted by the brain's decoder differ from those selected for evaluating choice correlations (see the "Methods" subsection "Nonlinear choice correlation to analyze an unknown nonlinearity"). The prediction is valid as long as the brain's nonlinearity can be expressed as a linear combination of the tested nonlinearities (see the "Methods" subsection "Nonlinear choice correlation to analyze an unknown nonlinearity"). Since the brain needs complicated nonlinearities for complicated tasks, it may be difficult to find a suitable basis set for truly natural conditions; feature spaces from deep networks trained on comparable tasks might provide a

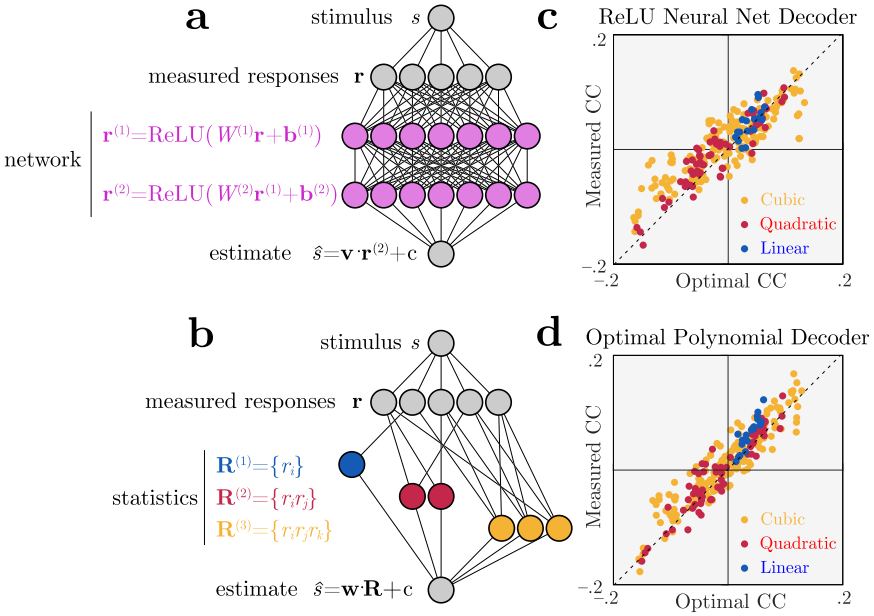

Fig. 5 Identifying optimal nonlinear decoding by a generic neural network using nonlinear choice correlations. Neural responses **r** are constructed to encode stimulus information in polynomial sufficient statistics up to cubic order (see the "Methods" section Eq. (13)). These responses are decoded 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 (see the "Methods" subsection "Nonlinear choice correlation to analyze an unknown nonlinearity"), 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)).

useful basis[38,52]. For the modest, controlled-complexity tasks used in most neuroscience experiments[53–56], polynomials or other simple bases may be sufficient, even when individual neurons do not use polynomial operations.

Indeed, Fig. 5 shows a situation where information is encoded by linear, quadratic and cubic sufficient statistics of neural responses, but 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.

Decoding efficiency revealed by choice correlations. Even if decoding is not strictly optimal, Eq. (7) can be approximately 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

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

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, $\alpha = \epsilon/\sigma_{\hat{s}}^2$. This slope α therefore provides an estimate of the efficiency of the brain's decoding.

Figure 4 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.

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[57], especially for unnatural tasks[58], so it is scientifically interesting to discover near-optimal nonlinear computation even in a redundant code.

Coarse versus fine, estimation versus classification. Our conceptual framework and predictions are most simply expressed for fine estimation tasks, where here we define 'fine' as a stimulus range over which the noise statistics do not vary with the stimulus. Some minor details of our predictions change when moving to binary classification instead of continuous estimation: this introduces a correction factor that depends on the response distribution (Section S.6.4).

More details change when moving to 'coarse' tasks, which we define as when noise statistics do 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, this relationship is slightly more complicated for coarse discrimination. For this reason we introduce a slightly more complicated measure of choice correlation that we call Normalized Average Conditional Choice Correlation (NACCC, Eq. (17)), which removes the stimulus-induced covariation between neuron and choice, and isolates only the remaining shared fluctuations that reflect the

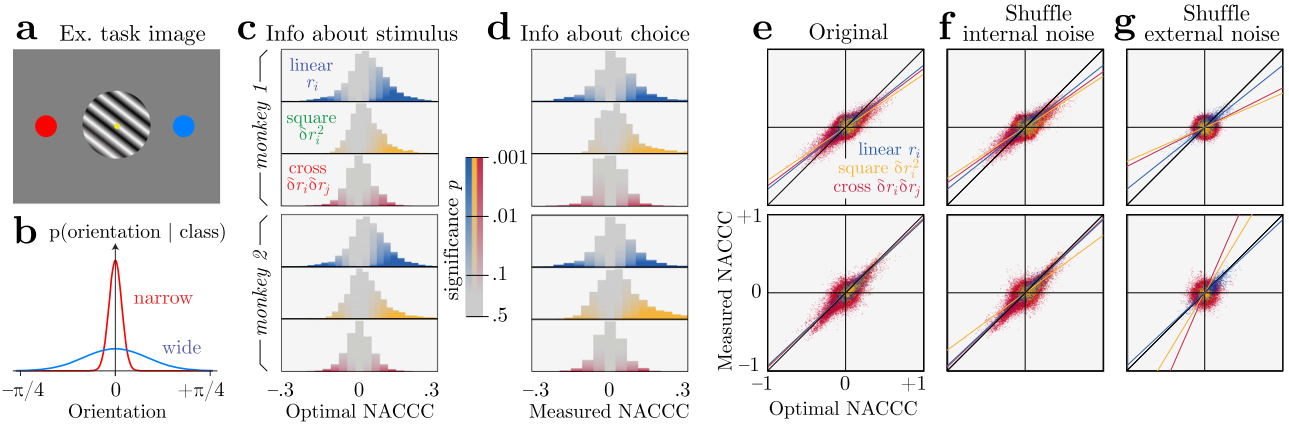

Fig. 6 Nonlinear information and choice correlations in a variance discrimination task, for neural data from two monkeys. a Example oriented grating and saccade targets. **b** The orientations of the gratings were drawn from a narrow or wide distribution, and the monkey had to guess which by saccading 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 (see the "Methods" subsection "Application to neural data"). **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). The coefficient of determination, R-squared is 0.50, 0.33, 0.12 for linear, square and cross terms for monkey 1; 0.61, 0.64, 0.40 for monkey 2. 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 (see the "Methods" subsection "Application to neural data") 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.

brain's processing. However, the end result is the same: choice correlations for optimal decoding are equal to the ratio of discriminabilities (Eq. (7); Supplemental Information S.5). As for fine estimation, there is a correction factor of order 1 for binary choices instead of continuous estimation (see the "Methods" subsection "Optimality test", Supplemental Information S.6.4, Eq. (185)).

Evidence for optimal nonlinear computation in macaque brains. We applied our optimality test to data recorded with Utah arrays from primate visual cortex (V1) during a nonlinear decoding task. Monkeys performed a Two-Alternative Forced Choice task (2AFC) in which they categorized an oriented drifting grating based on whether it came from a wide or narrow distribution of orientations[59] (Fig. 6a, b). The categorical target variable s is therefore the variance of the orientation distribution. This coarse binary discrimination task is a simplified version of a task that might arise in nature when identifying a surface texture or material[60]; the orientation of the material would be a nuisance variable independent of the material type. Here the observable variable, the orientation, is the product of the target variable and a nuisance variable v. An additional nuisance variable was the stimulus contrast varying independently of the stimulus variance, although here we only analyze the highest contrast.

Below we analyze whether the trial-by-trial nonlinear statistics of V1 multi-unit neural responses to these stimuli provide information about the task-relevant category and the behavioral choice, and whether these two informations are correlated as predicted by our optimal decoding theory. Since the marginal response statistics depend substantially on the stimulus category, at least in part because the corresponding nuisance distributions differ, this is a coarse binary discrimination task. For this reason we test suitable optimal decoding predictions about nonlinear choice correlations in coarse tasks, using the NACCC measure that we outlined in the section "Coarse versus fine, estimation versus classification" and derived in the "Methods" subsection "Application to neural data" and Supplemental Information S.6.

V1 responses contain information about orientation[61]. Here we found that V1 responses also contain some linear information about the orientation *variance* (Fig. 6c, blue; d' calculated by Eq. (20)). This implies that within their receptive field they have already performed some nonlinear transformations of the input that are useful for estimating the orientation variance. However, we expect that nonlinear computations downstream can extract still more information. Note that for a fixed contrast, an optimal computation based on the stimulus is simply to threshold the squared deviation from the mean orientation. Because neural responses in this brain area can be linearly decoded to compute orientation, a good downstream 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 $\delta r_i \delta r_j$ (Fig. 6c, red and green), verifying that downstream nonlinear computations could extract additional information from the neural responses. To isolate the nonlinear information we eliminated the linear stimulus dependence of the response, computing neural nonlinear statistics according to $\delta r_i = r_i - \langle r_i | \hat{s}_1 \rangle$, where $\hat{s}_1 = \mathbf{w}_{\text{opt}} \cdot \mathbf{r} + c$ is the optimal estimate decoded only from a linear combination of available neural responses.

These quadratic statistics also contained substantial nonlinear information about the behavioral choice (Fig. 6d). In general, there is no guarantee that the particular nonlinear statistics that are informative about the stimulus are also informative about the choice. However, 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 stimulus information (Fig. 6e). Remarkably, when we compare the measured nonlinear choice correlations to the ratio of discriminabilities after adjusting for the binary data (see the

"Methods" subsection "Application to neural data"), the slopes of this relationship for the two animals were near the value of 1 that Eq. (7) predicts for optimal decoding (Fig. 6e).

Monkey 2 performed slightly worse than an ideal observer, with a probability correct of 0.76, compared to the ideal of 0.82 (see the "Methods" subsection "Application to neural data")—even while its decoding was near-optimal, with an efficiency according to Eq. (9) of 0.96 ± 0.04 (mean ± 95% confidence intervals). Even at the level of individual sessions, this is consistent with optimal decoding, with efficiencies not significantly different from 1 ($p = 0.26$, one-tailed t-test). This suggests that information is lost in the encoding stage somewhere between the stimulus and the recorded neurons, and not downstream of those neurons. Monkey 1 had similar overall performance (probability correct of 0.74) but worse decoding efficiency (0.75 ± 0.08). Across sessions with reliable slopes (positive coefficients of determination, 77/119 sessions), the efficiencies were significantly different from 1 ($p < 10^{-6}$, one-tailed t-test). This suggests the second monkey's task performance has limitations arising downstream of the recorded neurons.

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, for each neuron we independently selected responses to high-contrast trials with matched target stimulus (variance), nuisance (orientations within ±1.5), and choice, and repeated our analysis on these shuffled data (Fig. 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 (Fig. 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.

We looked directly for stimulus-dependent internal noise correlations by conditioning on both the signal and the nuisance variable (which here is simply the single number, orientation) and measuring orientation-dependent response covariances. The resultant nonlinear tuning was quite weak compared with the trial-to-trial variability in those nonlinear statistics, and available nonlinear information arose largely in changing variances rather than covariances; likely arising from Poisson statistics and tuning of the mean firing (Supplementary Fig. S5A, B). Internal noise fluctuations in those directions were not significantly correlated with choice (Supplementary Fig. S5C, $p = 0.088$, 0.830, 0.969 for linear, square, and cross terms for monkey 1; $p = 0.073$, 0.094, 0.573 for linear, square, and cross terms for monkey 2 using a two-sample Kolmogorov–Smirnov test).

Recent analyses of these same data found that internal noise did in fact influence the monkeys' behavioral choices[62], but this effect was subtle and only apparent when examining the entire neural population simultaneously with a complex trained nonlinearity. 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. Additionally, in this work we analyzed a subset of trials with the highest contrasts and it is possible that at lower contrasts internal noise has a greater influence. 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.

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 shows how correlated fluctuations between neural activity and behavioral choices could reveal the efficiency of the brain's decoding. Unlike previous theories of nonlinear population codes[2,28], ours remains consistent with biological constraints due to the large cortical expansion of sensory representations by incorporating redundancy through a nonlinear generalization of information-limiting correlations[3]. Also unlike past work which largely concentrates on *encoding* efficiency, we provide mathematical methods to quantify the brain's nonlinear *decoding* efficiency. When we applied this method to the neural responses of monkeys performing a discrimination task in which neural statistics were dominated by nuisance variation, we found quantitative results consistent with efficient nonlinear decoding of V1 activity.

The best condition to apply our optimality test is in a task of modest complexity but still possessing fundamentally nonlinear structure. Some interesting examples where our test could have practical relevance include motion detection using photoreceptors[63], visual search with distractors (XOR-type tasks)[30,64], sound localization in early auditory processing before the inferior colliculus[65], or context switching in higher-level cortex[55].

