RICE UNIVERSITY

Essential nonlinear properties in neural decoding

by

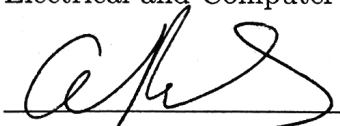
Qianli Yang

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

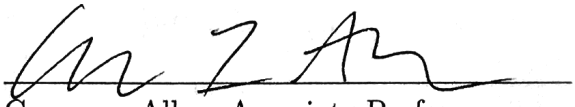
Ph.D.

Approved, Thesis Committee:

Xaq Pitkow, Chair, Assistant Professor,
Electrical and Computer Engineering



Ankit Patel, Assistant Professor,
Electrical and Computer Engineering



Genevera Allen, Associate Professor,
Statistics

Houston, Texas

May, 2018

ABSTRACT

Essential nonlinear properties in neural decoding

by

Qianli Yang

The sensory data about most natural task-relevant variables is confounded by task-irrelevant sensory variations, called nuisance variables. To be useful, the sensory signals that encode the relevant variables must be untangled from the nuisance variables through nonlinear recoding transformations, before the brain can use or decode them to drive behaviors. The information to be untangled is represented in the cortex by the activity of large populations of neurons, constituting a nonlinear population code.

In this thesis I provide three major contributions in theoretical neuroscience.

First, I provide a new way of thinking about nonlinear population codes and nuisance variables, leading to a theory of nonlinear feedforward decoding of neural population activity. This theory obeys fundamental mathematical limitations on information content that are inherited from the sensory periphery, producing redundant codes when there are many more cortical neurons than primary sensory neurons.

Second, and critically for experimental testing, I provide a theory that predicts a simple, easily computed quantitative relationship between fluctuating neural activity and behavioral choices if the brain uses its nonlinear population codes optimally:

more informative patterns should be more correlated with choices. To validate this theory, I show that when primates discriminate between a wide or narrow distribution from which oriented images could be sampled, quadratic statistics of primary visual cortex activity match this predicted pattern.

Third, I contribute new concepts and methods to characterize behaviorally relevant nonlinear computation downstream of recorded neurons. Since many neural transformations can generate the same behavioral output, I will define a new concept of equivalence classes for neural transformations based on the degeneracy of the decoding. This suggests that we can understand the neural transformations by picking a convenient nonlinear basis that approximates the actual neural transformation up to an equivalence relation given by the intrinsic uncertainty, instead of trying to reproduce the biophysical details. Then I extend the concept of redundant codes [1] to a more general scenario: when different subsets of neural response statistics contain limited information about the stimulus. This extension allows us understand the neural computation at the representational level [2] — extracting representations for different subsets of neural nonlinear statistics, characterizing how these representations transform the information about task-relevant variables and studying the coarse-grained computations on these representations.

Acknowledgement

I want first to thank my family. They are my firmest support. I cannot become who I am without the love from them. My father teaches me to be frank and honest with the friends and myself. My mother is optimistic all the time, no matter how challenging the situation is, which inspires to finish this thesis. They shape two sides of my sprit—one is outward towards the world, and another is inward to my mind. My grandmothers raised me when I was a child. Both of them passed away during my Ph.D. pursuing. I'm sorry that I didn't make their funeral. They will always be in my heart. It's also a great honor to meet my wife's family. Thanks for their trust and understanding so that I can join them as a new member of the family.

I was extraordinarily fortunate to have Xaq Pitkow as my adviser. I've learned so much from Xaq. Thinking about principles of the brain with Xaq is a fantastic experience.

I also want to thank my defense committees for valuable comments and suggestions on my work. Many thanks to Ankit Patel and Genevera Allen.

It's a tremendous honor to be the first member in Xaq's lab. Everyone in the lab contributes different components to shape the lab. I learned a vast variety of perspectives from the discussions. Many thanks to all the postdocs: Aram Giahi-Saravani, James Bridgewater, KiJung Yoon, Zhengwei Wu, Emin Orhan, Baptiste Caziot, Zhi Li, Giuseppe Vinci. Many thanks to all the graduate students: Rajkumar Raju, Kaushik Lakshminarasimhan, Saurabh Daptardar, Yicheng Fei.

Many thanks to all the collaborators: Andreas Tolias, Edgar Walker, James Cotton, Valentin Dragoi, Ming Hu, Arun Parajuli. I learned so much by combining my theory with their amazing experiments. This drives me to think more deeply about

how the brain works.

Last but not least, I want to thank my wife, Jing Xia. I was blessed to have her accompanying me during the journey of pursuing Ph.D.

Contents

Chapter 1

Introduction

Our percepts are not a direct copy of elements in the outside world. Instead, the goal of our brains is to make sense of its sense data and extract information about these external world variables. Hermann von Helmholtz was an early proponent of the idea that our brain uses unconscious inference to reach this goal: the perceptual systems of our brain construct neural representations of causal variables that account for the data collected by our senses [3]. Perception is challenging because the relevant variables are often ambiguous and rarely directly observable in sense data. To decode task-relevant variables from sensory observations, the brain must eliminate task-irrelevant variables, often called 'nuisance variables', that affect those observations. Nuisance variable is defined as the property in the world that alters how task-relevant stimuli appear but is, itself, irrelevant for the current task. For natural tasks, the varying nuisance variable make the neural representations of the task-relevant variables entangled. Disentangling them generally requires nonlinear computation [4]. For example, accurately detecting an object boundary in an image requires contrast invariance: an edge appears when the foreground object is darker *or* lighter than the background, yet any linear function will exhibit opposite responses in these two cases.

To understand how the brain realizes its goal of extracting task-relevant causal variables, we need to build a theory of how the brain implements these computations. My theory will need to consider key anatomical constraints of the brain. Furthermore,

we need to provide an experimental test of the theory.

In this thesis, I contribute a general mathematical framework in which distributed nonlinear computation can be understood and analyzed. This framework describes three major topics:

- Robust nonlinear neural codes
- Nonlinear choice correlation
- Coarse-grained nonlinear computation

The first part of this thesis will illustrate the theory of robust nonlinear neural codes. Compared to the previous work in nonlinear population codes [5, 6], this theory obeys the fundamental mathematical limitations on information content that is inherited from the sensory periphery. This inheritance structure produces a redundant code when there are many more cortical neurons than sensory neurons [7–9]. I generalize the recently-introduced concept of information-limiting correlation [1] to repair the error of violating the data processing inequality—i.e., the information in the cortex should not exceed the information in the sensory inputs.

The second part of the thesis will introduce a new statistical measure, nonlinear choice correlation, which is defined as the correlation coefficient between nonlinear functions of neural activity and the stimulus estimate. Choice correlation is widely used in neuroscience experiments as a proxy for how much a neuron contributes to an animal's behavior [10–18]. This proxy measure is confounded by correlations between neurons, which can account for why neurons might appear to be correlated with a behavioral choice. Previous work recognized that knowledge of neural correlations could help disambiguate between correlation and influence [19, 20], and moreover could reveal the efficiency of decoding [19]. However, this approach only worked for

linear computation. The nonlinear choice correlation test generalizes their methods to the regime of nonlinear population codes. It predicts a simple, easily computed relationship between fluctuating neural activity and behavioral choices if the brain uses its nonlinear population codes optimally. The optimality relationship predicted by this test holds even if the specific nonlinearities used by the brain differ from the exact nonlinearity defined by the neural encoding, as long as the decoder optimally uses all information. Even if the decoding is not strictly optimal, the information redundancy in the neural population can makes the test less stringent.

The third part of this thesis will develop concepts and methods to characterize behaviorally relevant nonlinear computation downstream of recorded neurons. I will start by proposing a new concept of equivalence classes for neural transformations. The number of equivalent computations expands in the presence of uncertainty, which makes it both more difficult and less relevant to distinguish fine features of the neural representations. This suggests that we can understand the neural transformations by picking a convenient nonlinear basis that approximates the actual neural transformation up to an equivalence relation given by the intrinsic uncertainty, instead of trying to reproduce the biophysical details. Then I extend the concept of redundant codes [1] to a more general scenario: when different subsets of neural response statistics contain limited information about the stimulus. This extension allows us understand the neural computation at the representational level [2] — extracting representations for different subgroup of neural nonlinear statistics, characterizing how these representations transform the information about task-relevant variables and studying the coarse-grained computations on these representations [21].

Chapter 2

Robust nonlinear neural codes

曲径通幽处，禅房花木深。

—《题破山寺后禅院》常建

The winding path leads to a secluded place, where the Zen Garden is hidden deep among lush flowery vegetation.

— 'The Rear Zen Garden of a ruined Mountain Temple', Chang Jian

2.1 Introduction

When the average responses of neurons are well-tuned to a stimulus of interest, it is easy for an animal to use, or 'decode', their relevant information. 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 task variable: natural variation in other sensory stimuli also influence their responses. This can dilute or even abolish the mean tuning to the relevant stimulus. Like I cannot simply enjoy the beauty of Zen without walking through the winding path, the brain cannot decode the information about the stimulus without performing nonlinear computation.

To see this problem in a simple case, imagine a simplified model of a visual neuron that includes an oriented edge-detecting linear filter followed by additive noise, with a Gabor receptive field like simple cells in the primary visual cortex (Figure 2.1A). If

an edge is presented to this model neuron, different rotation angles will change the overlap, producing a different mean. This neuron is tuned to orientation.

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

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

To estimate the variance, and thereby the orientation itself, the brain can compute the square of the linear responses. This would allow the brain to estimate the orientation independently from polarity. This is consistent with the well-known energy model of complex cells in primary visual cortex, which use squaring nonlinearities to achieve invariance to the polarity of an edge [22].

Generalizing from this example, I identify edge polarity as a 'nuisance variable' — a property in the world that alters how task-relevant stimuli appear but is, itself, irrelevant for the current task (here, perceiving orientation). Other examples of nuisance variables include the illuminant for guessing surface color, position for object recognition, expression for face identification, or pitch for speech recognition. Nuisance variables generally make it hard to extract the task-relevant variables from sense data, which is the central task of perception [4, 23–25]. (Of course, what is a nuisance for one task might be a target variable in another task, and vice versa.)

The prevailing neuroscience view of this disentangling process is deterministic: the output of a complex (often multi-stage) nonlinear function identifies the variables of interest [4, 23, 26]. Here I take a statistical perspective: the brain learns from its history of sensory inputs which statistics of its many sense data can be used to extract the task-relevant variable. In the orientation estimation task above, the relevant statistic was not the mean but the variance.

In this chapter, I will provide a new way of thinking about nonlinear population codes and nuisance variables, leading to a theory of robust nonlinear neural codes. I will start by specifying the mathematical framework by modeling a feedforward processing chain for the brain. Then I will define signal and noise elements in the neural responses according to their dependence on the stimulus. This naturally leads me to broaden the linear neural model to the nonlinear neural model. Last, I will generalize the recently-introduced concept of information-limiting correlation [1] to repair the error of violating the data processing inequality in the previous work on nonlinear population codes [5, 6].

2.2 Results

2.2.1 Task, stimulus, neural responses, action

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

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

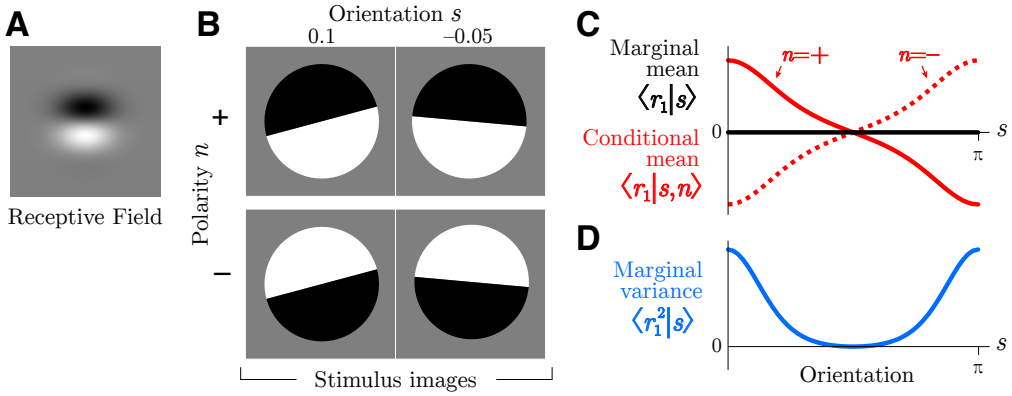


Figure 2.1 : Simple nonlinear code for orientation induced by two polarities. (**A**) Receptive field for a linear neuron. (**B**) Four example images, each with an orientation $s \in [0, \pi)$ and a polarity $\boldsymbol{n} \in \{-1, +1\}$. (**C**) Even though the mean response of the linear neuron is tuned to orientation if polarity were specified (conditional mean, red), the mean response is untuned when the polarity is unknown and could take any value (marginal mean, black). (**D**) Tuning is recovered by the marginal variance even if the polarity is unknown (blue).

$\boldsymbol{n}$ stands for nuisance). Together, these stimulus properties determine a complete sensory input that drives some responses $\boldsymbol{r}$ in a population of N neurons according to the distribution $p(\boldsymbol{r}|s, \boldsymbol{n})$. The neural responses $\boldsymbol{r}$ can be any measures of neural activity. Here I take here to be the spike counts of each neuron in a specified time window.

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

$$(s, \boldsymbol{n}) \rightarrow \boldsymbol{r} \rightarrow \boldsymbol{R} \rightarrow \hat{s} \tag{2.1}$$

I model the brain's estimate as a linear function of the downstream responses $\boldsymbol{R}$. Ultimately these estimates are used to generate an action that the experimenter can observe. Here I assume that the task is local or fine-scale estimation: the subject must

directly report its estimate $\hat{s}$ for the relevant stimuli near a reference s_0. I measure performance by the variance of this estimate, $\sigma^2_{\hat{s}}$.

I assume that I have recorded activity only from some of the upstream neurons, so I don't have direct access to $\boldsymbol{R}$, only $\boldsymbol{r}$. Nonetheless I would like to learn something about the downstream computations used in decoding. In Section 2, I show how to use the statistics of cofluctuations in $\boldsymbol{r}$ and $\hat{s}$ to estimate the quality of nonlinear decoding.

2.2.2 Signal and noise

The population response reflects both *signal* and *noise*, where signal is the repeatable stimulus-dependent aspects of the response, and noise reflects trial-to-trial variation. Conventionally in neuroscience, the signal is often thought to be the stimulus dependence of the *average* response, *i.e.* the tuning curve $\boldsymbol{f}(s) = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s) = \langle \boldsymbol{r}|s \rangle$ (angle brackets denote an average over all responses given the condition after the vertical bar). Below I 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 the response distribution $p(\boldsymbol{r}|s, \boldsymbol{n})$ when everything about the stimulus is fixed. This could also include uncontrolled variation in internal states [27–30], like attention, motivation, or wandering thoughts. External noise is variability generated by the external world, or nuisance variables, such as the positions of all dots in a random dot kinematogram [31] or the polarity of an edge (Figure 2.1). External noise leads to a neural response distribution $p(\boldsymbol{r}|s)$ where only the relevant variables are held fixed. Both types of

noise can lead to uncertainty about the true stimulus.

Trial-to-trial variability can of course be correlated across neurons. Neuroscientists often measure two types of second-order correlations: signal correlations and noise correlations [6, 32–39]. Signal correlations measure shared variation in responses $\boldsymbol{r}$ averaged over the set of stimuli s: $\rho_{\text{signal}} = \text{Corr}(\boldsymbol{r})$. (Internal) noise correlations measure shared variation that persists even when the stimulus is completely identical, nuisance variables and all: $\rho_{\text{noise}}(s, \boldsymbol{n}) = \text{Corr}(\boldsymbol{r}|s, \boldsymbol{n})$.

For multidimensional stimuli, however, these are only two extremes on a spectrum, depending on how many stimulus aspects are fixed across the trials to be averaged. I propose an intermediate type of correlation: *nuisance correlations*. Here I fix the task-relevant stimulus variable(s) s, and average over the nuisance variables $\boldsymbol{n}$: $\rho_{\text{nuisance}} = \text{Corr}(\boldsymbol{r}|s)$. Just as signal correlations don't mean correlations between signals, nuisance correlations are not correlations between nuisance variables, but rather between neural responses induced by the external noise or nuisance variation. Of course nuisance correlations will be task-dependent, since the task determines which variables are nuisance and which are relevant [40, 41].

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 [42] such as the variance of the response covariance (Figure 2.2D, Table 2.1). Stimulus-dependent correlations, principally due to nuisance variation, lead naturally to nonlinear population codes, as I will explain below.

2.2.3 Nonlinear encoding by neural populations

Most accounts of neural population codes actually address *linear* codes, in which the mean response is tuned to the variable of interest and completely captures all signal about it [1, 19, 43–45]. I 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, Section 2.5.2) is a linear function of the response. Linear codes for different variables may arise early in sensory processing, or after many stages of computation [4, 25].

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

It is illuminating to take a statistical view: unlike a linear code, the information is not encoded in mean neural responses but instead by higher-order statistics of responses [5, 6]. These functional and statistical views are naturally linked because estimating higher-order statistics requires nonlinear operations. For instance, information from a stimulus-dependent covariance $Q(s) = \langle \boldsymbol{r}\boldsymbol{r}^\top | s \rangle$ can be decoded by quadratic operations $\boldsymbol{R} = \boldsymbol{r}\boldsymbol{r}^\top$ [41, 46, 47]. Table 2.1 compares the relevant neural response properties for linear and nonlinear codes.

A simple example of a nonlinear code is the exclusive-or (XOR) problem. Given the responses of two binary neurons, r_1 and r_2, I would like to decode the value of a task-relevant signal $s = \mathrm{XOR}(r_1, r_2)$ (Figure 2.2A). I don't care about the specific value of r_1 by itself, and in fact r_1 alone tells us nothing about s. The same is true for r_2. The signal is actually reflected in the trial-by-trial *correlation* between r_1 and r_2: when they are the same then $s = -1$, and when they are opposite then $s = +1$. The correlation, and thus the relevant variable s, can be estimated nonlinearly from r_1 and r_2 as $\hat{s} = -r_1 r_2$.

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

Table 2.1 : Neural response properties relevant for linear and nonlinear codes. In each case, the brain must estimate the stimulus from a single example of neural data, but the relevant function of that data is linear for linear codes, and nonlinear for nonlinear codes. The noise and signal can be quantified by the corresponding covariance and stimulus-dependent changes in the corresponding means (*i.e.* the tuning curve slope).

Some experiments have reported stimulus-dependent internal noise correlations that depend on the signal, even for a completely fixed stimulus without any nuisance variation [48–52]. Other experiments have turned up evidence for nonlinear population codes by characterizing the nonlinear selectivity directly [26, 53, 54].

More typically, however, stimulus-dependent noise correlations arise from external noise, or nuisance correlations. In the introduction (Figure 2.1) I showed a simple orientation estimation example in which fluctuations of an unknown contrast eliminate the orientation tuning of mean responses, relegating the tuning to variances. Figure 2.2B–E shows a slightly more sophisticated version of this example, where instead of two image polarities, I introduce spatial phase as a continuous nuisance variable. This again eliminates mean tuning, but introduces nuisance covariances that are orientation tuned.

One might object that although the nuisance covariance is tuned to orientation, a subject cannot compute the covariance on a single trial because it does not experience all possible nuisance variables to average over. This objection stems from a conceptual error that conflates the tuning (signal) with the raw sense data (signal+noise). In

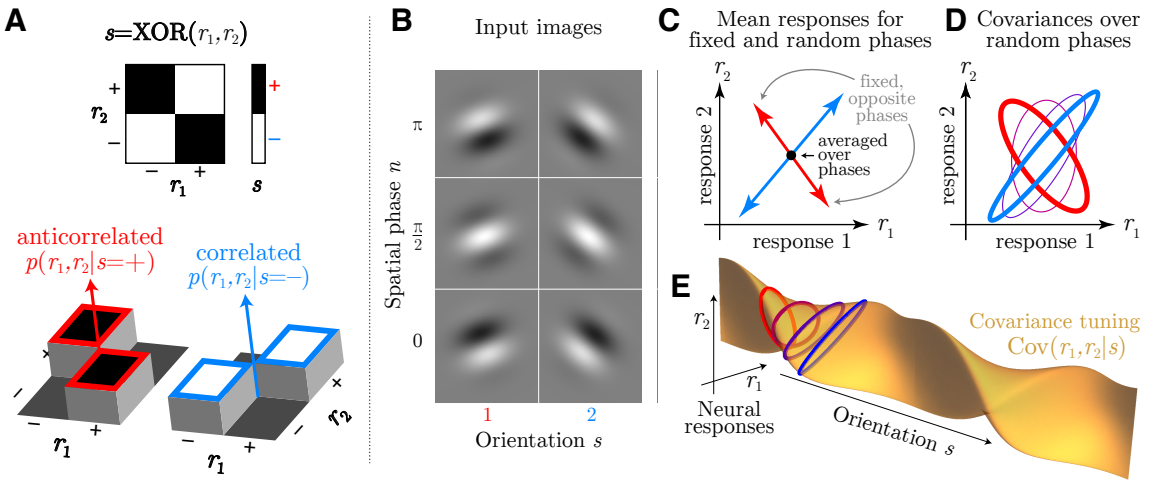


Figure 2.2 : Nonlinear codes. **A**: Simple example in which a stimulus s is the XOR of two neural responses (top). Conditional probabilities $p(r_1, r_2|s)$ of those responses (bottom) show they are anti-correlated when $s = +1$ (red) and positively correlated when $s = -1$ (blue). This stimulus-dependent correlation between responses creates a nonlinear code. The remaining panels show that a similar stimulus-dependent correlation emerges in orientation discrimination with unknown spatial phase. **B**: Gabor images with two orientations and three spatial phases. **C**: Mean responses of linear neurons with Gabor receptive fields are sensitive to orientation when phase is fixed (arrows), but point in different directions for different spatial phases. When phase is an unknown nuisance variable, this mean tuning therefore vanishes (black dot). **D**: The response covariance $\mathrm{Cov}(r_1, r_2|s)$ between these linear neurons is tuned to orientation even when averaging over spatial phase. Response covariances for four orientations are depicted by ellipses. **E**: A continuous view of the covariance tuning to orientation for a pair of neurons.

linear codes, the subject does not have access to the tuned mean response $\langle \boldsymbol{r}|s\rangle$, just a noisy single-trial version of the mean, namely $\boldsymbol{r}$. Analogously, the subject does not need access to the tuned covariance, just a noisy single-trial version of the covariance, $\boldsymbol{r}\boldsymbol{r}^\top$ (Table 2.1). In this simple example, the nuisance variable ensures that this quadratic statistic contains relevant information.