Optimal nonlinearities extract the sufficient statistics about the relevant stimulus. These statistics depend not only on the task but also on the nuisance variables. In complex tasks, like recognizing objects from images, nuisance variables push most of the relevant information into higher-order statistics which require more complex nonlinearities to extract. In such high-dimensional cases, our proposed test is unlikely to be useful. This is because our method expresses stimulus estimates as sums of nonlinear functions, and while that is universal in principle[66], 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 important but small contributions to the behavioral output. 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 (Fig. 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 varies 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).

Our approach is currently limited to spatial feedforward processing, which unquestionably oversimplifies cortical processing. The approach can be generalized to recurrent networks by considering spatiotemporal statistics[67]. 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[68–70], 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.

In principle our method can also be applied to temporal neural response properties like spike timing or time series. For optimal processing, spike timing that is tuned to task-relevant stimuli should also be correlated with resultant choices, even if the timing is converted to rate codes by downstream processing. On the other hand, it is more difficult to track behavioral consequences of spatiotemporal correlations that evolve through a recurrent network[67] with dynamic outputs, as in motor control applications. It should be fruitful to develop this theory further for more complex tasks involving time sequences of actions.

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[71–73].

Methods

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, $G(\mathbf{x}|s,\nu)$, where $\mathbf{x}$ is spatial position in the image, and s and ν are the orientation and spatial phase (nuisance) of the image, respectively (Supplementary 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 $\mathbf{r} \sim P(\mathbf{r}|s,\nu) = N(\mathbf{r}|\mathbf{f}(s,\nu), \epsilon I)$, where ϵ is the noise variance and the mean $\mathbf{f}(s,\nu) = \langle \mathbf{r}|s,\nu\rangle = \sum_{\mathbf{r}} \mathbf{r}\, p(\mathbf{r}|s,\nu)$ 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: $\mathbf{f}(s) = \langle \mathbf{r}|s\rangle = \sum_{\mathbf{r}} \mathbf{r}\, p(\mathbf{r}|s) = \int d\nu \sum_{\mathbf{r}} \mathbf{r}\, p(\mathbf{r}|s,\nu)p(\nu)$. Since each spatial phase can be paired with another phase π radians away that inverts the linear response, the phase-averaged mean is $\mathbf{f}(s) = 0$. Thus the brain cannot estimate orientation by decoding these neurons linearly; nonlinear computation is necessary.

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

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

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 $\mathrm{Cov}_{ij}(\mathbf{r}|s)$ is tuned to s.

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,

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

where $\mathbf{R}(\mathbf{r})$ is a vector of sufficient statistics for the natural parameter $\mathbf{H}(s)$, $b(\mathbf{r})$ 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 Supplementary Material S.1.2.2, we show that the optimal local estimation can be achieved by linearly decoding the nonlinear sufficient statistics, $\hat{s} = \mathbf{w}^\top\mathbf{R}(\mathbf{r}) + c$. The decoding weights minimize the variance of an unbiased decoder,

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

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

Quadratic encoding. In a quadratic coding model, the distribution of neural responses is described by the exponential family with up to quadratic sufficient statistics, $\mathbf{R}(\mathbf{r}) = \{r_i, r_ir_j\}$ for $i,j \in \{1,\ldots,N\}$. A familiar example is the Gaussian distribution with stimulus-dependent covariance $\Sigma(s)$. In order to demonstrate the coding properties of a purely nonlinear neural code, here we assume that the mean tuning curve $f(s)$ 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 Eq. (61).

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

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

$$p(\mathbf{z}|s) = \frac{1}{Z}\exp\left(-\|\mathbf{z}\|^4 + \gamma s\, z_iz_jz_k\right) \tag{13}$$

We approximate this distribution by a mixture of four Gaussians. The mixture is chosen to reproduce the tetrahedral symmetry of the cubic distribution (Supplementary Fig. 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 Supplementary Fig. 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 $\mathbf{z}$ in a manner dependent on s:

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

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

Nonlinear choice correlations. For fine discrimination tasks, the nonlinear choice correlation between the stimulus estimate $\hat{s} = \mathbf{w}^\top\mathbf{R} + c$ and one nonlinear function R_k (the kth element of the vector $\mathbf{R}$) of recorded neural activity $\mathbf{r}$ is

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

where $\mathbf{w}^\top\Gamma\mathbf{w} = \sigma_{\hat{s}}^2$ 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 the section "Evidence for optimal nonlinear computation in macaque brains"), the relevant quantities change slightly. The optimal linear decoder of nonlinear

statistics, $\hat{s} = \mathbf{w} \cdot \mathbf{R} + c$, has weights obtained through linear regression:

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

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

These differences are reflected in a slightly modified measure of correlation that we call normalized average conditional choice correlations (NACCC),

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

B_{R_k} is actually a correlation coefficient based on the average conditional covariance $\bar{\Gamma}$, and is bounded in absolute value by 1. As the stimulus range in a coarse task decreases, and the noise distribution $p(\mathbf{R}|s)$ 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 (Supplementary Materials S.6.1).

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

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

where $d'_{R_k(\mathbf{r})} = F'_k \Delta s/\sqrt{\Gamma_{kk}}$ is the fine discriminability provided by $R_k(\mathbf{r})$ 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 $B^{\mathrm{opt}}_{R_k(\mathbf{r})} = d'_{R_k}/d'$.

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 $d' = \Delta s\sqrt{J}$ [74]; estimator variance is bounded by the inverse of the Fisher information, $\sigma^2_{\hat{s}} \geq 1/J$; discrimination threshold is proportional to the estimator standard deviation, $\theta = \sqrt{\sigma^2_{\hat{s}}}$ 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:

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

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, $D_{R_k} = \mathrm{Corr}(R_k, s)$, or a continuously varying behavioral choice $\hat{s}$ and the stimuli, $D_{\hat{s}} = \mathrm{Corr}(\hat{s}, s)$:

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

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, $\delta \approx 1.2 \pm 0.2$.

Nonlinear choice correlation to analyze an unknown nonlinearity. In Fig. 5, we generated neural responses given sufficient statistics that are polynomials up to third order, $\mathbf{R}(\mathbf{r}) = \{r_i, r_i r_j, r_i r_j r_k\}$ (see the "Methods" subsection "Cubic encoding"). 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 20,000 training pairs $(\mathbf{r}, s)$ generated by the cubic encoding model. This trained neural network extracted 91% of the information available to an optimal decoder.

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 $\mathbf{r} \sim p(\mathbf{r}|s + ds)$, 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, $\Gamma = \mathrm{Cov}(\mathbf{R}|s)$ conditioned on s; the information-limiting component arises specifically from the signal covariance $\mathrm{Cov}(\mathbf{F}(s)|s)$. Since the signal for local estimation of stimuli near a reference s_0 is $\mathbf{F}'(s) = \frac{d}{ds}\langle \mathbf{R}(\mathbf{r})|s \rangle$, the information-limiting component of the covariance is proportional to $\mathbf{F}'\mathbf{F}'^\top$:

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

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

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

where $J_0 = \mathbf{F}'\Gamma_0^{-1}\mathbf{F}'$ 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_∞.

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 12 h light/dark cycle, along with around 10 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 $\sigma_+ = 15°$ and $\sigma_- = 3°$. 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 300 ms within 0.5°–1° visual angle. The stimulus appeared at the center of the screen. After 500 ms, 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[12,75]. 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 neural data from the two monkeys is of comparable quality, although the monkey with higher task accuracy (monkey 2) performs more trials and has more significantly tuned neurons. When we resampled from both datasets to control for the number of trials and tuned neurons, and then used comparable datasets to do nonlinear choice correlation analysis, we found similar decoding efficiencies for two monkeys as was reported in the section "Evidence for optimal nonlinear computation in macaque brains" (data not shown).

The task-relevant stimulus s is the large or small variance $s_\pm = \sigma^2_\pm$ of the distribution over orientations. The orientation ϕ is a variable jointly determined by the task-relevant stimulus and a multiplicative nuisance variable ν through $\phi = \sqrt{s}\nu$, with $\nu \sim \mathcal{N}(0, 1)$. 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 according to $\hat{s} = \hat{\phi}^2$. A binary classification of the variance is given by $\hat{s}_\pm = \mathrm{sgn}\ (\hat{\phi}^2 - \theta^2)$ where θ is the animal's orientation threshold. This threshold is optimal where the two stimuli are equally probable: $p(\phi|s_+) = p(\phi|s_-)$, implying that $\theta^2_{\mathrm{opt}} = \left(\log s_+ - \log s_-\right)/\left(s_-^{-1} - s_+^{-1}\right)$. 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, $p(\hat{s}_\pm = +|s_+) = 2\int_{\theta_{\mathrm{opt}}}^{\infty} d\phi\ p(\phi|s_+) = \mathrm{erfc}\left(\theta_{\mathrm{opt}}/\sqrt{2s_+}\right)$; similarly, $p(\hat{s}_\pm = -|s_-) = 1 - \mathrm{erfc}\left(\theta_{\mathrm{opt}}/\sqrt{2s_-}\right)$. 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 $r_i \sim p(r_i|s,\phi,\hat{s})$ (Fig. 6f) or $r_i \sim p(r_i|s,\hat{s})$ (Fig. 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, $p = 1 - \frac{1}{2}\mathrm{erfc}\left(-|x|/\sqrt{2\sigma_x}\right)$ where x is the quantity of interest and σ_x is its standard deviation (Fig. 6c, d).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

All custom code used for electrophysiology data collection and data processing are made publicly available at github.com/atlab. Experimental data for Fig. 6 and code used for analysis and figure generation are available for download from github.com/xaqlab/nonlinear_choice_correlation.

Received: 3 May 2020; Accepted: 21 September 2021;
Published online: 16 November 2021