2.2.4 Redundant codes

The previous work in nonlinear population codes [5,6] derive extensive nonlinear information. It violates the data processing inequality—i.e., the information in the cortex should not exceed the information in the sensory inputs. This error is because they [5,6] treat large cortical populations in isolation from the smaller sensory population that would naturally provide its input. However, when a network inherits information from a much smaller input population, the expanded neural code becomes highly redundant: the brain cannot have more information than it receives. Noise in the input is processed by the same pathway as the signal, and this generates noise correlations that can never be averaged away, which is named as 'information-limiting correlations' in [1].

Previous work [1] characterized linear information-limiting correlations for fine discrimination tasks by decomposing the noise covariance into $\Sigma = \Sigma_0 + \epsilon \boldsymbol{f}'\boldsymbol{f}'^\top$, where ϵ is the variance of the information-limiting component and Σ_0 is noise that can be averaged away with many neurons.

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

2.2.2). Analogous to the linear case, these correlations can be locally decomposed as

$$\Gamma = \mathrm{Cov}(\boldsymbol{R}(\boldsymbol{r})|s) = \Gamma_0 + \epsilon \boldsymbol{F}' \boldsymbol{F}'^{\top} \tag{2.2}$$

where ϵ is again the variance of the information-limiting component, and Γ_0 is any other covariance which can be averaged away in large populations. The information-limiting noise bounds the estimator variance $\sigma^2_{\hat{s}}$ to no smaller than ϵ even with optimal decoding. Neither additional neurons nor additional decoded statistics can improve performance beyond this bound.

2.3 Discussion

This study introduced a theory of nonlinear population codes, grounded in the natural computational task of separating relevant and irrelevant variables. The theory considers both encoding and decoding — how stimuli drive neurons, and how neurons drive behavioral choices. Unlike previous theories [6,46], mine remains consistent with biological constraints due to the large cortical expansion of sensory representations by incorporating redundancy through information-limiting correlations.

2.3.1 Nonlinear decoding or switched linear decoding?

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

A discrimination task with a changing class boundary [32,55,56] is, in principle, a nonlinear task. But if the class boundary is changed too slowly, perhaps changing only on different days, then the brain may well re-learn its weights rather than performing some nonlinear decoding of recent activity. A better experimental design for revealing nonlinear computation for task context would be randomly changing the tasks, either cued [57] or even uncued [58], on a short enough time scale that the recent neural activity affects the class boundary.

2.4 Methods

2.4.1 Orientation estimation with varying spatial phase

Figure 2.1 illustrates how nuisance variation can eliminate a neuron's mean tuning to relevant stimulus variables, relegating the neural tuning to higher-order statistics like covariances. In this example, the subject estimates the orientation of a Gabor image, $G(\boldsymbol{x}|s,n)$, where $\boldsymbol{x}$ is spatial position in the image, and s and n are the orientation and spatial phase of the image, respectively (Section 2.5.3). The model visual neurons are linear Gabor filters like idealized simple cells in primary visual cortex, corrupted by additive white Gaussian noise. Their responses are thus distributed as $\boldsymbol{r} \sim P(\boldsymbol{r}|s,n) = \mathcal{N}(\boldsymbol{r}; \boldsymbol{f}(s,n), \epsilon I)$, where ϵ is the noise variance and the mean $\boldsymbol{f}(s,n) = \langle \boldsymbol{r}|s,n\rangle = \sum_{\boldsymbol{r}} \boldsymbol{r}\, p(\boldsymbol{r}|s)$ is determined by the overlap between the image and the receptive field.

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

When the spatial phase varies, however, each mean response to a fixed orien-

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

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

$$\mathrm{Cov}_{ij}(\boldsymbol{r}|s) = \int dn\, (\mathrm{Cov}_{ij}(\boldsymbol{r}|s,n) + \delta f_i(s,n)\delta f_j(s,n))p(n) \tag{2.3}$$

where $\delta f_i(s,n) = f_i(s,n) - \langle f_i(s,n)\rangle_n$. Section 2.5.3 shows in detail how $\mathrm{Cov}_{ij}(\boldsymbol{r}|s)$ is tuned to s.

2.4.2 Exponential family distribution and sufficient statistics

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

$$p(\boldsymbol{r}|s) = b(\boldsymbol{r}) \exp(\boldsymbol{H}(s) \cdot \boldsymbol{R}(\boldsymbol{r}) - A(s)) \tag{2.4}$$

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

Estimation and inference are closely connected in the exponential family. In Section 2.5.2, I show that the optimal local estimation can be achieved by linearly de-

coding the nonlinear sufficient statistics, $\hat{s} = \boldsymbol{w}^T \boldsymbol{R}(\boldsymbol{r}) + c$. The following decoding weights minimize the variance of an unbiased decoder,

$$\boldsymbol{w}_{\text{opt}} = \frac{\boldsymbol{H}'(s)}{J} = \frac{\Gamma^{-1}\boldsymbol{F}'}{\boldsymbol{F}'^{\top}\Gamma^{-1}\boldsymbol{F}'} \tag{2.5}$$

where $\boldsymbol{F}' = \partial \langle \boldsymbol{R}(\boldsymbol{r})|s \rangle / \partial s$ is the sensitivity of the statistics to changing inputs, and $\Gamma = \text{Cov}(\boldsymbol{R}|s)$ is the stimulus-conditioned response covariance which generally includes nuisance correlations (Sections 2.2.2).

The variance of this unbiased local estimator from the neural responses is lower-bounded by the inverse Fisher information. For exponential family distributions with nonlinear sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$, the Fisher information is [42] (Section 2.5.1.1)

$$J = \boldsymbol{F}'^{\top}\Gamma^{-1}\boldsymbol{F}' \tag{2.6}$$

2.4.3 Quadratic encoding

In a quadratic coding model, the distribution of neural responses is described by the exponential family with up to quadratic sufficient statistics, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j\}$ for $i, j \in \{1, \ldots, N\}$. A familiar example is the Gaussian distribution with stimulus-dependent covariance $\Sigma(s)$. In order to demonstrate the coding properties of a purely nonlinear neural code, here I assume that the mean tuning curve $f(s)$ and the stimulus-conditional covariances $\Sigma_{ij}(s)$ depend smoothly on the stimulus. I can quantify the information content of the neural population using Equation 2.6.

2.4.4 Cubic encoding

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

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

$$p(\boldsymbol{z}|s) = \tfrac{1}{Z} \exp\left(-\|\boldsymbol{z}\|^4 + \gamma s\, z_i z_j z_k\right) \tag{2.7}$$

I approximate this distribution by a mixture of four Gaussians. The mixture is chosen to reproduce the tetrahedral symmetry of the cubic distribution (Figure 2.4), which allows the cubic statistics of responses to be stimulus dependent, leaving stimulus-independent quadratic and linear statistics.

To generate larger multivariate cubic codes (Figure 2.4), for simplicity, I assume that 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, I shift and scale these cubic samples $\boldsymbol{z}$ in a manner dependent on s:

$$\boldsymbol{r} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\boldsymbol{z} \tag{2.8}$$

where $\boldsymbol{f}(s)$ and $\Sigma(s)$ describes the desired signal-dependent mean and covariance (see Section 2.5.5).

2.4.5 Information-limiting correlations

Only specific correlated fluctuations limit the information content of large neural populations [59]. These fluctuations can ultimately be referred back to the stimulus as $\boldsymbol{r} \sim p(\boldsymbol{r}|s + ds)$, where ds is zero mean noise, whose variance $1/J_\infty$ determines the asymptotic variance of any stimulus estimator. These information-limiting correlations for nonlinear computation can be characterized by the covariance of the sufficient

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

$$\Gamma = \Gamma_0 + 1/J_\infty \boldsymbol{F}(s)'\boldsymbol{F}(s)'^{\top} \tag{2.9}$$

Here Γ_0 is any covariance of $\boldsymbol{R}$ that does *not* limit information in large populations. Substituting this expression into (2.6) for the nonlinear Fisher Information, I obtain

$$J = \boldsymbol{F}'\Gamma^{-1}\boldsymbol{F}' = \frac{1}{1/J_\infty + 1/J_0} \tag{2.10}$$

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

2.5 Supplemental material

2.5.1 Exponential family distributions

The probability density function of an exponential family distribution is

$$p(\boldsymbol{r}|s) = b(\boldsymbol{r}) \exp\left(\boldsymbol{H}(s)^{\top}\boldsymbol{R}(\boldsymbol{r}) - A(s)\right) \tag{2.11}$$

where $\boldsymbol{H}(s)$ are the natural parameters, $\boldsymbol{R}(\boldsymbol{r})$ are the sufficient statistics, $A(s)$ and $Z(\boldsymbol{r})$ are the log normalizer and underlying base measure. The statistics $\boldsymbol{R}(\boldsymbol{r})$ are called sufficient it contains all the information needed to compute the estimate of the natural parameter $\boldsymbol{H}(s)$.

2.5.1.1 Fisher information

In this section, I compute the Fisher information $J(s)$ for a stimulus-conditioned response distribution $p(\boldsymbol{r}|s)$ in the exponential family with sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$. I can denote the mean of the sufficient statistics as $\boldsymbol{F}(s) = \langle \boldsymbol{R}(\boldsymbol{r})|s \rangle$. The Fisher information is given by

$$J = -\left\langle \frac{\partial^2}{\partial s^2} \log p(\boldsymbol{r}|s) \right\rangle_{p(\boldsymbol{r}|s)} \tag{2.12}$$

$$= \left\langle \left(\frac{\partial}{\partial s} \log p(\boldsymbol{r}|s) \right)^2 \right\rangle_{p(\boldsymbol{r}|s)} \tag{2.13}$$

The mean of the sufficient statistics, $\langle \boldsymbol{R}|s \rangle$ can be obtained by differentiating $A(s)$ by the natural parameters $\boldsymbol{H}(s)$

$$\boldsymbol{F} = \frac{\partial A(s)}{\partial \boldsymbol{H}(s)} \tag{2.14}$$

Equation 2.14 can give us the first and second derivatives of $A(s)$ with respect to s

$$A' = \sum_i \frac{\partial A}{\partial H_i} \frac{dH_i}{ds} = \boldsymbol{H}'^{\top} \boldsymbol{F} \tag{2.15}$$

$$A'' = \boldsymbol{H}''^{\top} \boldsymbol{F} + \boldsymbol{H}'^{\top} \boldsymbol{F}' \tag{2.16}$$

Thus, I can express the Fisher information in two ways:

$$J = -\left\langle \frac{\partial^2}{\partial s^2} \log P(\boldsymbol{r}|s) \right\rangle_{P(\boldsymbol{r}|s)} \tag{2.17}$$

$$= A'' - \boldsymbol{H}''^{\top} \boldsymbol{F} \tag{2.18}$$

$$= \boldsymbol{H}'^{\top} \boldsymbol{F}' \tag{2.19}$$

and

$$J = \left\langle \left(\frac{\partial}{\partial s} \log P(\boldsymbol{r}|s) \right)^2 \right\rangle_{P(\boldsymbol{r}|s)} \tag{2.20}$$

$$= \boldsymbol{H}'^{\top} (\langle \boldsymbol{R}\boldsymbol{R}^{\top} \rangle - \boldsymbol{F}\boldsymbol{F}^{\top}) \boldsymbol{H}' \tag{2.21}$$

$$= \boldsymbol{H}'^{\top} \Gamma \boldsymbol{H}' \tag{2.22}$$

where $\Gamma = \mathrm{Cov}[\boldsymbol{R}(\boldsymbol{r})|s]$. Combining the two expressions, I have

$$\boldsymbol{H}' = \Gamma^{-1} \boldsymbol{F}' \tag{2.23}$$

Substituting Equation 2.23 into Equation 2.22, I obtain the Fisher Information for the exponential family distribution [42]

$$J = \boldsymbol{F}'^{\top} \Gamma^{-1} \boldsymbol{F}'. \tag{2.24}$$

2.5.2 Estimation in the exponential family

Assume that the likelihood is exponential family distribution with nonlinear sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$(Eq 2.11). I want to compute the maximum likelihood estimate, $\hat{s}$, near a reference s_0.