References

1. Shamir, M. & Sompolinsky, H. Nonlinear population codes. *Neural Comput.* **16**, 1105–1136 (2004).
2. Ecker, A. S., Berens, P., Tolias, A. S. & Bethge, M. The effect of noise correlations in populations of diversely tuned neurons. *J. Neurosci.* **31**, 14272–14283 (2011).
3. Moreno-Bote, R. et al. Information-limiting correlations. *Neuroscience* **17**, 1410–1417 (2014).
4. Adelson, E. H. & Bergen, J. R. Spatiotemporal energy models for the perception of motion. *Josa a* **2**, 284–299 (1985).
5. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
6. Pinto, N., Cox, D. D. & DiCarlo, J. J. Why is real-world visual object recognition hard? *PLoS Comput. Biol.* **4**, e27 (2008).
7. Rust, N. C. & DiCarlo, J. J. Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area v4 to it. *J. Neurosci.* **30**, 12978–12995 (2010).
8. Pagan, M., Urban, L. S., Wohl, M. P. & Rust, N. C. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nat. Neurosci.* **16**, 1132 (2013).
9. Meyers, E. M., Borzello, M., Freiwald, W. A. & Tsao, D. Intelligent information loss: the coding of facial identity, head pose, and non-face information in the macaque face patch system. *J. Neurosci.* **35**, 7069–7081 (2015).
10. Anselmi, F., Patel, A. & Rosasco, L. Neurally plausible mechanisms for learning selective and invariant representations. *J. Math. Neurosci.* **10**, 1–15 (2020).
11. Ecker, A. S. et al. Decorrelated neuronal firing in cortical microcircuits. *Science* **327**, 584–587 (2010).
12. Ecker, A. S. et al. State dependence of noise correlations in macaque primary visual cortex. *Neuron* **82**, 235–248 (2014).
13. Denfield, G. H., Ecker, A. S., Shinn, T. J., Bethge, M., & Tolias, A. S. Attentional fluctuations induce shared variability in macaque primary visual cortex. *Nat. commun.* **9**, 1–14 (2018).
14. Ecker, A. S., Denfield, G. H., Bethge, M. & Tolias, A. S. On the structure of neuronal population activity under fluctuations in attentional state. *J. Neurosci.* **36**, 1775–1789 (2016).
15. Bondy, A. G., Haefner, R. M. & Cumming, B. G. Feedback determines the structure of correlated variability in primary visual cortex. *Nat. Neurosci.* **21**, 598–606 https://doi.org/10.1038/s41593-018-0089-1 (2018).
16. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358 (2006).
17. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811 (2011).
18. Kohn, A., Coen-Cagli, R., Kanitscheider, I. & Pouget, A. Correlations and neuronal population information. *Annu. Rev. Neurosci.* **39**, 237–256 (2016).
19. Abbott, L. F. & Dayan, P. The effect of correlated variability on the accuracy of a population code. *Neural Comput.* **11**, 91–101 (1999).
20. Cohen, M. R. & Maunsell, J. H. Attention improves performance primarily by reducing interneuronal correlations. *Nat. Neurosci.* **12**, 1594 (2009).
21. Cohen, M. R. & Newsome, W. T. Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *J. Neurosci.* **29**, 6635–6648 (2009).
22. Gawne, T. J. & Richmond, B. J. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **13**, 2758–2771 (1993).
23. Beck, J., Bejjanki, V. R. & Pouget, A. Insights from a simple expression for linear fisher information in a recurrently connected population of spiking neurons. *Neural Comput.* **23**, 1484–1502 (2011).
24. Paradiso, M. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biol. Cybern.* **58**, 35–49 (1988).
25. Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
26. Sompolinsky, H., Yoon, H., Kang, K. & Shamir, M. Population coding in neuronal systems with correlated noise. *Phys. Rev. E* **64**, 051904 (2001).
27. Pitkow, X., Liu, S., Angelaki, D. E., DeAngelis, G. C. & Pouget, A. How can single sensory neurons predict behavior? *Neuron* **87**, 411–423 (2015).
28. Shamir, M. & Sompolinsky, H. Implications of neuronal diversity on population coding. *Neural Comput.* **18**, 1951–1986 (2006).
29. Burge, J. & Jaini, P. Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS Comput. Biol.* **13**, e1005281 (2017).
30. Pagan, M., Simoncelli, E. P. & Rust, N. C. Neural quadratic discriminant analysis: nonlinear decoding with v1-like computation. *Neural Comput.* **28**, 2291–2319 (2016).
31. Gutnisky, D. A. & Dragoi, V. Adaptive coding of visual information in neural populations. *Nature* **452**, 220 (2008).
32. Kohn, A. & Smith, M. A. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J. Neurosci.* **25**, 3661–3673 (2005).
33. Averbeck, B. B. & Lee, D. Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* **95**, 3633–3644 (2006).
34. Ohiorhenuan, I. E. et al. Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* **466**, 617 (2010).
35. Ponce-Alvarez, A., Thiele, A., Albright, T. D., Stoner, G. R. & Deco, G. Stimulus-dependent variability and noise correlations in cortical mt neurons. *Proc. Natl Acad. Sci. USA* **110**, 13162–13167 (2013).
36. Rigotti, M. et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* **497**, 585–590 (2013).
37. Pagan, M. & Rust, N. C. Dynamic target match signals in perirhinal cortex can be explained by instantaneous computations that act on dynamic input from inferotemporal cortex. *J. Neurosci.* **34**, 11067–11084 (2014).
38. Yamins, D. L. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
39. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Vis. Neurosci.* **13**, 87–100 (1996).
40. Shadlen, M. N., Britten, K. H., Newsome, W. T. & Movshon, J. A. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J. Neurosci.* **16**, 1486–1510 (1996).
41. Dodd, J. V., Krug, K., Cumming, B. G. & Parker, A. J. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area mt. *J. Neurosci.* **21**, 4809–4821 (2001).
42. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292 (2010).
43. de Lafuente, V. & Romo, R. Neuronal correlates of subjective sensory experience. *Nat. Neurosci.* **8**, 1698 (2005).
44. Treue, S. & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575 (1999).
45. Roitman, J. D. & Shadlen, M. N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *J. Neurosci.* **22**, 9475–9489 (2002).
46. Gu, Y., Angelaki, D. E. & DeAngelis, G. C. Neural correlates of multisensory cue integration in macaque mstd. *Nat. Neurosci.* **11**, 1201 (2008).
47. Purushothaman, G. & Bradley, D. C. Neural population code for fine perceptual decisions in area mt. *Nat. Neurosci.* **8**, 99 (2005).
48. Haefner, R. M., Gerwinn, S., Macke, J. H. & Bethge, M. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nat. Neurosci.* **16**, 235–242 (2013).

49. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Vis. Neurosci.* **13**, 87–100 (1996).
50. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (John Wiley, 1966).
51. Kanitscheider, I., Coen-Cagli, R. & Pouget, A. Origin of information-limiting noise correlations. *Proc. Natl Acad. Sci. USA* **112**, E6973–E6982 (2015).
52. Khaligh-Razavi, S. M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
53. Haag, J., Denk, W. & Borst, A. Fly motion vision is based on reichardt detectors regardless of the signal-to-noise ratio. *Proc. Natl Acad. Sci. USA* **101**, 16333–16338 (2004).
54. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
55. Saez, A., Rigotti, M., Ostojic, S., Fusi, S. & Salzman, C. Abstract context representations in primate amygdala and prefrontal cortex. *Neuron* **87**, 869–881 (2015).
56. Walker, E. Y., Cotton, R. J., Ma, W. J. & Tolias, A. S. A neural basis of probabilistic computation in visual cortex. *Nat. Neurosci.* **23**, 122–129 (2020).
57. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–9 (2012).
58. Nienborg, H. & Cumming, B. G. Psychophysically measured task strategy for disparity discrimination is reflected in v2 neurons. *Nat. Neurosci.* **10**, 1608 (2007).
59. Qamar, A. T. et al. Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proc. Natl Acad. Sci. USA* **110**, 20332–20337 (2013).
60. Karklin, Y. & Lewicki, M. S. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* **457**, 83–86 (2009).
61. Hubel, D. H. & Wiesel, T. N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **160**, 106–154 (1962).
62. Walker, E. Y., Cotton, R. J., Ma, W. J., & Tolias, A. S. A neural basis of probabilistic computation in visual cortex. *Nat. Neurosci*. **23**, 122–129 (2020).
63. Poggio, T. & Koch, C. Synapses that compute motion. *Sci. Am.* **256**, 46–53 (1987).
64. Ma, W. J., Navalpakkam, V., Beck, J. M., Van Den Berg, R. & Pouget, A. Behavior and neural basis of near-optimal visual search. *Nat. Neurosci.* **14**, 783 (2011).
65. Davis, K. A., Ramachandran, R. & May, B. J. Auditory processing of spectral cues for sound localization in the inferior colliculus. *J. Assoc. Res. Otolaryngol.* **4**, 148–163 (2003).
66. Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* **4**, 251–257 (1991).
67. Lakshminarasimhan, K., Pouget, A., DeAngelis, G., Angelaki, D. & Pitkow, X. Inferring decoding strategies for multiple correlated neural populations. *PLoS Comput. Biol.* **14**, e1006371 (2018).
68. Haefner, R. M., Berkes, P. & Fiser, J. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron* **90**, 649–660 (2016).
69. Graf, A. B., Kohn, A., Jazayeri, M. & Movshon, J. A. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat. Neurosci.* **14**, 239 (2011).
70. Maynard, E. et al. Neuronal interactions improve cortical population coding of movement direction. *J. Neurosci.* **19**, 8083–8093 (1999).
71. Pitkow, X. & Angelaki, D. E. Inference in the brain: statistics flowing in redundant population codes. *Neuron* **94**, 943–953 (2017).
72. Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A. & Poeppel, D. Neuroscience needs behavior: correcting a reductionist bias. *Neuron* **93**, 480–490 (2017).
73. Niv, Y. The primacy of behavioral research for understanding the brain. *Behav Neurosci.* **135**, 601–609 https://doi.org/10.1037/bne0000471 (2021).
74. Berens, P., Ecker, A. S., Gerwinn, S., Tolias, A. S. & Bethge, M. Reassessing optimal neural population codes with neurometric functions. *Proc. Natl Acad. Sci. USA* **108**, 4423–4428 (2011).
75. Tolias, A. S. et al. Recording chronically from the same neurons in awake, behaving primates. *J. Neurophysiol.* **98**, 3780–3790 (2007).

Acknowledgements

The authors thank Jeff Beck, Valentin Dragoi, Arun Parajuli, Alex Pouget, Nicole Rust, and Haim Sompolinsky for helpful conversations. This work was supported by NSF CAREER grant 1552868 to X.P., by NeuroNex grant 1707400 to X.P. and A.T., and by NSF Grant No. PHY-1748958, NIH Grant No. R25GM067110, the Gordon and Betty Moore Foundation Grant No. 2919.01. Q.Y. was supported in part by National Natural Science Foundation of China grant No. 32100832 and by the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province China Grant No. 19KJD520001. X.P. and A.T. were supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: the views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

Author contributions

X.P. and Q.Y. conceived the theoretical framework. X.P. and Q.Y. designed and performed the mathematical analyses. Q.Y. performed the simulations. E.W., R.J.C., and A.S.T. designed the experiments for a study with Wei Ji Ma; E.W., R.J.C., and A.S.T. performed the experiments; E.W. preprocessed the neural data; Q.Y. and X.P. analyzed the neural data. Q.Y. and X.P. wrote the manuscript; all authors discussed the results and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-26793-9.

Correspondence and requests for materials should be addressed to Xaq Pitkow.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Information

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 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

$$G(\mathbf{x}|s,\nu) = e^{-\|\mathbf{x}\|^2}\cos\left(\mathbf{k}\cdot\mathbf{x}+\nu\right) \tag{1}$$

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

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

$$\mathbf{k}_j = \kappa(\cos s_j, \sin s_j) \tag{3}$$

where each neuron has a preferred orientation s_j, spatial phase ν_j, and spatial frequency $\mathbf{k}_j$. 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:

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

This expression can be written in the form:

$$f_j(s,\nu) = A_j(s)\cos\left(\nu+\psi_j(s)\right) \tag{5}$$

using the stimulus-dependent response amplitude

$$A_j(s) = C\sqrt{2\cosh 2\beta_j(s) + 2\cos 2\nu_j} \tag{6}$$

and phase

$$\psi_j(s) = \nu_j - \alpha_j(s) \tag{7}$$

where we define the quantities

$$C = \frac{\pi}{4}\exp\left(-\frac{1}{4}\kappa^2\right) \tag{8}$$

$$\beta_j(s) = \frac{1}{4}\kappa^2\cos(s-s_j) \tag{9}$$

$$\alpha_j(s) = \tan^{-1}\frac{\exp\left(\beta_j(s)\right)\sin 2\nu_j}{\exp\left(-\beta_j(s)\right)+\exp\left(\beta_j(s)\right)\cos 2\nu_j} \tag{10}$$

Equation 5 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

$$P(\mathbf{r}|\nu, s) = \mathcal{N}(\mathbf{f}(s,\nu), \mathbf{I}) \tag{11}$$

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 $\mathbf{f}(s)$ can be expressed as

$$\begin{aligned} f_j(s) = \langle r_j|s\rangle &= \int r_j\, p(r_j|s) dr_j && (12)\\ &= \iint r_j\, p(r_j|s,\nu)\, p(\nu)\, dr_j\, d\nu && (13)\\ &= \int f_j(s,\nu) p(\nu)\, d\nu && (14)\\ &= \frac{1}{2\pi}\int f_j(s,\nu)\, d\nu && (15)\\ &= \frac{A_j(s)}{2\pi}\int_0^{2\pi} \cos\left(\nu + \psi_j(s)\right) d\nu && (16)\\ &= 0 && (17) \end{aligned}$$

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 $\mathrm{Cov}_{ij}[\mathbf{r}|s,\nu]$ as the neural covariance (noise correlations) when everything in the image is fixed, and $\mathrm{Cov}_{ij}[\mathbf{r}|s]$ as the neural covariance when the nuisance is unknown and free to vary (nuisance correlations). Then $\mathrm{Cov}_{ij}[\mathbf{r}|s]$ is

$$\begin{aligned} \mathrm{Cov}_{ij}[\mathbf{r}|s] &= \langle (r_i - f_i(s))(r_j - f_j(s))|s\rangle && (18)\\ &= \langle r_i r_j|s\rangle = \iint r_i r_j\, p(\mathbf{r}|s) dr_i dr_j && (19)\\ &= \int d\nu \iint r_i r_j\, p(\mathbf{r}|s,\nu) p(\nu) dr_i dr_j && (20)\\ &= \int d\nu\, p(\nu) \langle r_i r_j|s,\nu\rangle && (21)\\ &= \int d\nu\, p(\nu) \left(\mathrm{Cov}_{ij}[\mathbf{r}|s,\nu] + f_i(s,\nu) f_j(s,\nu)\right) && (22)\\ &= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi}\int d\nu f_i(s,\nu) f_j(s,\nu) && (23)\\ &= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi} D_{ij}(s) && (24) \end{aligned}$$

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

$$\begin{aligned} D_{ij}(s) &= \int d\nu f_i(s,\nu) f_j(s,\nu)\\ &= \int d\nu\, A_i(s)\cos\left(\nu + \psi_i(s)\right) A_j(s)\cos\left(\nu + \psi_j(s)\right)\\ &= \pi\cos(\psi_i(s) - \psi_j(s)) A_i(s) A_j(s) \end{aligned} \tag{25}$$