$$\hat{s} = \underset{s}{\mathrm{argmax}}\, p(\boldsymbol{r}|s) \tag{2.25}$$

$$= \underset{s}{\mathrm{argmax}}\ \log p(\boldsymbol{r}|s) \tag{2.26}$$

$$= \underset{s}{\mathrm{argmax}}\, \boldsymbol{H}(s)^{\top} \boldsymbol{R}(\boldsymbol{r}) - A(s) \tag{2.27}$$

I use a Taylor expansion around a reference stimulus s_0

$$\begin{aligned} &\boldsymbol{H}(s)^{\top} \boldsymbol{R}(\boldsymbol{r}) - A(s) \\ &\approx [\boldsymbol{H}^{\top} \boldsymbol{R} - A] \\ &+ [\boldsymbol{H}'^{\top} \boldsymbol{R} - A'](s - s_0) \\ &+ \tfrac{1}{2}(s - s_0)^{\top} [\boldsymbol{H}''^{\top} \boldsymbol{R} - A''](s - s_0) + \cdots \end{aligned} \tag{2.28}$$

where all functions and derivatives are evaluated at s_0, and then find the maximum by differentiating with respect to s and setting to zero:

$$0 = [\boldsymbol{H}'^{\top}\boldsymbol{R} - A'] + (s - s_0)[\boldsymbol{H}''^{\top}\boldsymbol{R} - A''] \tag{2.29}$$

The solution is

$$s = s_0 - \frac{\boldsymbol{H}'^{\top}\boldsymbol{R} - A'}{\boldsymbol{H}''^{\top}\boldsymbol{R} - A''} \tag{2.30}$$

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

$$s = \boldsymbol{w}^{\top}\boldsymbol{R} + \boldsymbol{c} \tag{2.31}$$

where

$$\boldsymbol{w} = \frac{\boldsymbol{H}'}{J} = \frac{\Gamma^{-1}\boldsymbol{F}'}{\boldsymbol{F}'^{\top}\Gamma^{-1}\boldsymbol{F}'} \tag{2.32}$$

where I used the results from Equations 2.6 and 2.23, with $\Gamma = \mathrm{Cov}(\boldsymbol{R}|s_0)$ and $\boldsymbol{F} = \langle \boldsymbol{R}|s_0\rangle$. Thus, in this limit, the optimal estimator for s is a linear decoding of the sufficient statistics $\boldsymbol{R}(\boldsymbol{r})$.

2.5.3 Orientation estimation task with varying spatial phase

In Figure 2.2B, the subject's task is to estimate orientation s near a reference s_0, based on images G of Gabor patterns given by

$$G(\boldsymbol{x}|s, n) = e^{-\|\boldsymbol{x}\|^2}\cos\left(\boldsymbol{k}\cdot\boldsymbol{x} + n\right) \tag{2.33}$$

where $\boldsymbol{k} = \kappa(\cos s, \sin s)$. Here the target s is the orientation of the pattern, $\boldsymbol{n}$ is a nuisance variable reflecting the spatial phase, $\boldsymbol{x}$ is the pixel location in the image,

and $\boldsymbol{k}$ is a spatial frequency vector with amplitude $\kappa = \|\boldsymbol{k}\|$. I assume the spatial receptive field of simple cell j in primary visual cortex is also described by a Gabor function

$$\begin{aligned} \mathrm{RF}_j(\boldsymbol{x}, s_j, n_j) &= e^{-\|\boldsymbol{x}\|^2} \cos\left(\boldsymbol{k}_j \cdot \boldsymbol{x} + n_j\right) \\ \boldsymbol{k}_j &= \kappa(\cos s_j, \sin s_j) \end{aligned} \tag{2.34}$$

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

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

$$\begin{aligned} f_j(s, n) &= \int d\boldsymbol{x}\, G(\boldsymbol{x}|s, n) \mathrm{RF}_j(\boldsymbol{x}, s_j, n_j) \\ &= \Big[e^{-\frac{1}{4}\kappa^2 \cos(s - s_j)} \cos\left(n + n_j\right) \\ &\quad + e^{+\frac{1}{4}\kappa^2 \cos(s - s_j)} \cos\left(n - n_j\right) \Big] \frac{\pi}{4} e^{-\frac{1}{4}\kappa^2} \end{aligned} \tag{2.35}$$

This expression can be written in the form:

$$f_j(s, n) = A_j(s) \cos\left(n + \psi_j(s)\right) \tag{2.36}$$

using the stimulus-dependent response amplitude

$$A_j(s) = C\sqrt{2\cosh 2\beta_j(s) + 2\cos 2n_j} \tag{2.37}$$

and phase

$$\psi_j(s) = n_j - \alpha_j(s) \tag{2.38}$$

where I define the constants

$$C = \frac{\pi}{4} \exp\left(-\frac{1}{4}\kappa^2\right) \tag{2.39}$$

$$\beta_j(s) = \frac{1}{4}\kappa^2 \cos(s - s_j) \tag{2.40}$$

$$\alpha_j(s) = \tan^{-1} \frac{\exp\left(\beta_j(s)\right) \sin 2n_j}{\exp\left(-\beta_j(s)\right) + \exp\left(\beta_j(s)\right) \cos 2n_j} \tag{2.41}$$

Equation 2.36 reveals that the mean response of each neuron traces out a sinusoidal oscillation in n, where the amplitude and phase depend on s and the specific neuron j. The mean tuning for each pair of neurons therefore traces out an ellipse as a function of the nuisance variable, the input's spatial phase. When I *average* over the ellipse generated by the nuisance variable n, the mean tuning to s is abolished — but the response *covariances* (nuisance correlations) remain tuned to s.

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

$$P(\boldsymbol{r}|n, s) = \mathcal{N}(\boldsymbol{f}(s, n), \boldsymbol{I}) \tag{2.42}$$

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

$$f_j(s) = \langle r_j|s\rangle = \int r_j \, p(r_j|s) dr_j \tag{2.43}$$

$$= \iint r_j \, p(r_j|s, n) \, p(n) \, dr_j \, dn \tag{2.44}$$

$$= \int f_j(s, n) p(n) \, dn \tag{2.45}$$

$$= \frac{1}{2\pi} \int f_j(s, n) \, dn \tag{2.46}$$

$$= \frac{A_j(s)}{2\pi} \int_0^{2\pi} \cos\left(n + \psi_j(s)\right) dn = 0 \tag{2.47}$$

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

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

$$\begin{aligned}
\mathrm{Cov}_{ij}[\boldsymbol{r}|s] &= \langle (r_i - f_i(s))(r_j - f_j(s))|s\rangle && (2.48)\\
&= \langle r_i r_j|s\rangle = \iint r_i r_j\, p(\boldsymbol{r}|s) dr_i dr_j && (2.49)\\
&= \int dn \iint r_i r_j\, p(\boldsymbol{r}|s,n) p(n) dr_i dr_j && (2.50)\\
&= \int dn\, p(n) \langle r_i r_j|s,n\rangle && (2.51)\\
&= \int dn\, p(n) \left(\mathrm{Cov}_{ij}[\boldsymbol{r}|s,n] + f_i(s,n) f_j(s,n)\right) && (2.52)\\
&= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi}\int dn f_i(s,n) f_j(s,n) && (2.53)\\
&= \frac{1}{2\pi}\delta_{ij} + \frac{1}{2\pi} D_{ij}(s) && (2.54)
\end{aligned}$$

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

$$\begin{aligned}
D_{ij}(s) &= \int dn f_i(s,n) f_j(s,n)\\
&= \int dn\, A_i(s) \cos\left(n + \psi_i(s)\right) A_j(s) \cos\left(n + \psi_j(s)\right) && (2.55)\\
&= \pi \cos(\psi_i(s) - \psi_j(s)) A_i(s) A_j(s)
\end{aligned}$$

Here when I compute Equation 2.55, I used the trigonometric identity: $2\cos(x)\cos(y) = \cos(x+y) + \cos(x-y)$, and $\int \cos(2n + \psi_i + \psi_j) dn = 0$.

This demonstrates that the neural covariance $\mathrm{Cov}_{ij}[\boldsymbol{r}|s]$ depends on the orientation s. While linear computation is useless for estimating orientation since the mean responses are untuned (2.43), quadratic (or higher-order) nonlinear computations can be used to estimate the orientation.

2.5.4 Quadratic coding model

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

As a concrete example I construct a covariance that rotates with stimulus s (Figure 2.3A). Any covariance matrix needs to be positive semidefinite. I build $\Sigma(s)$ by setting the eigenvalues to be positive and s-independent and eigenvectors to form an orthogonal basis that rotates with s:

$$\Sigma(s) = V(s)\Lambda V(s)^\top \tag{2.56}$$

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.

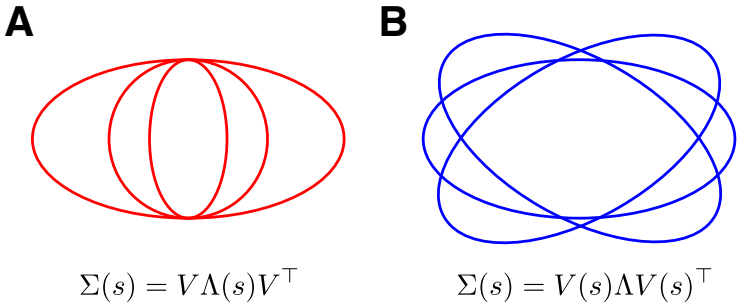


Figure 2.3 : Examples for stimulus-dependent covariance. **A**: Covariance that scales with stimulus s. **B**: Covariance that rotates with stimulus s.

To calculate the Fisher Information (Equation 2.6), I need to first calculate the derivative of the mean $\boldsymbol{F}' = \frac{\partial}{\partial s}\langle \boldsymbol{R}(\boldsymbol{r})|s\rangle$ and covariance $\Gamma = \mathrm{Cov}[\boldsymbol{R}(\boldsymbol{r})|s]$ of the quadratic sufficient statistics.

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

$$\Sigma'(s) = U e^{\Omega s}(\Omega X - X\Omega)e^{-\Omega s}U^\top \tag{2.57}$$

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

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

$$\begin{aligned} \langle r_i r_j r_k r_n | s \rangle = &\langle r_i r_j | s \rangle \langle r_k r_n | s \rangle + \langle r_j r_n | s \rangle \langle r_i r_n | s \rangle \\ &+ \langle r_i r_n | s \rangle \langle r_j r_k | s \rangle \end{aligned} \tag{2.58}$$

where

$$\langle r_i r_j | s \rangle = \Sigma_{ij} + f_i f_j \tag{2.59}$$

Substitution of the response covariance (Equation 2.56) into Equation 2.58 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.

2.5.5 Cubic codes

In Figure 2.4 I assume the brain encodes the stimulus using a cubic code. A simple cubic code in $\boldsymbol{z} = (z_i, z_j, z_k) \in \mathbb{R}^3$ can be written as

$$p(\boldsymbol{z}|s) = \tfrac{1}{Z} \exp\left(\gamma(s) z_i z_j z_k - |\boldsymbol{z}|^4\right) \tag{2.60}$$

where I include the base measure $e^{-|\boldsymbol{z}|^4}$ to ensure normalizability (Figure 2.4A).

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

$$p(\boldsymbol{z}|s) \approx \sum_{a=1}^{4} p(a)p(\boldsymbol{z}|a) \tag{2.61}$$

$$= \sum_{a} \frac{1}{4}\mathcal{N}(\boldsymbol{z}|\mu_a, \Sigma_a) \tag{2.62}$$

where

$$\mu_a = \frac{s}{\sqrt{1+s^2}}\boldsymbol{v}_a \tag{2.63}$$

and

$$\Sigma_a = \frac{1}{(1+s^2)^2}(I + s^2\boldsymbol{v}_a\boldsymbol{v}_a^\top) \tag{2.64}$$

The vectors $\boldsymbol{v}_a$ reflect the four corners of the tetrahedron, $v_{a,i} = \pm 1$, to match the tetrahedral symmetry of the pure cubic code (Equation 2.60, Figure 2.4). To sample from this distribution, I randomly choose a component a and then sample from the gaussian $\mathcal{N}(\boldsymbol{z}|\mu_a, \Sigma_a)$ conditioned on that component.

This distribution has zero mean and identity covariance but a nontrivial skewness tensor, and qualitatively matches the corresponding distribution for the true exponential family distribution with cubic sufficient statistics (Figure 2.4).

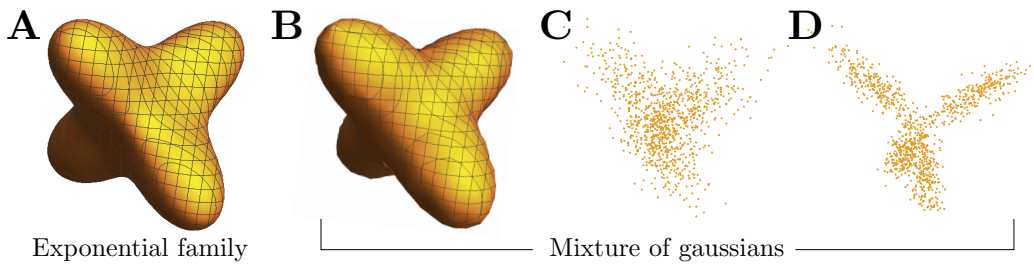


Figure 2.4 : Multivariate skewed distributions. (**A**) Isoprobability contour of an exponential family distribution with cubic statistics in three dimensions, drawn from $p(\boldsymbol{z}|s) \propto \exp\left(sz_1z_2z_3 - \|\boldsymbol{z}\|^4\right)$. (**B**) Isoprobability contour for a mixture of four gaussians (Eq 2.62). (**C,D**) Samples drawn from the mixture form, with $s = 1, 2$.

For simplicity, I consider pure cubic codes with non-overlapping cliques of three

variables.

$$p(\boldsymbol{z}|s) = \prod_{\alpha} p(\boldsymbol{z}_\alpha|s) = \prod_{\alpha} p(z_{\alpha_1}, z_{\alpha_2}, z_{\alpha_3}|s) \tag{2.65}$$

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

$$\boldsymbol{r} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\, \boldsymbol{z}\, \Sigma^{1/2}(s) \tag{2.66}$$

$$\boldsymbol{z} \sim \frac{1}{Z(s)} \exp\left[\sum_{ijk} \gamma_{ijk}(s) z_i z_j z_k - |\boldsymbol{z}|^4\right] \tag{2.67}$$

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

$$p(\boldsymbol{r}|a) = \mathcal{N}(\boldsymbol{r}|\boldsymbol{f}(s) + \boldsymbol{m}_a(s), \Sigma^{1/2}(s) S_a(s) \Sigma^{1/2}(s) \tag{2.68}$$

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

2.5.6 Information-limiting correlations

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

$$\begin{aligned} \Gamma =& \mathrm{Cov}_{\boldsymbol{r},ds}(\boldsymbol{R}(\boldsymbol{r})|s) \\ =& \left\langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s, ds) \right\rangle_{ds} + \mathrm{Cov}_{ds} \left\langle \boldsymbol{R}(\boldsymbol{r})|s, ds \right\rangle_{\boldsymbol{r}} \end{aligned} \tag{2.69}$$

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

$$\langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s, ds)\rangle_{ds} = \langle \Gamma(s+ds)\rangle_{ds} \tag{2.70}$$

$$\approx \langle \Gamma_0 + ds\,\Gamma'\rangle_{ds} \tag{2.71}$$

$$= \Gamma_0 \tag{2.72}$$

Here I denote the covariance of $\boldsymbol{R}(\boldsymbol{r})$ given s and ds as $\Gamma(s+ds)$. The second equality used a Taylor expansion of $\Gamma(s+ds)$ around s. The third equality used the fact that the mean of ds is zero. Γ_0 is the covariance of $\boldsymbol{R}$ in the absence of information-limiting correlations. The second term in Equation 2.69 can be expressed as

$$\mathrm{Cov}_{ds}\,\langle \boldsymbol{R}(\boldsymbol{r})|s, ds\rangle_{\boldsymbol{r}} \tag{2.73}$$

$$= \mathrm{Cov}_{ds}(\boldsymbol{F}(s+ds)) \tag{2.74}$$

$$\approx \mathrm{Cov}_{ds}(\boldsymbol{F}(s) + ds\,\boldsymbol{F}'(s)) \tag{2.75}$$

$$= \frac{1}{J_\infty}\boldsymbol{F}'(s)\boldsymbol{F}'(s)^\top \tag{2.76}$$

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

Equation 2.69 can therefore be written as

$$\Gamma = \Gamma_0 + \frac{1}{J_\infty}\boldsymbol{F}(s)'\boldsymbol{F}(s)'^\top \tag{2.77}$$

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

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

$$\Gamma^{-1} = \Gamma_0^{-1} - \frac{\Gamma_0^{-1}\boldsymbol{F}'\boldsymbol{F}'^\top\Gamma_0^{-1}}{J_\infty + \boldsymbol{F}'\Gamma_0^{-1}\boldsymbol{F}'^\top} \tag{2.78}$$

Substituting these equations into the nonlinear Fisher Information (Equation 2.6) and simplifying, I obtain

$$J_{R(\boldsymbol{r})} = \frac{1}{1/J_\infty + 1/J_0} \tag{2.79}$$

Here $J_0 = \boldsymbol{F'}^\top \Gamma_0^{-1} \boldsymbol{F'}$ is the nonlinear Fisher Information in the absence of information-limiting correlations. When the population size grows, the term J_0 grows proportionally [5, 6], so for large populations the output information saturates at J_∞.

Chapter 3

Nonlinear choice correlation

高尚是高尚者的墓志铭，卑鄙是卑鄙者的通行证。

—《回答》北岛

Debasement is the password of the base, nobility the epitaph of the noble.

— 'The answer', Bei Dao

3.1 Introduction

As shown in the quotation, a choice made during hard times reveals the spirit and the morality of a person. Likewise, the behavioral tasks — tasks requiring the subjects to behave properly when they are shown a specific stimulus — are widely used in neuroscience studies, for the purpose of unraveling the brain's computational strategy. In this chapter, we examine how relationships between patterns of neural activity and patterns of choices can help us understand neural computation.

Just because a neural population encodes information about the stimulus, it does not mean that the brain decodes it all. 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 [10–18]. However, this choice-related activity is hard to interpret, because it may reflect decoding of the recorded neurons, or merely correlations between them and other neurons that are decoded instead [20].

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

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

In this chapter, I will start by reviewing the previous work [19] in the case of linear neural codes, which evaluates the brain's decoding efficiency from the testable predictions about choice-related activity. Then I will generalize it in the context of nonlinear neural codes–nonlinear choice correlation test. Then I show the optimality relationship predicted by this test holds even if the specific nonlinearities used by the brain differ from the exact nonlinearity defined by the neural encoding, as long as the decoder optimally uses all information. Even if the decoding is not strictly optimal, the information redundancy in the neural population can makes the test less stringent. To validate this theory, I show that when primates discriminate between a wide or narrow distribution from which oriented images could be sampled, quadratic statistics of primary visual cortex activity match the prediction from the nonlinear choice correlation test.

3.2 Results

3.2.1 Choice correlations predicted for optimal linear decoding

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

$$C_{r_k} = \mathrm{Corr}(r_k, \hat{s}|s) \tag{3.1}$$

Here we compute the choice correlation conditioned on the stimulus. This is because we want to find the choice fluctuations purely driven by the neurons, instead of the co-fluctuations driven by the stimulus. The latter tells little about the statistical interactions between the response and the stimulus estimate. One can always find significant correlation between the response and the stimulus estimate without conditioning on the stimulus.

This choice correlation is a conceptually simpler and more convenient measure than the more conventional statistic, 'choice probability' [10], but it has almost identical properties (Section 3.4) [19, 20].

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 [19] (Derivation in Section 3.4 and 3.4.2 where substituting $\boldsymbol{R} = \boldsymbol{r}$):

$$C_{r_k}^{\mathrm{opt}} = \sqrt{\frac{J_{r_k}}{J}} \tag{3.2}$$

where J and J_{r_k} are, respectively, the Fisher Information based on the entire population $\boldsymbol{r}$ or on neuron k's response r_k (Section 3.4). This relationship holds for a locally

optimal linear estimator,

$$\hat{s} = \boldsymbol{w} \cdot \boldsymbol{r} + c \tag{3.3}$$

regardless of the structure of noise correlations.

Another way to test for optimal linear decoding would be to measure whether the animal's behavioral discriminability matches the discriminability for an ideal observer of the neural population response. Yet this approach is not feasible, as it requires one to measure simultaneous responses of many, or even all, relevant neurons. In contrast, the optimality test (Eq 3.2) requires measuring only single neuron responses, which is vastly easier. Neural recordings in the vestibular system are consistent with optimal decoding according to this prediction [19].

3.2.2 Nonlinear choice correlations for optimal decoding

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

This situation is exactly the same one that gives rise to nonlinear codes. A natural generalization of Equation 3.2 can reveal the quality of neural computation on nonlinear codes. I simply define a '*nonlinear* choice correlation' between the stimulus estimate $\hat{s}$ and nonlinear functions of neural activity $\boldsymbol{R}(\boldsymbol{r})$:

$$C_R = \mathrm{Corr}(R(\boldsymbol{r}), \hat{s}|s) \tag{3.4}$$

(Section 3.4), where $R(\boldsymbol{r})$ is a nonlinear function of the neural responses. If the brain optimally decodes the information encoded in the nonlinear statistics of neural activity, according to the simple nonlinear extension to Eq 3.3,

$$\hat{s} = \boldsymbol{w} \cdot \boldsymbol{R}(\boldsymbol{r}) + c \tag{3.5}$$

then the nonlinear choice correlation satisfies the equation

$$C^{\text{opt}}_{R(\boldsymbol{r})} = \sqrt{\frac{J_{R(\boldsymbol{r})}}{J}} \tag{3.6}$$

where $J_{R(\boldsymbol{r})}$ is the Fisher Information in $R(\boldsymbol{r})$ (Section 3.4.2).