Here when we compute Equation 25, 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 $\mathrm{Cov}_{ij}[\mathbf{r}|s]$ depends on the orientation s. While linear computation is useless for estimating orientation since the mean responses are untuned (12), 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 $\mathbf{r}$, the conditional probability is a member of the exponential family when

$$p(\mathbf{r}|s) = b(\mathbf{r}) \exp\left(\mathbf{\Theta}(s)^\top \mathbf{R}(\mathbf{r}) - A(s)\right) \tag{26}$$

where $\mathbf{\Theta}(s)$ are the natural parameters, $\mathbf{R}(\mathbf{r})$ are the sufficient statistics, $A(s)$ and $b(\mathbf{r})$ are the log normalizer and base measure. The statistics $\mathbf{R}(\mathbf{r})$ 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 $J(s)$ [1–3,6–8]. The Fisher information is given by

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

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

$$\mathbf{F} = \frac{\partial A(s)}{\partial \mathbf{\Theta}(s)} \tag{29}$$

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

$$A' = \sum_i \frac{\partial A}{\partial \Theta_i}\frac{d\Theta_i}{ds} = \boldsymbol{\Theta}'^\top \mathbf{F} \tag{30}$$

$$A'' = \boldsymbol{\Theta}''^\top \mathbf{F} + \boldsymbol{\Theta}'^\top \mathbf{F}' \tag{31}$$

Thus we can compute two definitions of Fisher information.

$$J = -\left\langle \frac{\partial^2}{\partial s^2} \log P(\mathbf{r}|s) \right\rangle_{P(\mathbf{r}|s)} \tag{32}$$

$$= A'' - \boldsymbol{\Theta}''^\top \mathbf{F} \tag{33}$$

$$= \boldsymbol{\Theta}'^\top \mathbf{F}' \tag{34}$$

and

$$J = \left\langle \left(\frac{\partial}{\partial s} \log P(\mathbf{r}|s) \right)^2 \right\rangle_{P(\mathbf{r}|s)} \tag{35}$$

$$= \boldsymbol{\Theta}'^\top (\langle \mathbf{R}\mathbf{R}^\top \rangle - \mathbf{F}\mathbf{F}^\top) \boldsymbol{\Theta}' \tag{36}$$

$$= \boldsymbol{\Theta}'^\top \Gamma \boldsymbol{\Theta}' \tag{37}$$

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

Since the two definition are equivalent, we have

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

Substituting Equation 38 into Equation 37, we find the Fisher Information for the exponential family [5]

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

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:

$$\hat{s} = \operatorname*{argmax}_s p(\mathbf{r}|s) \tag{40}$$

$$= \operatorname*{argmax}_s \log p(\mathbf{r}|s) \tag{41}$$

$$= \operatorname*{argmax}_s \boldsymbol{\Theta}(s)^\top \mathbf{R}(\mathbf{r}) - A(s) \tag{42}$$

A Taylor expansion around the reference yields

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

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:

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

The solution is

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

Since $\mathbf{r}$ is a random quantity, we can express $\mathbf{R}$ as a mean and a deviation away from that mean: $\mathbf{R} = \langle \mathbf{R}|s_0 \rangle + \delta\mathbf{R} = \mathbf{F} + \delta\mathbf{R}$. In this case, $\boldsymbol{\Theta}''^\top \mathbf{R} - A'' = \boldsymbol{\Theta}''^\top \mathbf{F} - A'' + \boldsymbol{\Theta}''^\top \delta\mathbf{R}$, 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

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

where

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

and where we used the results from Equations 39 and 38, with $\Gamma = \mathrm{Cov}(\mathbf{R}|s_0)$ and $\mathbf{F} = \langle \mathbf{R}|s_0 \rangle$. Thus, in this limit, the optimal estimator for s is a linear decoding of the sufficient statistics $\mathbf{R}(\mathbf{r})$.

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, $p(\mathbf{r}|s) \sim \exp\left[\boldsymbol{\Theta}(s)^\top \mathbf{R}(\mathbf{r})\right]$ where $\mathbf{R}(\mathbf{r}) = (\ldots, r_i r_j, \ldots)$. A familiar example is a Gaussian distribution with stimulus-dependent covariance: $p(\mathbf{r}|s) = N(\mathbf{f}, \Sigma(s))$.

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:

$$\Sigma(s) = V(s) \Lambda V(s)^\top \tag{48}$$

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 39), we need to first calculate the derivative of the mean

$\mathbf{F}' = \frac{\partial}{\partial s}\langle \mathbf{R}(\mathbf{r})|s\rangle$ and covariance $\Gamma = \mathrm{Cov}[\mathbf{R}(\mathbf{r})|s]$ of the quadratic sufficient statistics.

Because the mean of $\mathbf{r}$ 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 $\mathbf{r}$,

$$\Sigma'(s) = Ue^{\Omega s}(\Omega X - X\Omega)e^{-\Omega s}U^\dagger \tag{49}$$

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:

$$\begin{aligned}\langle r_i r_j r_k r_n|s\rangle = &\langle r_i r_j|s\rangle \langle r_k r_n|s\rangle + \langle r_j r_n|s\rangle \langle r_i r_k|s\rangle \\ &+ \langle r_i r_n|s\rangle \langle r_j r_k|s\rangle\end{aligned} \tag{50}$$

where

$$\langle r_i r_j|s\rangle = \Sigma_{ij} + f_i f_j \tag{51}$$

Substitution of the response covariance (Equation 48) into Equation 50 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 $\mathbf{z} = (z_i, z_j, z_k) \in \mathbb{R}^3$ can be written as

$$p(\mathbf{z}|s) = \tfrac{1}{Z}\exp\left(\gamma(s) z_i z_j z_k - \|\mathbf{z}\|^4\right) \tag{52}$$

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

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

$$p(\mathbf{z}|s) \approx \sum_{a=1}^{4} p(a)p(\mathbf{z}|a,s) \tag{53}$$

$$= \sum_a \frac{1}{4}\mathcal{N}(\mathbf{z}|\mu_a(s), M_a(s)) \tag{54}$$

where

$$\mathbf{m}_a(s) = \frac{s}{\sqrt{1+s^2}}\mathbf{v}_a \tag{55}$$

and

$$M_a(s) = \frac{I + s^2 \mathbf{v}_a \mathbf{v}_a^\top}{(1+s^2)^2} \tag{56}$$

The vectors $\mathbf{v}_a$ reflect the four corners of the tetrahedron, $v_{a,i} = \pm 1$, to match the tetrahedral symmetry of the pure cubic code (Equation 52, Figure S1). To sample from this distribution, we randomly choose a component a and then sample from the gaussian $\mathcal{N}(\mathbf{z}|\mathbf{m}_a(s), M_a(s))$ 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).

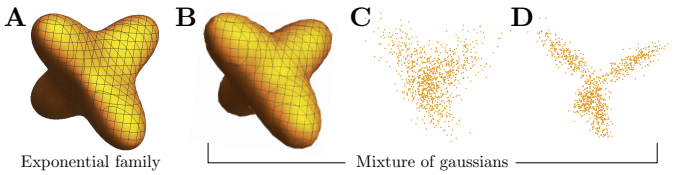

Figure S1: Multivariate skewed distributions. (**A**) Isoprobability contour of an exponential family distribution with cubic statistics in three dimensions, drawn from $p(\mathbf{z}|s) \propto \exp\left(sz_1z_2z_3 - \|\mathbf{z}\|^4\right)$. (**B**) Isoprobability contour for a mixture of four gaussians (Eq. 54). (**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.

$$p(\mathbf{z}|s) = \prod_\alpha p(\mathbf{z}_\alpha|s) = \prod_\alpha p(z_{\alpha_1}, z_{\alpha_2}, z_{\alpha_3}|s) \tag{57}$$

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:

$$\mathbf{r} = \mathbf{f}(s) + \Sigma^{1/2}(s)\,\mathbf{z}\,\Sigma^{1/2}(s) \tag{58}$$

$$\mathbf{z} \sim \frac{1}{Z(s)}\exp\left[\sum_{ijk}\gamma_{ijk}(s) z_i z_j z_k - \|\mathbf{z}\|^4\right] \tag{59}$$

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

$$p(\mathbf{r}|a,s) = \mathcal{N}(\mathbf{r}|\mathbf{f}(s) + \mathbf{m}_a(s), \Sigma^{1/2}(s) M_a(s) \Sigma^{1/2}(s)) \tag{60}$$

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 $\mathbf{r} \sim p(\mathbf{r}|s + ds)$, 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 $\mathbf{R}(\mathbf{r})$ conditioned on the stimulus into two parts:

$$\begin{aligned}\Gamma =& \mathrm{Cov}(\mathbf{R}(\mathbf{r})|s) \\ &= \langle \mathrm{Cov}(\mathbf{R}(\mathbf{r})|s, ds)\rangle_{ds} + \mathrm{Cov}\,\langle \mathbf{R}(\mathbf{r})|s, ds\rangle_{\mathbf{r}}\end{aligned} \tag{61}$$

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

$$\begin{aligned}\langle \mathrm{Cov}(\mathbf{R}(\mathbf{r})|s, ds)\rangle_{ds} &= \langle \Gamma(s + ds)\rangle_{ds} && (62)\\ &\approx \left\langle \Gamma_0 + ds\,\Gamma'\right\rangle_{ds} && (63)\\ &= \Gamma_0 && (64)\end{aligned}$$

Here we denote the covariance of $\mathbf{R}(\mathbf{r})$ 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 $\mathbf{R}$ in the absence of information-limiting correlations. The second term in Equation 61 can be expressed as

$$\begin{aligned}&\mathrm{Cov}\,\langle \mathbf{R}(\mathbf{r})|s, ds\rangle_{\mathbf{r}} && (65)\\ &= \mathrm{Cov}(\mathbf{F}(s + ds)|s) && (66)\\ &\approx \mathrm{Cov}(\mathbf{F}(s) + ds\,\mathbf{F}'(s)|s) && (67)\\ &= \frac{1}{J_\infty}\mathbf{F}'(s)\mathbf{F}'(s)^\top && (68)\end{aligned}$$

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

Equation 61 can therefore be written as

$$\Gamma = \Gamma_0 + \frac{1}{J_\infty}\mathbf{F}(s)'\mathbf{F}(s)'^\top \tag{69}$$

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

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

$$\Gamma^{-1} = \Gamma_0^{-1} - \frac{\Gamma_0^{-1}\mathbf{F}'\mathbf{F}'^\top\Gamma_0^{-1}}{J_\infty + \mathbf{F}'\Gamma_0^{-1}\mathbf{F}'^\top} \tag{70}$$

Substituting these equations into the nonlinear Fisher Information (Equation 39) and simplifying, we obtain

$$J_{R(\mathbf{r})} = \frac{1}{1/J_\infty + 1/J_0} \tag{71}$$

Here $J_0 = \mathbf{F}'^\top\Gamma_0^{-1}\mathbf{F}'$ 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 $\mathbf{r}$, *i.e.* a Taylor series expansion of the neural nonlinearities, $\mathbf{\Psi}(\mathbf{r}) = (r_i, r_ir_j, r_ir_jr_k, ...)$.

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

$$\hat{s}_{\mathrm{opt}} = \mathbf{w} \cdot \mathbf{R}(\mathbf{r}). \tag{72}$$

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

$$\hat{s}_{\mathrm{brain}} = \mathbf{v} \cdot \mathbf{g}(\mathbf{r}). \tag{73}$$

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 $\Psi(\mathbf{r})$ to check the optimality condition (Equation 7).