As an example of this relationship, I return to the orientation example. Here the response covariance $\Sigma(s) = \text{Cov}(\boldsymbol{r}|s)$ depends on the stimulus, but the mean $\boldsymbol{f} = \langle \boldsymbol{r}|s \rangle = \langle \boldsymbol{r} \rangle$ does not. In this model, optimally decoded neurons would have no linear correlation with behavioral choice. Instead, the choice should be driven by the product of the neural responses, $\boldsymbol{R}(\boldsymbol{r}) = \text{vec}(\boldsymbol{r}\boldsymbol{r}^\top)$, where $\text{vec}(\cdot)$ is a vectorization that flattens an array into a one-dimensional list of numbers. Figure 3.1 shows linear and nonlinear choice correlations for pairs of neurons, defined as $C_{r_i r_j} = \text{Corr}(r_i r_j, \hat{s}|s)$. When decoding is linear, linear choice correlations are strong while nonlinear choice correlations are near zero (Figure 3.1A,B). When the decoding is quadratic, here mediated by an intermediate layer that multiplies pairs of neural activity, the nonlinear choice correlations are strong while the linear ones are insignificant (Figure 3.1C,D).

3.2.3 Which nonlinearity?

This optimality relationship holds even if the specific nonlinearities used by the brain differ from those selected for testing in Eq 3.6 (Section 3.4.3), as long as the decoder optimally uses all information. This is equivalent to expressing the same nonlinearities in a different basis (Section 3.4.3). Figure 3.2 shows a situation where information is encoded by quadratic and cubic sufficient statistics of neural responses, while a simulated brain decodes them near-optimally using a generic neural network rather than a set of nonlinearities matched to the sufficient statistics. Despite this mismatch I can successfully identify that the brain is near-optimal by applying Eq 3.6, even without knowing the simulated brain's true nonlinear transformations.

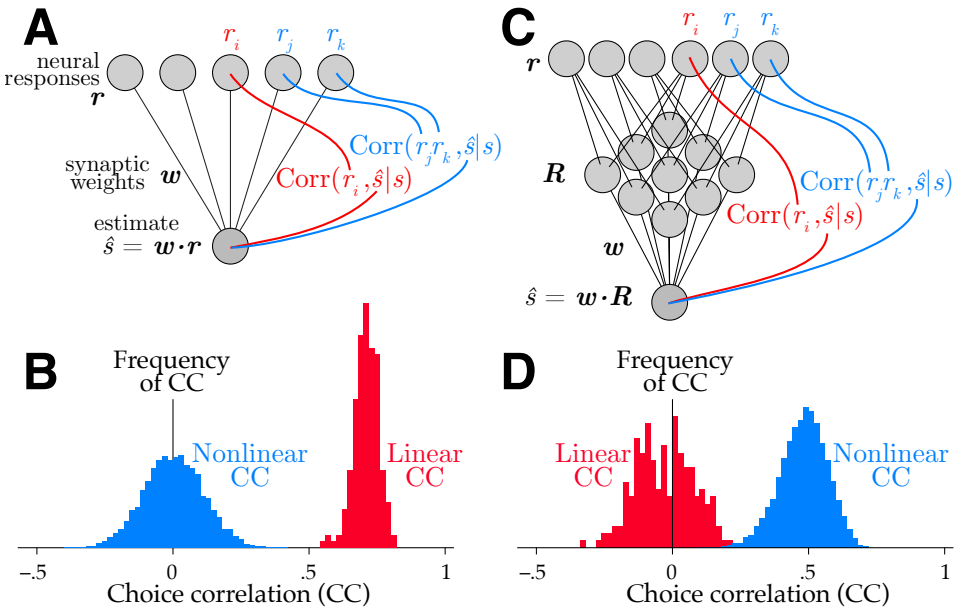


Figure 3.1 : Linear and nonlinear choice correlations successfully distinguish network structure. A linearly decoded population (**A**) produces nonzero linear choice correlations (**B**), while the nonlinear choice correlations are randomly distributed around zero. The situation is reversed for a nonlinear 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, $C_{jk} = \mathrm{Corr}(r_j r_k, c)$.

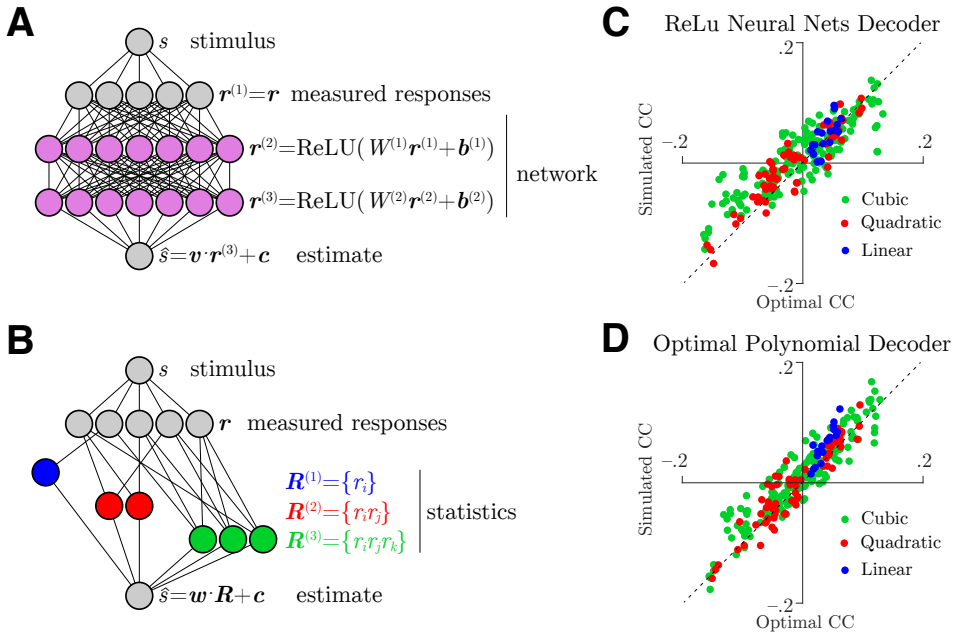


Figure 3.2 : Identifying optimal nonlinear decoding by a generic neural network using nonlinear choice correlation. **A**: Responses encode stimulus information in polynomial sufficient statistics up to cubic, simulated brain using ReLu nonlinearities trained to extract that information **B**: Simulated brain using matched polynomial nonlinearities to extract the information from responses. **C,D**: Choice correlations with polynomial nonlinear statistics show that the network computations are consistent with optimal nonlinear decoding, regardless if the tested statistics match(Network **B**) or not match(Network **A**) the actual decoder(Section 3.4.3).

3.2.4 Decoding efficiency revealed by choice correlations

Even if decoding is not strictly optimal, Eq. 3.6 can be satisfied due to information-limiting correlations. Decoders that seem substantially suboptimal because they fail to avoid the largest noise components in Γ_0 can be nonetheless dominated by the bound from information-limiting correlations. This will occur whenever the variability from suboptimally decoding Γ_0 is smaller than ϵ. Just as I can decompose the nonlinear noise correlations into information-limiting and other parts, I can decompose nonlinear choice correlations into corresponding parts as well, with the result that

$$C_R^{\text{sub}} \approx \alpha C_R^{\text{opt}} + \zeta_R \tag{3.7}$$

where ζ_R depends on the particular type of suboptimal decoding (Section 3.5.2). 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 therefore provides an estimate of the efficiency of the brain's decoding.

Figure 3.3 shows an example of a decoder that would be highly suboptimal without considering redundancy, but is nonetheless close to optimal when information limits are inherited.

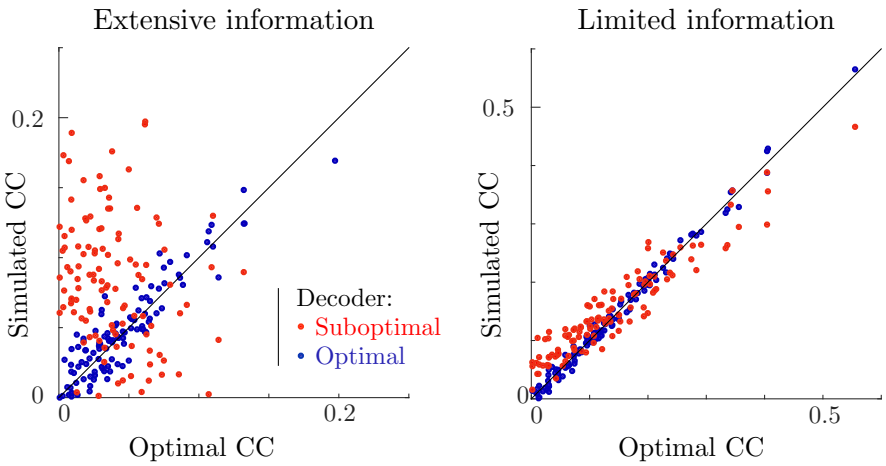


Figure 3.3 : Information-limiting noise makes a network more robust to suboptimal decoding. (Left) A simulated optimal decoder produces choice correlations that match my optimal predictions (blue, on diagonal). In contrast, a suboptimal decoder, such as one that is blind to higher-order correlations ($\boldsymbol{w} \propto \boldsymbol{F}'$), exhibits a suboptimal pattern of choice correlations (red, off-diagonal) in the presence of noise Γ_0 that permits the population to have extensive information. (Right) When information is limited, the same decoding weights are less detrimental, and thus exhibit a similar pattern of choice correlations as an optimal decoder.

In realistically redundant models that have more cortical neurons than sensory neurons, many decoders could be near-optimal [19]. However, even in redundant codes there may be substantial inefficiencies, especially for unnatural tasks [60].

3.2.5 Application to neural data

We applied our optimality test to data recorded from primate visual cortex (V1) during a nonlinear decoding task and found supportive results. Monkeys faced a Two-Alternative Forced Choice task (2AFC) in which they categorized an oriented grating based on whether it came from a wide or narrow distribution of orientations (Section 3.4.4.1). While some orientations could arise from either distribution, they were not equally likely.

For this task, the target variable s is the variance of the distribution. If neural responses can be linearly decoded to compute orientation, then the locally optimal decoder for the variance is quadratic in those responses. Indeed there is substantial information about the variance in both linear and quadratic statistics of V1 responses (Figure 3.4A), suggesting that these neural responses have already performed some useful nonlinear transformations of the input from their receptive field, as expected from the energy model of complex cells [22]. Furthermore, we found nonlinear choice correlations that were highly correlated with the corresponding information content of each statistic (Figure 3.4A), as predicted by Equation 3.6. The slope of this relationship was 0.96 ± 0.01, near the value of 1 predicted for optimal decoding.

To calculate choice correlations associated solely with internal noise, we would have to eliminate all nuisance variables that modulate both neuronal responses and behavior. In this experiment, however, the stimulus is the variance of the nuisance, so the nuisance variation cannot be frozen for many trials without disturbing the task. Nonetheless, we can distinguish whether nonlinear choice correlation arises from nuisance variation or internal noise variation by performing two control analyses.

First we break the correlations between the internal noise and fluctuations in choice, by shuffling the subjects' choices randomly between trials with matched s and n, and then perform the same analysis on these shuffled data. Since the orientations were not generally identical across trials, we shuffle choices for trials with similar orientations (within 1°). Figure 3.4B shows this shuffle control does not eliminate the choice correlations, demonstrating that the internal noise was not responsible.

In a second control, we break the correlations between choices and the external nuisance variable. To do so, we shuffled the subjects' choices while fixing only the distribution width, *i.e.* matching a specific stimulus, while allowing the nuisance

variable to take any value from the stimulus-conditioned distribution $p(n|s)$. Figure 3.4C shows that this control completely eliminates the nonlinear choice correlation.

The same pattern of choice correlations also was seen in simulated data for this task (Figure 3.4D–F, Methods 3.4.4.1).