In Figure 4, we assumed that the optimal nonlinear basis function $\mathbf{R}$ is polynomial nonlinearity up to third order, $\mathbf{R}(\mathbf{r}) = (r_i, r_ir_j, r_ir_jr_k, ...)$. We used cubic codes described in Methods **Cubic encoding** to generate neural responses for which $\mathbf{R}(\mathbf{r})$ 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,

$$\hat{s}_{\mathrm{brain}} = \mathbf{v} \cdot \mathbf{r}^{(3)} + \mathbf{b}^{(3)} \tag{74}$$

$$\mathbf{r}^{(3)} = \mathrm{ReLU}(\mathbf{W}^{(2)}\mathbf{r}^{(2)} + \mathbf{b}^{(2)}) \tag{75}$$

$$\mathbf{r}^{(2)} = \mathrm{ReLU}(\mathbf{W}^{(1)}\mathbf{r}^{(1)} + \mathbf{b}^{(1)}) \tag{76}$$

$$\mathbf{r}^{(1)} = \mathbf{r} \tag{77}$$

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 $\{\mathbf{W}^{(\ell)}\}$, biases $\{\mathbf{b}^{(\ell)}\}$, and readout vector $\mathbf{v}$ 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 $\Psi(\mathbf{r})$ up to third order, *e.g.* r_i, $r_i r_j$, $r_i^2 r_j$, r_k^3, etc. The simulated choice correlations were calculated by Equation 5, where $\mathbf{R}(\mathbf{r}) = \Psi(\mathbf{r})$ 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 $\sqrt{J_{\Psi(\mathbf{r})}} = d'_\Psi/\Delta s = \frac{\Delta \mathbf{F}_\Psi}{\Delta s\, \sigma_\Psi}$, and $\sqrt{J} \approx 1/\sigma_{\hat{s}_{\mathrm{brain}}}$. We computed $\Delta\mathbf{F}_\Psi$ based on neural population responses $\mathbf{r}_+$ and $\mathbf{r}_-$ driven by stimuli $s_+ = s_0 \pm \Delta s/2$. The change in mean was $\Delta\mathbf{F}_\Psi = \langle\Psi(\mathbf{r}_+)\rangle - \langle\Psi(\mathbf{r}_-)\rangle$, and the average variance was $\sigma^2_\Psi = \frac{1}{2}\mathrm{Var}(\Psi(\mathbf{r}_+)) + \frac{1}{2}\mathrm{Var}(\Psi(\mathbf{r}_-))$. 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 [9]:

$$C_{R_k} = \frac{(\Gamma\mathbf{w})_k}{\sigma_k \sigma_{\hat{s}}} = \frac{(\Gamma_0\mathbf{w} + \frac{1}{J_\infty}\mathbf{F}'\mathbf{F}'^\top\mathbf{w})_k}{\sigma_k \sigma_{\hat{s}}} \tag{78}$$

For unbiased decoding, $\mathbf{w}^\top\mathbf{F}' = 1$. Some manipulation gives [9]

$$C_{R_k} = \frac{(\Gamma_0\mathbf{w})_k}{\Gamma_{0k}\sigma_{0\hat{s}}}\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}}\frac{\Gamma_{0k}}{\Gamma_k} + \frac{F'_k}{\sigma_k}\sigma_{\hat{s}}\frac{1/J_\infty}{\sigma^2_{\hat{s}}} \tag{79}$$

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 $\mathbf{w}$ 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:

$$\sigma^2_{\hat{s}} = \mathbf{w}^\top\Gamma\mathbf{w} = \mathbf{w}^\top\Gamma_0\mathbf{w} + \frac{1}{J_\infty}\mathbf{w}^\top\mathbf{F}'\mathbf{F}'^\top\mathbf{w} \tag{80}$$

$$= \sigma^2_{0\hat{s}} + 1/J_\infty \tag{81}$$

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

$$\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} = \sqrt{1 - \frac{1/J_\infty}{\sigma^2_{\hat{s}}}} = \sqrt{1-\alpha} \tag{82}$$

with $\alpha = \frac{1/J_\infty}{\sigma^2_{\hat{s}}}$. Substituting these into (Eq. 79) 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:

$$C_R^{\mathrm{sub}} \approx \alpha C_R^{\mathrm{opt}} + C_R^{\mathrm{sub}}\sqrt{1-\alpha} \tag{83}$$

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 $p(\mathbf{r}|s)$. 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 $\mathbf{R}(\mathbf{r})$ are linearly tuned to both the stimulus s and a scalar nuisance variable ν,

$$\mathbf{R}(\mathbf{r}) = \mathbf{F}'s + \mathbf{G}'\nu + \eta \tag{84}$$

where $\mathbf{F}'$ and $\mathbf{G}'$ characterize the sensitivity of $\mathbf{R}(\mathbf{r})$ to stimulus s and nuisance ν, and an internal noise source η 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

$$\Gamma = H + \xi \mathbf{G}'\mathbf{G}'^{\top} \tag{85}$$

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

$$\mathbf{w} = \frac{\Gamma^{-1}\mathbf{F}'}{J_1} \tag{86}$$

where the denominator $J_1 = \mathbf{F}'^{\top}\Gamma^{-1}\mathbf{F}'$ 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 $J_1 = 1/\sigma^2_{\hat{s}}$ [12]. The normalization by J_1 ensures the decoding is locally unbiased. These weights are used to estimate the stimulus according to

$$\hat{s} = \mathbf{w}^{\top}\mathbf{R}(\mathbf{r}) + b \tag{87}$$

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

$$c^{\text{sub}}_{R_k} = \text{Corr}(R_k, \hat{s}|s, \nu) \tag{88}$$

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:

$$c^{\text{sub}}_{R_k} = \frac{\text{Cov}(R_k, \hat{s}|s, \nu)}{\sigma_{R_k|s,n}\sigma_{\hat{s}|s,n}} \tag{89}$$

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

$$\text{Cov}(\mathbf{R}, \hat{s}|s, \nu) = \langle \mathbf{R}\hat{s}|s, n\rangle \tag{90}$$

$$= \langle \mathbf{R}\mathbf{R}^{\top}|s, n\rangle \mathbf{w} \tag{91}$$

$$= \frac{H\Gamma^{-1}\mathbf{F}'}{J_1} \tag{92}$$

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:

$$\Gamma^{-1} = (H + \xi\mathbf{G}'\mathbf{G}'^{\top})^{-1} \tag{93}$$

$$= H^{-1} - \frac{H^{-1}\mathbf{G}'\mathbf{G}'^{\top}H^{-1}}{1/\xi + \mathbf{G}'^{\top}H^{-1}\mathbf{G}'} \tag{94}$$

Substituting this inverse covariance into Equation 90, we obtain

$$\text{Cov}(\mathbf{R}, \hat{s}|s, \nu) \tag{95}$$

$$= \frac{1}{J_1}H(H^{-1} - \frac{H^{-1}\mathbf{G}'\mathbf{G}'^{\top}H^{-1}}{1/\xi + \mathbf{G}'H^{-1}\mathbf{G}'^{\top}})\mathbf{F}' \tag{96}$$

$$= \frac{1}{J_1}(\mathbf{F}' - \frac{\mathbf{G}'\mathbf{G}'^{\top}H^{-1}\mathbf{F}'}{1/\xi + \mathbf{G}'H^{-1}\mathbf{G}'^{\top}}) \tag{97}$$

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

$$\mathbf{J}(s, \nu) = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{F}'^{\top}H^{-1}\mathbf{F}' & \mathbf{F}'^{\top}H^{-1}\mathbf{G}' \\ \mathbf{G}'^{\top}H^{-1}\mathbf{F}' & \mathbf{G}'^{\top}H^{-1}\mathbf{G}' \end{bmatrix} \tag{98}$$

With these substitutions, we have

$$\text{Cov}(\mathbf{R}, \hat{s}|s, \nu) = \frac{1}{J_1}\left(\mathbf{F}' - \frac{J_{12}}{1/\xi + J_{22}}\mathbf{G}'\right) \tag{99}$$

The denominator of Equation 89 involves the variance of the sufficient statistics,

$$\sigma^2_{R_k|s,n} = H_{kk} \tag{100}$$

and the variance of the brain's decoder,

$$\begin{aligned}\sigma^2_{\hat{s}} &= \mathbf{w}^\top H \mathbf{w} \\ &= \mathbf{w}^\top (\Gamma - \xi \mathbf{G}' \mathbf{G}'^\top) \mathbf{w} \\ &= \frac{1}{J_1} - \frac{J_{12}^2}{\xi J_1^2} \frac{1}{(1/\xi + J_{22})^2}\end{aligned} \tag{101}$$

where we used the following results:

$$\begin{aligned}\mathbf{w}^\top \mathbf{G}' \mathbf{G}'^\top \mathbf{w} &= \left(\frac{\mathbf{F}' \Gamma^{-1}}{J_1} \mathbf{G}' \right)^2 \\ &= \frac{1}{J_1^2} \left(\mathbf{F}' H^{-1} \mathbf{G}' - \frac{\mathbf{F}' \Gamma^{-1} \mathbf{G}' \mathbf{G}' H^{-1} \mathbf{G}'}{1/\xi + \mathbf{G}' H^{-1} \mathbf{G}'} \right)^2 \\ &= \frac{1}{J_1^2} \left(J_{12} - \frac{J_{12} J_{22}}{1/\xi + J_{22}} \right)^2 \\ &= \frac{J_{12}^2}{\xi^2 J_1^2} \frac{1}{(1/\xi + J_{22})^2}\end{aligned} \tag{102}$$

Combining the results from Equation 99, 101 and 100, we can compute Equation 89

$$\begin{aligned}c^{\text{sub}}_{R_k} &= \text{Corr}(R_k, \hat{s}|s, \nu) \\ &= \frac{\text{Cov}(R_k, \hat{s}|s, \nu)}{\sigma_{R_k|s,n} \sigma_{\hat{s}|s,n}} \\ &= \frac{\frac{1}{J_1} \left(F'_k - \frac{J_{12}}{1/\xi + J_{22}} G'_k \right)}{\sqrt{H_{kk}} \sigma_{\hat{s}|s,n}}\end{aligned} \tag{103}$$

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

$$C^{\text{opt}}_{R_k} = \sqrt{\frac{J_{1,R_k}}{J_1}} = \frac{F'_k}{\sigma_{R_k|s} \sqrt{J_1}} \tag{104}$$

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

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

$$c^{\text{sub}}_{R_k} = \beta_k C^{\text{opt}}_{R_k} - \gamma_k \tag{105}$$

where we have defined the following constants:

$$\begin{aligned}\beta_k &= \frac{\sigma_{R_k|s}}{\sigma_{R_k|s,n}} \frac{1}{\sqrt{J_1} \sigma_{\hat{s}|s,n}} \\ &= \sqrt{\frac{H_{kk} + \xi G'^2_k}{H_{kk}}} \frac{1}{\sqrt{J_1} \sigma_{\hat{s}|s,n}} \\ &= \sqrt{\frac{H_{kk} + \xi G'^2_k}{H_{kk}}} \frac{1}{\sqrt{1 - \frac{J_{12}^2}{\xi J_1} \frac{1}{(1/\xi + J_{22})^2}}},\end{aligned} \tag{106}$$

and

$$\gamma_k = \frac{G'_k}{\sqrt{H_{kk}}} \frac{J_{12}}{(1/\xi + J_2) J_1 \sigma_{\hat{s}|s,n}} \tag{107}$$

The slope β_k and offset γ_k of the relationship between these two types of choice correlations (Equation 105) 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 $\mathbf{R}(\mathbf{r})$, 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 $\mathbf{R}(\mathbf{r}) = \mathbf{r}$ for simplicity. Neural responses were generated from a Gaussian distribution with a stimulus-dependent mean and identity covariance $H = I$: $p(\mathbf{r}|s, \nu) = \mathcal{N}(\mathbf{F}' s + \mathbf{G}' \nu, I)$. In Figure S2A, $\mathbf{F}'$ and $\mathbf{G}'$ are set to be orthogonal to ensure $J_{12} = \mathbf{F}'^\top H^{-1} \mathbf{G}' = 0$. 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 $\mathbf{F}'$ and $\mathbf{G}'$ 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

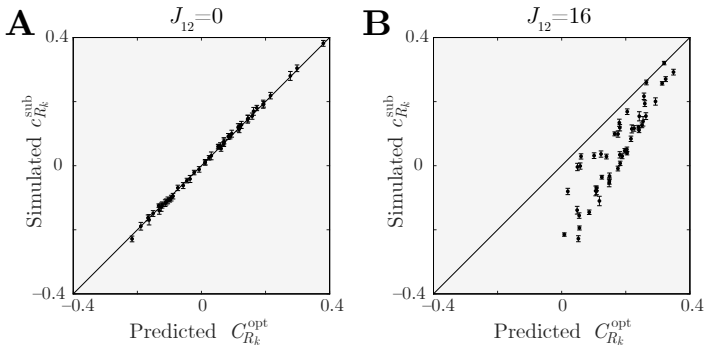

Figure S2: Comparing choice correlations caused by internal and external noise. (**A**) When estimates of nuisance variables are independent of estimates of task-relevant signals, the optimal choice correlations driven by internal noise, $c^{\rm sub}_{R_k}$, match the optimal pattern $C^{\rm opt}_{R_k}$ 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. Means and standard deviations (denoted by error bars) for simulated choice correlations were compute by repeating the tests for 10 times independently.

variation with zero mean and variance $\xi = 2$. The decoding weights follow Equation 86, and the stimulus is estimated using Equation 87. Choice correlations in this fixed-nuisance experiment are computed by Equation 88 (vertical axis in Figure S2). The predicted optimal choice correlation is computed by Equation 104 (horizontal axis in Figure S2). In this setting, $\beta_k \approx 1$ when $J_{12} = 0$.