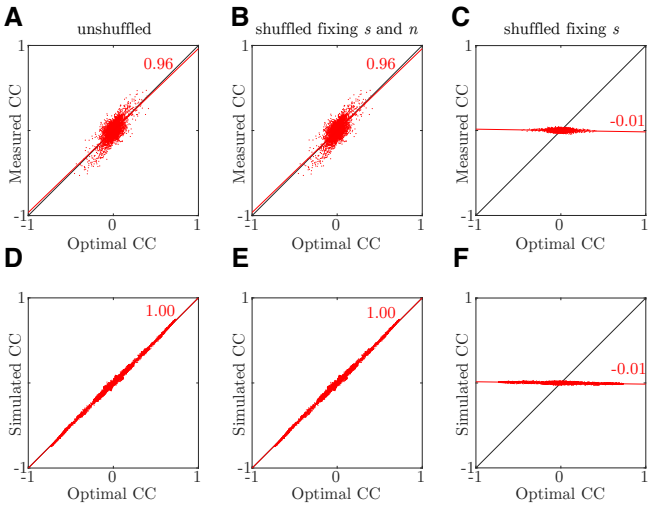


Figure 3.4 : Nonlinear choice correlations in a variance discrimination task, for both real and simulated neural data. **A**: There is significant nonlinear information (horizontal axis), and the nonlinear choice correlation (vertical axis) is strongly correlated with the predictions of Equation 3.6. **B**: Shuffling choices while matching both the stimulus and nuisance variables s and n preserves the correlation between prediction and nonlinear choice, implying that internal noise correlations are not responsible. **C**: Shuffling choices and nuisance values while matching the stimulus eliminates the correlation, implying that nuisance correlations create the nonlinear code. **D**,**E**,**F**: As for A–C, but with simulated data (Methods 3.4.4.1).

Combining the conclusion from the shuffle controls, we find no evidence in this task that the brain decodes any stimulus-dependent internal noise correlations. Instead the nonlinear information decoded by the brain is generated by nuisance variation.

3.3 Discussion

In this chapter, I provide a remarkably simple test to determine if downstream nonlinear computation decodes all that is encoded. This test showed how correlated fluctuations between neural activity and behavioral choices could reveal properties of the brain's decoding.

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

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

3.3.1 Which nonlinearities should I test?

If all neural signals are decoded optimally, then all choice correlations for any function of those signals should also be consistent with optimal decoding. Yet for the wrong or incomplete nonlinearities that do not fully disentangle the task-relevant variables from the nuisance variables, the test may be inconclusive, just as it was for linear decoding of a nonlinear code (Figure 3.2): the chosen nonlinear functions may extract little

linearly decodable information and have correspondingly little choice correlation.

The optimal nonlinearities would be those that collectively extract the sufficient statistics about the relevant stimulus, which will depend on both the task and the nuisance variables. In complex tasks, like recognizing object from images with many nuisance variables, most of the relevant information lives in higher-order statistics, and therefore require more complex nonlinearities to extract. In such high-dimensional cases, my proposed test is unlikely to be useful. This is because my method expresses stimulus estimates as sums of nonlinear functions, and while that is universal in principle [61], that is not a compact way to express the complex nonlinearities of deep networks. Alternatively, with good guidance from trained neural network models my method could potentially judge whether those nonlinearities provide a good description of neural decoding. This decoding perspective would complementing studies that argue for a good match between encoding by convolutional neural networks [26].

The best condition to apply my optimality test is in tasks of modest complexity but still possessing fundamentally nonlinear structure. Some interesting examples where my test could have practical relevance include motion detection using photoreceptors [62], visual search with distractors (XOR-type tasks) [47, 63], sound localization in early auditory processing before the inferior colliculus [64], or context switching in higher-level cortex [58].

my test for optimal nonlinear decoding really amounts to testing for optimal linear decoding of nonlinear functions of recorded neural data. If we had access to some putative downstream neurons that computed these nonlinear functions, we could just test whether the brain linearly decoded those neurons optimally. Yet that would circumvent the most interesting and crucial nonlinear aspects of neural computation. Alternatively, if we could record from neurons at different levels of the processing

chain, we could try to characterize that nonlinear recoding between them directly, without reference to a behavioral choice. But this would not easily relate these computations to their functional role. The method proposed here allows us to skip these intermediate steps and directly test the optimality of all accumulated downstream nonlinearities.

3.3.2 Limitations of the approach

For efficient decoding in a learned task, the optimality test (3.6) is necessary but not sufficient. If the brain neglects some of the informative sufficient statistics, and we don't test these neglected statistics either, then we could find that the brain is consistent with my optimal decoding test, yet still be suboptimal. Only if the test is passed for *all* statistics will the test be conclusive. For an extreme example, a single neuron might pass the test, but if other neurons don't, then the brain is not using its information well. On a broader scale, one might find that all individual responses r_k pass the optimality test, while products of responses $r_j r_k$ fail. This would be consistent with linear information being used well while distinct quadratic information is present but unused; on the other hand this outcome would not be consistent with quadratic statistics that are uninformative but decoded anyway, since that would increase the output variance beyond that expected from the linear information. In future work I will demonstrate how I can use nonlinear choice correlations to identify properties of suboptimal decoders [8].

My approach is currently limited to feedforward processing, which unquestionably oversimplifies cortical processing. Nonetheless, feedforward models do a fair job of capturing the representational structure of the brain [26].

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 [65], then this could appear as information-limiting correlations, enhancing the match with Eq 3.6. This situation could be disambiguated by measuring the internal noise source providing the feedback, and of course this would require more simultaneous measurements.

3.3.3 Comparing choice correlations from internal and external noise

Since many stimulus-dependent response correlations are induced by external nuisance variation, not internal noise, we might not find informative stimulus-dependent noise correlations upon repeated presentations of a fixed stimulus. Those correlations may only be informative about a stimulus in the presence of natural nuisance variation. For example, if a picture of a face is shown repeatedly without changing its pose, then small expression changes can readily be identified by linear operations; if the pose can vary then the stimulus is only reflected in higher-order correlations [25].

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. This influence will manifest as nonlinear choice correlations. In other words, stimulus-dependent noise correlations need not predict a fixed stimulus, but they may predict the choice.

For optimal decoding, the choice correlations measured using fixed nuisance variables will differ from Eq 3.6, which should strictly hold only when there is natural nuisance variation. This is implicit in Eq 3.6, since the relevant quantities are conditioned only on the relevant stimulus s while averaging over the nuisance variations $\boldsymbol{n}$.

However, under some conditions, a related prediction for nonlinear choice correlations holds even without averaging over nuisance variables (Section 3.5.4).

3.4 Methods

3.4.1 Estimating choice correlation

The nonlinear choice correlation between the stimulus estimate $\hat{s} = \boldsymbol{w}^T \boldsymbol{R} + c$ and a nonlinear function R_k of recorded neural activity $\boldsymbol{r}$ is

$$C_{R_k} = \mathrm{Corr}(R_k(\boldsymbol{r}), \hat{s}|s) = \frac{(\Gamma \boldsymbol{w})_k}{\sqrt{\Gamma_{kk} \boldsymbol{w}^\top \Gamma \boldsymbol{w}}} \tag{3.8}$$

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

To compute this quantity from neural responses to stimuli, I need to condition neural responses and behavior data on the same signal s, or on the same total input $(s, \boldsymbol{n})$ if I want to isolate the contribution of purely internal noise rather than nuisance variation (Section 3.5.4). I combine choice correlations calculated under different stimulus conditions by balanced z-scoring [66].

3.4.2 Optimality test

Locally optimal linear estimator weights for decoding statistics $\boldsymbol{R}$ are given by linear regression as $\boldsymbol{w} \propto \Gamma^{-1} \boldsymbol{F}'$. Substituting these weights into (3.8) $\hat{s} = \boldsymbol{w}^T \boldsymbol{R}(\boldsymbol{r}) + c$, the optimal nonlinear choice correlation becomes

$$C^{\mathrm{opt}}_{R_k(\boldsymbol{r})} = \frac{(\Gamma \Gamma^{-1} \boldsymbol{F}')_k}{\sqrt{\Gamma_{kk} \boldsymbol{F}'^\top \Gamma^{-1} \boldsymbol{F}'}} = \frac{F'_k}{\sqrt{\Gamma_{kk}}} \sigma_{\hat{s}} = \sqrt{\frac{J_{R_k(\boldsymbol{r})}}{J}} \tag{3.9}$$

where $J_{R_k(\boldsymbol{r})} = F'^2_k / \Gamma_{kk}$ is the linear Fisher Information in $R_k(\boldsymbol{r})$.

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

equivalent ways:

$$C_{R_k}^{opt} = \frac{d'_{R_k}}{d'} = \frac{\theta}{\theta_{R_k}} = \sqrt{\frac{\sigma_{\hat{s}}^2}{\sigma_{\hat{s},R_k}^2}} = \sqrt{\frac{J_{R_k}}{J}} \tag{3.10}$$

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

In my simulations with binary choices for fine discrimination, I calculate the optimal nonlinear choice correlation using d-prime [67]. d'_{R_k} is estimated from neural responses generated by stimuli $s_\pm = s_0 \pm \Delta s/2$ near a reference stimulus s_0:

$$d'_{R_k} = \frac{\Delta F_k}{\sigma_{R_k}} = \frac{F_k(s_+) - F_k(s_-)}{\sqrt{\frac{1}{2}\left(\sigma_{R_k|s_+}^2 + \sigma_{R_k|s_-}^2\right)}} \tag{3.11}$$

The discriminability for a decoded neural population is estimated from the unbiased decoder output's standard deviation, $d' = 1/\sigma_{\hat{s}_{\text{ref}}}$. In Section 3.2.5, the experimental task involves *coarse* discrimination, which requires a slight adjustment to the predicted choice correlations (Section 3.5.3).

3.4.3 Nonlinear choice correlation to analyze an unknown nonlinearity

In Figure 3.2, I generated neural responses given sufficient statistics that are polynomials up to third order, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j, r_i r_j r_k\}$ (Section 2.4.4). In contrast, my model brain decodes the stimulus using a cascade of linear-nonlinear transformations, with Rectified Linear Units ($\text{ReLU}(x) = \max(0, x)$) for the nonlinear activation functions. I used a fully-connected ReLU network with two hidden layers and 30 units

per hidden layer,

$$\boldsymbol{r}^{(1)} = \boldsymbol{r} \tag{3.12}$$

$$\boldsymbol{r}^{(2)} = \mathrm{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{r}^{(1)} + \boldsymbol{b}^{(1)}) \tag{3.13}$$

$$\boldsymbol{r}^{(3)} = \mathrm{ReLU}(\boldsymbol{W}^{(2)}\boldsymbol{r}^{(2)} + \boldsymbol{b}^{(2)}) \tag{3.14}$$

$$\hat{s}_{\mathrm{brain}} = \boldsymbol{v} \cdot \boldsymbol{r}^{(3)} + \boldsymbol{b}^{(3)} \tag{3.15}$$

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

3.4.4 Application to neural data

3.4.4.1 Orientation variance task and experiment

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

After the monkey was fully trained, we implanted a 96-electrode microelectrode array (Utah array, Blackrock Microsystems, Salt Lake City, UT, USA) with a shaft length of 1 mm over parafoveal area V1 on the right hemisphere. The neural signals

were pre-amplified at the head stage by unity gain preamplifiers (HS-27, Neuralynx, Bozeman MT, USA). These signals were then digitized by 24-bit analog data acquisition cards with 30 dB onboard gain (PXI-4498, National Instruments, Austin, TX) and sampled at 32 kHz. The spike detection was performed offline according to a previously described method [28, 68]. Code for spike detection is available online at `github.com/atlab/spikedetection`. Multiunit neural responses r_k were measured by spike counts in the 500 ms preceding the saccade target onset.