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 $\mathbf{r}$ as a linear weighted sum of nonlinear statistics $\mathbf{R}(\mathbf{r})$, using weights given by linear regression as

$$\mathbf{w} \propto \mathrm{Cov}(\mathbf{R})^{-1}\mathrm{Cov}(\mathbf{R}, s) \tag{108}$$

The latter factor reflects the signal strength,

$$\begin{aligned}\mathrm{Cov}(\mathbf{R}, s) &= \langle \mathbf{R}s\rangle - \langle \mathbf{R}\rangle\langle s\rangle &(109)\\ &= \tfrac{1}{2}(\mathbf{F}_+ - \mathbf{F}_-)ds = \tfrac{1}{2}\Delta\mathbf{F}\, ds &(110)\end{aligned}$$

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

$$\begin{aligned}\mathrm{Cov}\,\mathbf{R} &= \langle \mathrm{Cov}(\mathbf{R}|s)\rangle_s + \mathrm{Cov}\,\langle \mathbf{R}|s\rangle &(111)\\ &= \bar\Gamma + \tfrac{1}{4}\Delta\mathbf{F}\Delta\mathbf{F}^\top &(112)\end{aligned}$$

where the first term is the average noise covariance across the stimulus ensemble, $\bar\Gamma = \langle \mathrm{Cov}(\mathbf{R}|s)\rangle_s$, 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

$$\begin{aligned}\mathbf{w} &\propto \mathrm{Cov}(\mathbf{R})^{-1}\,\mathrm{Cov}(\mathbf{R}, s)\\ &= \left(\bar\Gamma + \tfrac{1}{4}\Delta\mathbf{F}\Delta\mathbf{F}^\top\right)^{-1}\tfrac{1}{2}\Delta\mathbf{F}\,\Delta s\\ &\propto \left(\bar\Gamma^{-1} - \frac{\frac{1}{4}\bar\Gamma^{-1}\Delta\mathbf{F}\Delta\mathbf{F}^\top\bar\Gamma^{-1}}{1 + \frac{1}{4}\Delta\mathbf{F}^\top\bar\Gamma^{-1}\Delta\mathbf{F}}\right)\tfrac{1}{2}\Delta\mathbf{F}\\ &\propto \bar\Gamma^{-1}\Delta\mathbf{F} &(113)\end{aligned}$$

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

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 [10, 13]. Mathematically, we examine the statistical relationship between neural responses and choices that remains after *conditioning* on the stimulus, via $p(\mathbf{R}, \hat{s}|s)$. Here we quantify this relationship through a conditional covariance, $\mathrm{Cov}(\mathbf{R}, \hat{s}|s)$. 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

$$B_{R_k} = \frac{\langle \mathrm{Cov}(R_k, \hat{s}|s)\rangle_s}{\sqrt{\langle \mathrm{Var}(R_k|s)\rangle_s \langle \mathrm{Var}(\hat{s}|s)\rangle_s}} \tag{114}$$

This nonlinear choice correlation can be rewritten as

$$\begin{aligned} B_{R_k} &= \frac{(\mathbf{w}^\top \langle \mathrm{Cov}(\mathbf{R}|s)\rangle_s)_k}{\sqrt{\bar{\Gamma}_{kk}\ \mathbf{w}^\top \langle \mathrm{Cov}(\mathbf{R}|s)\rangle_s \mathbf{w}}} \\ &= \frac{(\Delta\mathbf{F}^\top \bar{\Gamma}^{-1}\bar{\Gamma})_k}{\sqrt{\bar{\Gamma}_{kk}\ \Delta\mathbf{F}^\top \bar{\Gamma}^{-1}\bar{\Gamma}\bar{\Gamma}^{-1}\Delta\mathbf{F}}} \\ &= \frac{\Delta\mathbf{F}_k}{\sqrt{\bar{\Gamma}_{kk}\ \Delta\mathbf{F}^\top \bar{\Gamma}^{-1}\Delta\mathbf{F}}} \end{aligned} \tag{115}$$

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

$$B^{\mathrm{opt}}_{R_k} = \frac{d'_k}{d'} \tag{116}$$

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,

$$\begin{aligned} D_{R_k,s} &= \mathrm{Corr}(R_k, s) \qquad (117)\\ &= \frac{\mathrm{Cov}(R_k, s)}{\sigma_{R_k}\sigma_s} \\ &= \frac{\frac{1}{2}\Delta F_k\, ds}{\sqrt{(\bar{\Gamma}_{kk} + \frac{1}{4}\Delta F_k^2)\sigma_s^2}} \\ &= \frac{1}{\sqrt{\frac{4\bar{\Gamma}_{kk}}{\Delta F_k^2} + 1}} \\ &= \frac{1}{\sqrt{4{d'_k}^{-2} + 1}} \qquad (118) \end{aligned}$$

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. 118 to find

$$d'_k = \frac{2}{\sqrt{D^{-2}_{R_k,s} - 1}} \tag{119}$$

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, $D_{\hat{s},s} = \mathrm{Corr}(\hat{s}, s)$:

$$d' = \frac{2}{\sqrt{D^{-2}_{\hat{s},s} - 1}} \tag{120}$$

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:

$$B^{\mathrm{opt}}_{R_k} \approx \frac{D_{R,s}}{D_{\hat{s},s}} \tag{121}$$

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.

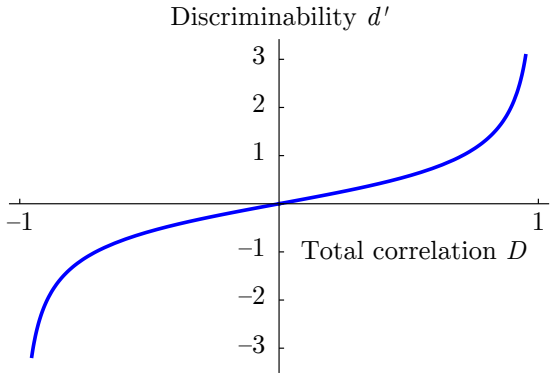

Figure S3: Stimulus discriminability d' for a response variable R, versus total correlation between that variable and the stimulus, $D = \mathrm{Corr}\,(R, s)$, according to Eq. 119.

S.6 Orientation variance discrimination task

S.6.1 Coarse tasks: Continuous estimation versus binary discrimination

The experiment of Section **Evidence for optimal nonlinear computation in macaque brains** 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 **Optimality test**, 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 dis-

crimination, 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 $\hat{s} = \hat{\phi}^2$, and a binary guess about the variance is given by $\hat{s}_\pm = \text{sgn}\,(\hat{\phi}^2 - \theta^2)$ where θ is the animal's orientation threshold. We assume that $\hat{\phi}$ is an unbiased estimate of the orientation ϕ, so $\langle \hat{\phi} | \phi \rangle = \phi$. We denote one neuron's mean response to the orientation by $\langle r|\phi\rangle = \mu(\phi)$ which we approximate linearly as $\mu(\phi) \approx \bar{\mu} + \mu'\phi$ 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 $p(r, \hat{\phi}|s)$ arises from both internal noise and nuisance variation, $p(r, \hat{\phi}|s) = \int d\phi\, p(r, \hat{\phi}|\phi) p(\phi|s)$. For a given orientation ϕ, the neural response r and orientation estimate $\hat{\phi}$ follow a bivariate normal distribution,

$$p(r, \hat{\phi}|\phi) = \mathcal{N}\left(\begin{matrix} r \\ \hat{\phi} \end{matrix} \middle| \begin{matrix} \mu(\phi) \\ \phi \end{matrix}; \begin{bmatrix} H_{rr|\phi} & H_{r\hat{\phi}|\phi} \\ H_{r\hat{\phi}|\phi} & H_{\hat{\phi}\hat{\phi}|\phi} \end{bmatrix} \right) \tag{122}$$

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

By design, the orientation variable ϕ is driven by a normally distributed nuisance variable ν, with $\phi = \sqrt{s}\nu$ and $p(\phi|s) = \mathcal{N}(\phi|0, s)$, so we can write the marginal distribution $p(r, \hat{\phi}|s)$ as

$$\begin{aligned} p(r, \hat{\phi}|s) &= \mathcal{N}\left(\begin{matrix} r \\ \hat{\phi} \end{matrix} \middle| \begin{matrix} \bar{\mu} \\ 0 \end{matrix}; \begin{bmatrix} H_{rr|\phi} + \mu'^2 s & H_{r\hat{\phi}|\phi} + \mu' s \\ H_{r\hat{\phi}|\phi} + \mu' s & H_{\hat{\phi}\hat{\phi}|\phi} + s \end{bmatrix} \right) \\ &= \mathcal{N}(\mu(s), \Sigma(s)) \end{aligned} \tag{123}$$

For now we suppress the explicit dependence on s.

The conditional covariance between the nonlinear statistic R and choice is

$$\begin{aligned} &\text{Cov}(\hat{s}_\pm, R|s) \\ &= \left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) r^2 \right\rangle_{r,\hat{\phi}} - \left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) \right\rangle_{\hat{\phi}} \left\langle r^2 \right\rangle_r \end{aligned} \tag{124}$$

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

$$\begin{aligned} &\left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) r^2 \right\rangle_{r,\hat{\phi}} \\ &= \left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) \left\langle r^2 | \hat{\phi} \right\rangle_{r|\hat{\phi}} \right\rangle_{\hat{\phi}} \\ &= \left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) (\Sigma_{r|\hat{\phi}} + \mu^2_{r|\hat{\phi}}) \right\rangle_{\hat{\phi}} \\ &= \left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) \left[\Sigma_{rr} - \frac{\Sigma^2_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} + \left(\mu + \frac{\Sigma_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} \hat{\phi} \right)^2 \right] \right\rangle_{\hat{\phi}} \end{aligned} \tag{125}$$

where we used the conditional distribution

$$p(r|\hat{\phi}) = \mathcal{N}\left(r \middle| \mu + \frac{\Sigma_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} \hat{\phi},\, \Sigma_{rr} - \frac{\Sigma^2_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} \right) \tag{126}$$

This can be written as

$$\left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2)(a\hat{\phi}^2 + b\hat{\phi} + c) \right\rangle_{\hat{\phi}} \tag{127}$$

for coefficients

$$a = \frac{\Sigma^2_{r\hat{\phi}}}{\Sigma^2_{\hat{\phi}\hat{\phi}}} \tag{128}$$

$$b = 2\frac{\Sigma_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} \mu \tag{129}$$

$$c = \Sigma_{rr} - \frac{\Sigma^2_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} + \mu^2 \tag{130}$$

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

$$\begin{aligned} &\left\langle \text{sgn}\,(\hat{\phi}^2 - \theta^2) \hat{\phi}^\alpha \right\rangle_{\hat{\phi}} \\ &= \left[\int_{-\infty}^{-\theta} - \int_{-\theta}^{\theta} + \int_{\theta}^{\infty} \right] \hat{\phi}^\alpha p(\hat{\phi}) d\hat{\phi} \\ &= \left[\int_{-\infty}^{-\theta} - \left(\int_{-\infty}^{\theta} - \int_{-\infty}^{-\theta} \right) + \left(\int_{-\infty}^{\infty} - \int_{-\infty}^{\theta} \right) \right] \hat{\phi}^\alpha p(\hat{\phi}) d\hat{\phi} \\ &= \left[2\int_{-\infty}^{-\theta} - 2\int_{-\infty}^{\theta} + \int_{-\infty}^{\infty} \right] \hat{\phi}^\alpha p(\hat{\phi}) d\hat{\phi} \end{aligned} \tag{131}$$

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

$$\int_{-\infty}^{\theta} d\hat{\phi}\, \hat{\phi}^0 \mathcal{N}(\hat{\phi}|0, \sigma^2_{\hat{\phi}}) = \tfrac{1}{2} \text{erfc}\left(\tfrac{\theta}{\sqrt{2}\sigma_{\hat{\phi}}} \right) \tag{132}$$

$$\int_{-\infty}^{\theta} d\hat{\phi}\, \hat{\phi}^1 \mathcal{N}(\hat{\phi}|0, \sigma^2_{\hat{\phi}}) = -p_{\hat{\phi}}(\theta) \sigma^2_{\hat{\phi}} \tag{133}$$

$$\int_{-\infty}^{\theta} d\hat{\phi}\, \hat{\phi}^2 \mathcal{N}(\hat{\phi}|0, \sigma^2_{\hat{\phi}}) = \tfrac{1}{2} \text{erfc}\left(\tfrac{\theta}{\sqrt{2}\sigma_{\hat{\phi}}} \right) \sigma^2_{\hat{\phi}} - p_{\hat{\phi}}(\theta) \sigma^2_{\hat{\phi}} \theta \tag{134}$$

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

Combining these with Eq. 131 we obtain

$$m = \left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2) \right\rangle = 2\,\mathrm{erfc}\Big(\tfrac{\theta}{\sqrt{2}\sigma}\Big) - 1 \tag{135}$$

$$\left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2)\hat{\phi} \right\rangle = 0 \tag{136}$$

$$\left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2)\hat{\phi}^2 \right\rangle = \sigma^2_{\hat{\phi}} m + 4\theta\sigma^2_{\hat{\phi}} p_{\hat{\phi}}(\theta) \tag{137}$$

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

Returning to Eq. 125, we have

$$\begin{aligned}
&\mathrm{Cov}(\hat{s}_\pm, R|s) \\
&= \left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2) r^2 \right\rangle_{r,\hat{\phi}} - \left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2) \right\rangle_{\hat{\phi}|s} \left\langle r^2 \right\rangle_r \\
&= \left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2)(a\hat{\phi}^2 + b\hat{\phi} + c) \right\rangle_{\hat{\phi}} \\
&\quad - \left\langle \mathrm{sgn}\,(\hat{\phi}^2 - \theta^2) \right\rangle_{\hat{\phi}} \left\langle r^2 \right\rangle_r \\
&= a\left[\sigma^2_{\hat{\phi}} m + 4\theta\sigma^2_{\hat{\phi}} p_{\hat{\phi}}(\theta)\right] m + cm - m(\Sigma_{rr} + \mu^2) \\
&= 4\theta \frac{\Sigma^2_{r\hat{\phi}}}{\Sigma_{\hat{\phi}\hat{\phi}}} p_{\hat{\phi}}(\theta)
\end{aligned} \tag{138}$$

Note that all of the erfc terms have canceled.

Compare that to the corresponding covariance for continuous estimation,

$$\mathrm{Cov}\,(\hat{s}, R|s) = 2\Sigma^2_{r\hat{\phi}} \tag{139}$$

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

$$\mathrm{Var}(\hat{s}_\pm|s) = 1 - \langle \hat{s}_\pm | s \rangle^2 \tag{140}$$

$$= 1 - m^2 \tag{141}$$

whereas, the variance for the continuous estimator is

$$\mathrm{Var}(\hat{s}|s) = \left\langle (\hat{\phi}^2 - \theta^2)^2 | s \right\rangle - \left\langle \hat{\phi}^2 - \theta^2 | s \right\rangle^2 \tag{142}$$

$$= 2\Sigma^2_{\hat{\phi}\hat{\phi}} \tag{143}$$

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,

$$\zeta = \frac{B^\pm_R}{B_R} = \frac{\frac{\langle \mathrm{Cov}(\hat{s}_\pm, R|s)\rangle_s}{\sqrt{\langle \mathrm{Var}(\hat{s}_\pm|s)\rangle_s \langle \mathrm{Var}(R|s)\rangle_s}}}{\frac{\langle \mathrm{Cov}(\hat{s}, R|s)\rangle_s}{\sqrt{\langle \mathrm{Var}(\hat{s}|s)\rangle_s \langle \mathrm{Var}(R|s)\rangle_s}}} \tag{144}$$

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

$$\zeta = \frac{\langle \mathrm{Cov}\,(\hat{s}_\pm, R|s)\rangle_s}{\langle \mathrm{Cov}\,(\hat{s}, R|s)\rangle_s} \sqrt{\frac{\langle \mathrm{Var}\,(\hat{s}|s)\rangle_s}{\langle \mathrm{Var}\,(\hat{s}_\pm|s)\rangle_s}} \tag{145}$$

$$= \frac{\sum_s p(s) 4\theta \frac{\Sigma^2_{r\hat{\phi}|s}}{\Sigma_{\hat{\phi}\hat{\phi}|s}} p_{\hat{\phi}}(\theta|s)}{\sum_s p(s) 2\Sigma^2_{r\hat{\phi}|s}} \sqrt{\frac{\sum_s p(s) 2\Sigma^2_{\hat{\phi}\hat{\phi}|s}}{\sum_s p(s)(1 - m_s^2)}} \tag{146}$$

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 $\Sigma_{rr} \approx \mu'^2 s$, $\Sigma_{r\hat{\phi}} \approx \mu' s$, and $\Sigma_{\hat{\phi}\hat{\phi}} \approx s$. Substituting these approximations into the expression above, we obtain

$$\zeta \approx \frac{\frac{1}{2}\sum_s 4\theta \frac{\mu'^2 s^2}{s} \frac{e^{-\theta^2/2s}}{\sqrt{2\pi s}}}{\frac{1}{2}\sum_s 2\mu'^2 s^2} \sqrt{\frac{\frac{1}{2}\sum_s 2s^2}{1 - \frac{1}{2}\sum_s m_s^2}} \tag{147}$$

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

$$\zeta \approx \frac{\frac{1}{2} 4\theta \sqrt{\frac{s_+}{2\pi}}}{\frac{1}{2} 2 s_+^2} \sqrt{\frac{\frac{1}{2} 2 s_+^2}{1 - \frac{1}{2}\sum_s m_s^2}} \tag{148}$$

$$= \frac{2}{\sqrt{\pi}} \frac{\theta}{\sqrt{s_+}} \frac{1}{\sqrt{1 - \frac{1}{2}\sum_s m_s^2}} \tag{149}$$

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

$$\zeta \approx \frac{2\theta}{\sqrt{\pi s_+}} \tag{150}$$

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. 150, we used logistic regression of choice $\hat{s}_\pm$ on the absolute value of the stimulus orientation, $|\phi|$, and assign the threshold θ to be the orientation where the probability of both choices was equal.

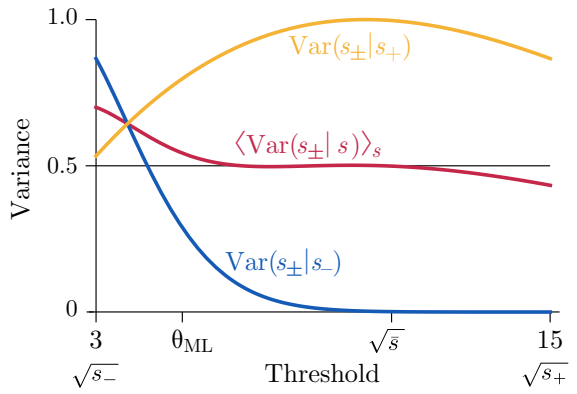

Figure S4: The average variance of $\hat{s}_\pm$ conditioned on the stimulus s (red) is approximately 1/2 over a wide range of thresholds.

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.

$$\zeta^{\text{fine}} = \frac{\mathrm{Cov}\left(\hat{s}_\pm, R|s\right)}{\mathrm{Cov}\left(\hat{s}, R|s\right)} \sqrt{\frac{\mathrm{Var}\left(\hat{s}|s\right)}{\mathrm{Var}\left(\hat{s}_\pm|s\right)}} \tag{151}$$

$$= \frac{4\theta \frac{\Sigma^2_{r\hat{\phi}|s}}{\Sigma_{\hat{\phi}\hat{\phi}|s}} p_{\hat{\phi}}(\theta|s)}{2\Sigma^2_{r\hat{\phi}|s}} \sqrt{\frac{2\Sigma^2_{\hat{\phi}\hat{\phi}|s}}{1-m_s^2}} \tag{152}$$

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

$$\zeta^{\text{fine}} = \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \sqrt{\frac{8}{1-(2\,\mathrm{erfc}(\frac{1}{\sqrt{2}})-1)^2}} \tag{153}$$

$$\approx 0.735 \tag{154}$$

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 **Optimality test** 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 $D_{\hat{s}_\pm,s}$.

$$\begin{aligned}
\mathrm{Cov}\left(\hat{s}, s\right) &= \left\langle \langle \hat{s}|s\rangle\, s\right\rangle_s - \langle \hat{s}\rangle \langle s\rangle \\
&= \left\langle \left\langle \hat{\phi}^2|s\right\rangle s\right\rangle_s - \left\langle \left\langle \hat{\phi}^2|s\right\rangle\right\rangle_s \bar{s} \\
&= \left\langle \Sigma_{\hat{\phi}\hat{\phi}|s} s\right\rangle_s - \left\langle \left\langle \Sigma_{\hat{\phi}\hat{\phi}}|s\right\rangle\right\rangle_s \bar{s} \\
&= \left\langle (H_{\hat{\phi}\hat{\phi}}+s)s\right\rangle_s - \left\langle (H_{\hat{\phi}\hat{\phi}}+s)\right\rangle_s \bar{s} \\
&= H_{\hat{\phi}\hat{\phi}}\bar{s} + \left\langle s^2\right\rangle_s - (H_{\hat{\phi}\hat{\phi}}+\bar{s})\bar{s} \\
&= \mathrm{Var}(s) \\
&= \tfrac{1}{2}(s_+^2+s_-^2) - \tfrac{1}{4}(s_++s_-)^2 \\
&= \tfrac{1}{4}\Delta s^2
\end{aligned} \tag{155}$$

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

$$\begin{aligned}
\mathrm{Cov}\left(\hat{s}_\pm, s\right) &= \left\langle \langle \hat{s}_\pm|s\rangle\, s\right\rangle_s - \langle \hat{s}_\pm\rangle \langle s\rangle \\
&= \langle m_s s\rangle_s - \langle m_s\rangle_s \bar{s} \\
&= \tfrac{1}{2}\left(m(s_+)s_+ + m(s_-)s_-\right) \\
&\quad - \tfrac{1}{4}\left(m(s_+)+m(s_-)\right)\bar{s} \\
&= \tfrac{1}{4}\Delta m \Delta s
\end{aligned} \tag{156}$$

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

$$\begin{aligned}
\mathrm{Var}\left(\hat{s}\right) &= \left\langle \hat{\phi}^4\right\rangle - \left\langle \hat{\phi}^2\right\rangle^2 \\
&= \left\langle \left\langle \hat{\phi}^4|s\right\rangle\right\rangle_s - \left\langle \left\langle \hat{\phi}^2|s\right\rangle\right\rangle_s^2 \\
&= \left\langle 3\Sigma^2_{\hat{\phi}\hat{\phi}|s}\right\rangle_s - \left\langle \Sigma_{\hat{\phi}\hat{\phi}|s}\right\rangle_s^2 \\
&= \tfrac{3}{2}\left(\Sigma^2_{\hat{\phi}\hat{\phi}|+} + \Sigma^2_{\hat{\phi}\hat{\phi}|-}\right) - \left(\tfrac{1}{2}\left(\Sigma_{\hat{\phi}\hat{\phi}|+} + \Sigma_{\hat{\phi}\hat{\phi}|-}\right)\right)^2 \\
&= \tfrac{1}{4}\left(5\Sigma^2_{\hat{\phi}\hat{\phi}|+} - 2\Sigma_{\hat{\phi}\hat{\phi}|+}\Sigma_{\hat{\phi}\hat{\phi}|-} + 5\Sigma^2_{\hat{\phi}\hat{\phi}|-}\right) \\
&= \tfrac{1}{4}(5(H_{\hat{\phi}\hat{\phi}}+s_+)^2 - 2(H_{\hat{\phi}\hat{\phi}}+s_+)(H_{\hat{\phi}\hat{\phi}}+s_-) \\
&\quad + 5(H_{\hat{\phi}\hat{\phi}}+s_-)^2) \\
&= \tfrac{1}{4}(8H^2_{\hat{\phi}\hat{\phi}} + H_{\hat{\phi}\hat{\phi}}(10s_+ - 2s_+ - 2s_- + 10s_-) \\
&\quad + 5s_+^2 - 2s_+s_- + 5s_-^2) \\
&= 2H^2_{\hat{\phi}\hat{\phi}} + 2H_{\hat{\phi}\hat{\phi}}(s_++s_-) \\
&\quad + \tfrac{1}{4}\left(5s_+^2 - 2s_+s_- + 5s_-^2\right)
\end{aligned} \tag{157}$$

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

$$\mathrm{Var}\left(\hat{s}\right) \approx \tfrac{5}{4}s_+^2 \tag{158}$$

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

$$\mathrm{Var}\left(\hat{s}_\pm\right) = 1 - \langle \hat{s}_\pm\rangle^2 = 1 - \bar{m}^2 \tag{159}$$

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

Combining these computations, we see that ratio of total correlations for binary $\hat{s}_\pm$ and continuous $\hat{s}$ is

$$\begin{aligned}\delta &= \frac{D_{\hat{s},s}}{D_{\hat{s}_\pm,s}} \\ &= \frac{\mathrm{Corr}(\hat{s}_\pm, s)}{\mathrm{Corr}\,(\hat{s}, s)} \\ &= \frac{\mathrm{Cov}(\hat{s}_\pm, s)}{\mathrm{Cov}\,(\hat{s}, s)} \sqrt{\frac{\mathrm{Var}\,(\hat{s})}{\mathrm{Var}\,(\hat{s}_\pm)}} \\ &\approx \sqrt{\tfrac{5}{4}} \frac{1}{1 - \frac{s_-}{s_+}} \frac{\Delta m}{\sqrt{1 - \bar{m}^2}} \end{aligned} \tag{160}$$

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. 121 and Eq. 144, we have

$$B_R^{\mathrm{opt},\pm} = \zeta \frac{D_{R,s}}{D_{\hat{s},s}} \tag{161}$$

From Eq. 160 we can adjust the

$$D_{\hat{s},s} = \delta D_{\hat{s}_\pm,s} \tag{162}$$

Combining these we have

$$B_R^{\mathrm{opt},\pm} = \frac{\zeta}{\delta} \frac{D_{R,s}}{D_{\hat{s}_\pm,s}} \tag{163}$$

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.

S.6.5 Nonlinear information in internal noise

S.7 Practical considerations in nonlinear choice correlation tests

S.7.1 Information and decoding efficiency in a redundant code

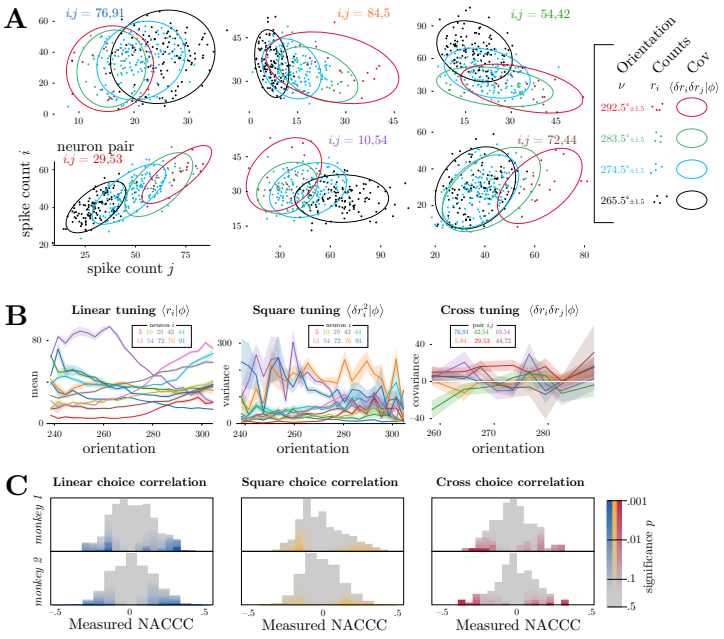

Figure S5: Internal noise covariance is only weakly tuned to orientation and insignificantly tuned to choice. **A**: Scatter plots of neural responses to multiple groups of trials, each group with nearly identical orientations (±1.5 deg). There are significant shifts in the means and variances of these response clouds, but changes in the correlations are not reliable. Cell pairs with strongest joint quadratic tuning were selected. **B**: For these same selected neurons, direct plots of linear and quadratic tunings confirm that the mean response is strongly tuned to orientation, the response variance is moderately tuned, and the internal noise cross-covariance is not tuned. Solid lines denote mean and shades denote standard deviations of each neural statistics. **C**: Histograms of choice correlations show that purely internal noise is not significantly correlated with choices ($p < 0.01$, two-sample Kolmogorov-Smirnov test for a choice-shuffled null distribution), unlike the nuisance-generated fluctuations seen in Figure 6D. To isolate the correlation of internal noise on choice, we compute the Normalized Average Conditional Choice Correlation (NACCC) where we condition on, and then average over, the *complete* stimulus (s, ν) rather than just on the task-relevant stimulus s as in Eq. 17. Individual choice correlations within the histograms are each colored by their significance according to their own null distribution (Methods **Application to neural data**).

When information is broadly distributed amongst many neurons, the contributions of individual neural responses or functions of those responses could be too small to measure. This potential problem is magnified for higher-order statistics, since there are so many of them. In particular, if there are N neurons in a population, there are $O(N^z)$ of them for a z-th order polynomial nonlinearity. However, for redundant codes, we show that vastly fewer higher-order statistics are required to capture the information.

Consider a redundant code arising from a linear cortical expansion [11]. When a large cortical population of size N inherits all of its information from a smaller upstream sensory neural population of size M, there will be significant redundancy in the recorded cortical neural population. For information contained in z-th order polynomial statistics, there will be only $\sim M^z$ degrees of freedom, which is much smaller than $\sim N^z$.

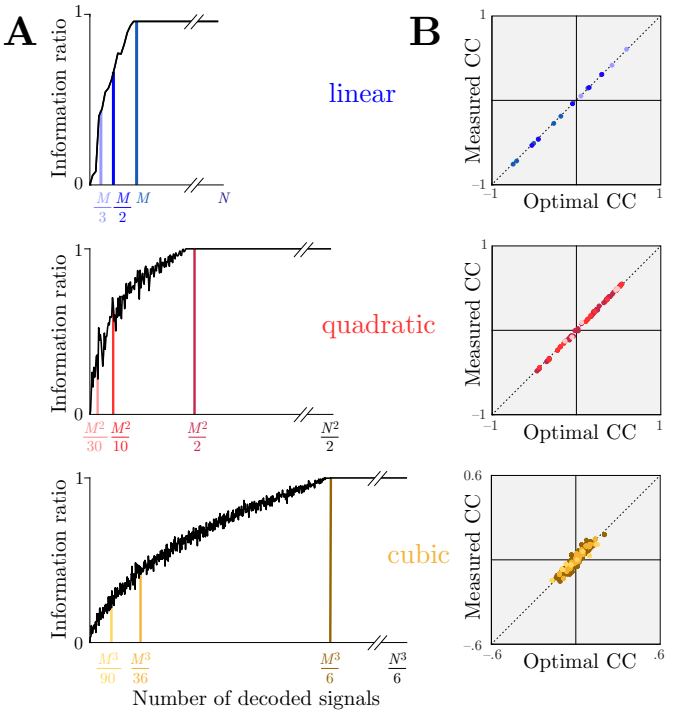

Figure S6: Consequences of redundancy on information and choice correlations. (**A**) Information decoded from various numbers of linear (blue), quadratic (red), and cubic (green) statistics of downstream neural population. Information saturates after decoding $\sim M^z$ units where M is the upstream population size and z is the order of the statistic. (**B**) Nonlinear choice correlations for recordings from a simulated optimal decoder under the same conditions. The brain is assumed to decode optimally while the experimenter only examines a subset of statistics. The slope of 1 predicted for optimal decoding is evident even when recording few informative neural statistics.

Figure S6 shows simulations of a cortical expansion from a smaller population of upstream neurons ($M = 15$) to a much larger population recorded downstream ($N = 100$), with 50000 trials. The upstream neural responses $\rho \in \mathbb{R}^M$ are first generated from cubic codes (Supplemental Materials S.1.4) with third-order sufficient statistics $\mathbf{T}(\rho) = \{\rho_i, \rho_i\rho_j, \rho_i\rho_j\rho_k\}$. This upstream population ρ expands noiselessly and linearly into the downstream population $\mathbf{r}$ according to $\mathbf{r} = A\rho$, where A is a 100×15 matrix whose elements are generated from a standard normal distribution. Because this expansion is linear, the sufficient statistics for $p(\mathbf{r}|s)$ are still polynomials up to third-order, $\mathbf{R}(\mathbf{r}) = \{r_i, r_ir_j, r_ir_jr_k\}$.

To compute the nonlinear information content of the downstream neurons, we estimate the stimulus from polynomial nonlinearities of $\mathbf{r}$. We first generate estimates $\hat{s}_1$ from $\mathbf{R}^{(1)} = \mathbf{r}$ directly. Then, to isolate information of second order, we removed this first-order information by subtracting the conditional mean responses $\langle r_i|\hat{s}_1\rangle$ given the first-order estimates from the full population (even unrecorded neurons) to obtain $\delta r_i = r_i - \langle r_i|\hat{s}_1\rangle$, and then compute second-order products $\mathbf{R}^{(2)} = \{\delta r_i \delta r_j\}$. To isolate third-order information, we remove the stimulus-dependent covariance $\Sigma_{\mathbf{r}}(\hat{s}_2) = \left\langle \delta\mathbf{r}\, \delta\mathbf{r}^\top | \hat{s}_2 \right\rangle$ with estimates $\hat{s}_2$ based on a quadratic decoder from the full population, to obtain whitened deviations $\mathbf{z} = \Sigma_{\mathbf{r}}(\hat{s}_2)^{-1}\delta\mathbf{r}$, and evaluate their skewness as $\mathbf{R}^{(3)} = \{z_iz_jz_k\}$. This approach is natural because it is essentially the inverse of the generation process we used for quadratic and cubic codes (Section **Quadratic encoding** and **Cubic encoding**). Note that this isolation is not necessary to evaluate the nonlinear information content or to apply our choice correlation test, but here it allows us to identify how information in different nonlinear orders scales with population size.

We then construct estimates from random subsets of these statistics and approximate their Fisher information content by their inverse variance over trials with the same stimulus, $\mathbf{R}^{(z)}_{\text{subset}}$, $J_{\mathbf{R}^{(z)}_{\text{subset}}} = 1/\sigma^2_{\hat{s}^{\text{subset}}_z}$. Figure S6A shows the fraction of the information extracted from the selected subset of statistics to the information in all of the z-th order statistics, $J_{\mathbf{R}^{(z)}_{\text{subset}}}/J_{\mathbf{T}_z}$, plotted against the number of the decoded units. From the simulation, we find that the information ratio generically saturates at 1 after decoding M^z units, so we do not need to decode all N^z statistics of order z from the large downstream population. If the cortical expansion introduces noise, then the information will not

have completely saturated by M^z statistics, but most of the information will be extracted by recordings of that size [3].

In principle, the number of recorded units might also affect the accuracy of our nonlinear choice correlation test in assessing decoding efficiency. Figure S6B plots measured versus predicted linear, quadratic, and cubic choice correlations. Color shading indicates different numbers of decoded statistics, as indicated in Figure S6A. We found that in this testing regime the decoding efficiency revealed by the slope in the choice correlation test is not substantially affected by changing the number of measured statistics.

S.7.2 Assessing optimal decoding from estimation versus discrimination

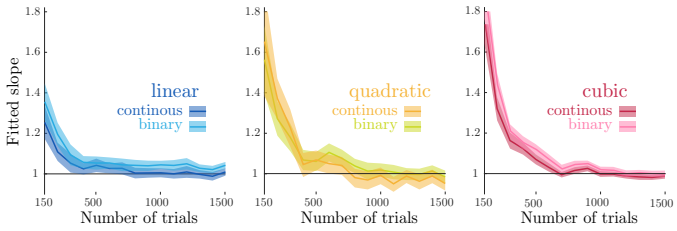

Figure S7: Optimal decoding can be revealed with equal fidelity from either fine continuous estimation or fine binary discrimination. We simulate neurons from random quadratic codes and decode them optimally to generate continuous estimates or binary choices. The three panels plot the slope of the relationship between predicted and measured nonlinear choice correlations, as a function of the number of trials. Different panels and colors denote the nonlinear types and task types. Means (denoted by solid lines) and 95% confidence interval (denoted by shades) for fitted slopes were compute by repeating the procedures for 30 times independently.

Here we use simulations to compare the outcomes of the nonlinear choice correlation test under continuous estimation versus binary discrimination. This simulation is based on quadratic codes (Supplemental Materials S.1.3) with 10 neurons whose response means and covariances contain information about the stimulus.

For the continuous estimation task, we assume the brain decodes neural activity through a weighted sum of linear and quadratic statistics $\mathbf{R}(\mathbf{r})$ minimizing the variance of an locally unbiased decoder, $\hat{s} = \mathbf{w}^\top \mathbf{R}(\mathbf{r}) + c$, with weights $\mathbf{w} \propto \Gamma^{-1}\mathbf{F}'$. Nonlinear choice correlations are measured directly in simulation, and the optimal decoding predictions are calculated by $C^{\text{opt}}_{R_k,\hat{s}} = \sqrt{\sigma^2_{\hat{s}}/\sigma^2_{\hat{s},R_k}}$ (Eq. 19), where $\sigma^2_{\hat{s}}$ is the variance of the locally optimal unbiased estimator based on the entire population, and $\sigma^2_{\hat{s},R_k}$ is the same for an estimator built only from R_k.

For binary discrimination, we assume the brain decodes neural activity using the same optimal decoding weights as used for the continuous estimation, but we now threshold it at a reference s_0 to obtain a binary output choice, $\hat{s}_\pm = \text{sgn}(\hat{s} - s_0)$. The measured choice correlations are now the correlations between the binary choice and nonlinear neural statistics. The optimal choice correlations are calculated by $C^{\text{opt}}_{R_k,\hat{s}_\pm} = \zeta^{\text{fine}} d'_{R_k}/d'$, where d' are sensitivities of the discriminators derived from the full population or just from R_k, and ζ^{fine} is the correction factor defined in Eqs. 152 and 154.

We then test the accuracy of the nonlinear choice correlation test in the two task settings while varying the number of trials. Figure S7 shows that the fitted slope between the measured and predicted choice correlations is biased to give a larger slope when there are few trials, with a slightly greater bias in binary discrimination than continuous estimation. These results reveal that binary tasks provide similar utility as continuous estimation in these settings.

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

[5] 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.

[6] Paradiso M (1988) A theory for the use of visual orientation information which exploits the colum-

nar structure of striate cortex. Biological cybernetics 58: 35–49.

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

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

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

[10] 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.

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

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

[13] 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.