3.4.4.2 Orientation variance analysis

The task-relevant stimulus s is the variance σ^2 of the distribution over orientations, and takes either a larger or smaller value ($s_\pm = \sigma^2_\pm$). The orientation itself then serves as a nuisance variable n, drawn from the distribution $p(n|s) = \mathcal{N}(n|0, s)$. The maximum log-likelihood s is therefore given by $\partial_s \left(-n^2/2s^2 - \frac{1}{2}\log s\right) = 0$, so the stimulus s can be estimated as a quadratic function of the orientation, $\hat{s} = n^2$. This means that 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.

To compute experimental choice correlations, we first resample experimental trials from both s_+ and s_- to create a perfectly ambiguous new stimulus condition, s_0, in which orientations are normally distributed with a variance equal to the decision boundary s_0 for the 2AFC task. The optimal decision boundary is given by the condition that $p(\hat{s} = s_0|s_+) = p(\hat{s} = s_0|s_-)$ (Supplementary Material **??**). We then calculate nonlinear choice correlations from this composite data set.

To predict the choice correlations for optimal decoding (Eq 3.6), we need to compute the Fisher Information at the decision boundary s_0, both for the whole popula-

tion and for the statistic $R_k(\boldsymbol{r})$ under consideration. For the population information, we assume that the intrinsic uncertainty of estimating a variance from orientation dominates the smaller uncertainty of estimating orientation from neural activity, so that $J \approx 1/2s^2$. The estimate of the Fisher information in the statistic $R_k(\boldsymbol{r})$ is more subtle for the coarse discrimination task than for the fine discriminations considered above (Section 3.4.2). Supplementary Material 3.5.3 shows how we use the mean orientation sensitivities $\boldsymbol{g}'$ and response covariance $\Sigma = \mathrm{Cov}(\boldsymbol{r}|s,n)$ to estimate the Fisher Information for $r_j r_k$:

$$J_{R_{jk}} = \frac{g_j'^2 g_k'^2}{(\Sigma_{jj} + s g_j'^2)(\Sigma_{kk} + s g_k'^2) - (\Sigma_{jk} + s g_j' g_k')^2} \tag{3.16}$$

which then gives us the prediction for the optimal nonlinear choice correlations.

To validate this analysis, we repeated it for simulated neural data created using random neural sensitivities g_i' that were normally distributed with zero mean and unit variance, and independent gaussian internal noise with variance 3. Simulated choices were generated by thresholding optimal orientation variance estimates (Equation 3.44).

3.5 Supplemental material

3.5.1 Using nonlinear choice correlation to analyze unknown nonlinearities

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

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

(Equation 2.31):

$$\hat{s}_{\text{opt}} = \boldsymbol{w} \cdot \boldsymbol{R}(\boldsymbol{r}). \tag{3.17}$$

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

$$\hat{s}_{\text{brain}} = \boldsymbol{v} \cdot \boldsymbol{g}(\boldsymbol{r}). \tag{3.18}$$

As long as the brain's nonlinear function spans the same function basis as the sufficient statistics, I can still get all of the information about stimulus from neural population. This allows us to use choice correlation between brain's estimate $\hat{s}_{\text{brain}}$ and my analysis nonlinearity $\Psi(\boldsymbol{r})$ to check the optimality condition (Equation 3.6).

In Figure 3.2, I assumed that the optimal nonlinear basis function $\boldsymbol{R}$ is polynomial nonlinearity up to third order, $\boldsymbol{R}(\boldsymbol{r}) = (r_i, r_i r_j, r_i r_j r_k, ...)$. I used cubic codes described in Section 2.4.4 to generate neural responses for which $\boldsymbol{R}(\boldsymbol{r})$ are sufficient statistics for the stimulus. In this simulation, 18 neuronal responses (six cliques of size 3) were generated using cubic codes.

My model brain decodes the stimulus using a cascade of linear-nonlinear transformations, with Rectified Linear Units ($\text{ReLU}(x) = \max(0, x)$) for the nonlinear activation functions. I used a fully-connected ReLU network with two hidden layers and 30 units per hidden layer,

$$\hat{s}_{\text{brain}} = \boldsymbol{v} \cdot \boldsymbol{r}^{(3)} + \boldsymbol{b}^{(3)} \tag{3.19}$$

$$\boldsymbol{r}^{(3)} = \text{ReLU}(\boldsymbol{W}^{(2)}\boldsymbol{r}^{(2)} + \boldsymbol{b}^{(2)}) \tag{3.20}$$

$$\boldsymbol{r}^{(2)} = \text{ReLU}(\boldsymbol{W}^{(1)}\boldsymbol{r}^{(1)} + \boldsymbol{b}^{(1)}) \tag{3.21}$$

$$\boldsymbol{r}^{(1)} = \boldsymbol{r} \tag{3.22}$$

I trained the neural network with 20000 response samples generated from a cubic code driven by stimuli near the reference s_0. I optimized the estimation performance

for the neural network using backpropagation to find weights $\{\boldsymbol{W}^{(\ell)}\}$, biases $\{\boldsymbol{b}^{(\ell)}\}$, and readout vector $\boldsymbol{v}$ that minimized the mean squared error. My 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 my simulated brain's true decoder, I used mononomial nonlinearities $\Psi(\boldsymbol{r})$ in the nonlinear choice correlation test (Equation 3.6). The simulated choice correlations were calculated by Equation 3.4, where $\boldsymbol{R}(\boldsymbol{r}) = \Psi(\boldsymbol{r})$ based on neural responses driven by the reference stimulus s_0, and the stimulus estimate was $\hat{s}_{\text{brain}}$. The optimal choice correlation is computed using Equation 3.6, where $\sqrt{J_{\Psi(\boldsymbol{r})}} = d'_\Psi/\Delta s = \frac{\Delta \boldsymbol{F}_\Psi}{\Delta s\, \sigma_\Psi}$, and $\sqrt{J} \approx 1/\sigma_{\hat{s}_{\text{brain}}}$. I computed $\Delta \boldsymbol{F}_\Psi$ based on neural population responses $\boldsymbol{r}_+$ and $\boldsymbol{r}_-$ driven by stimuli $s_+ = s_0 \pm \Delta s/2$. The change in mean was $\Delta \boldsymbol{F}_\Psi = \langle \Psi(\boldsymbol{r}_+) \rangle - \langle \Psi(\boldsymbol{r}_-) \rangle$, and the average standard deviation was $\sigma_\Psi = \sqrt{\frac{1}{2}\text{Var}(\Psi(\boldsymbol{r}_+)) + \frac{1}{2}\text{Var}(\Psi(\boldsymbol{r}_-))}$. $\sigma^2_{\hat{s}_{\text{brain}}}$ is the variance of estimate of reference stimulus s_0 using the trained neural network. Based on these quantities, Figure 3.2 shows that I can successfully identify that the brain is near-optimal.

3.5.2 Nonlinear choice correlation for suboptimal decoding

A decoder that would be suboptimal for one population code could be near-optimal in the presence of information-limiting noise. In this case, nonlinear choice correlations can be decomposed into a sum of two terms, one from the information-limiting component and the other from the rest of the noise [19]:

$$C_{R_k} = \frac{(\Gamma \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} = \frac{(\Gamma_0 \boldsymbol{w} + \frac{1}{J_\infty} \boldsymbol{F}' \boldsymbol{F}'^\top \boldsymbol{w})_k}{\sigma_k \sigma_{\hat{s}}} \tag{3.23}$$

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

$$C_{R_k} = \frac{(\Gamma_0 \boldsymbol{w})_k}{\Gamma_{0k} \sigma_{0\hat{s}}} \frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} \frac{\Gamma_{0k}}{\Gamma_k} + \frac{F'_k}{\sigma_k} \sigma_{\hat{s}} \frac{1/J_\infty}{\sigma^2_{\hat{s}}} \tag{3.24}$$

where $\Gamma_{0k} = (\Gamma_0)_{kk} \approx \Gamma_{kk}$ for small information-limiting noise variance $1/J_\infty \ll \Gamma_{0k}$ (which nonetheless can have a large effect on information despite the small variance), and where $\sigma_{0\hat{s}}$ is the standard deviation of the estimate produced by the same suboptimal decoder $\boldsymbol{w}$ in the absence of information-limiting correlations, *i.e.* when the covariance of the sufficient statistics is Γ_0. The variance of $\hat{s}$ can itself be decomposed into two terms as well:

$$\begin{aligned}\sigma_{\hat{s}}^2 &= \boldsymbol{w}^\top \Gamma \boldsymbol{w} = \boldsymbol{w}^\top \Gamma \boldsymbol{w} + \frac{1}{J_\infty} \boldsymbol{w}^\top \boldsymbol{F}' \boldsymbol{F}'^\top \boldsymbol{w} \\ &= \sigma_{0\hat{s}}^2 + 1/J_\infty\end{aligned} \tag{3.25}$$

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

$$\frac{\sigma_{0\hat{s}}}{\sigma_{\hat{s}}} = \sqrt{1 - \frac{1/J_\infty}{\sigma_{\hat{s}}^2}} = \sqrt{1-\alpha} \tag{3.26}$$

with $\alpha = \frac{1/J_\infty}{\sigma_{\hat{s}}^2}$. Substituting these into (Eq 3.24) I find that the choice correlation for a suboptimal decoder in the presence of information-limiting correlations is a weighted sum of the choice correlations for optimal and suboptimal decoding:

$$C_R^{\text{sub}} \approx \alpha C_R^{\text{opt}} + C_R^{\text{sub}} \sqrt{1-\alpha} \tag{3.27}$$

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.

3.5.3 Orientation variance task

Section 3.2.5 of the main text analyzes the quadratic decoding used to estimate whether an oriented visual stimulus was drawn from a distribution with high or low variance. Here I describe how I predict the choice correlations for optimal decoding when the stimulus is at the decision boundary.

First, I assume a neural model and then estimate the parameters in the model from the recorded neural data. Second, I compute the optimal decision boundary for a binary choice under this model. Third, I predict the choice correlations for continuous optimal estimation when the stimulus is at this decision boundary. Last, I relate the predicted choice correlation for the continuous estimate to the predicted choice correlation for the binary choice.

3.5.3.1 Neural model

I created a simulated brain to perform this orientation variance discrimination task, for comparison to the experimental data. For simplicity I model V1 neuronal responses as having a mean linearly tuned to orientation as $\langle \boldsymbol{r}|n\rangle = \boldsymbol{g}_0 + \boldsymbol{g}'n$, with additive Gaussian noise having covariance Σ: $p(\boldsymbol{r}|n,s) = p(\boldsymbol{r}|n) = \mathcal{N}(\boldsymbol{r}|n\boldsymbol{g}' + \boldsymbol{g}_0, \Sigma)$. I can then compute $p(\boldsymbol{r}|s)$ by marginalizing over n:

$$p(\boldsymbol{r}|s) = \int dn\, p(\boldsymbol{r}|n)\, p(n|s) \tag{3.28}$$

$$= \mathcal{N}(\boldsymbol{r}|\boldsymbol{g}_0, \Sigma + s\boldsymbol{g}'\boldsymbol{g}'^{\top}) \tag{3.29}$$

where the prior over orientations depends on the variance s as $p(n|s) = \mathcal{N}(n|0,s)$. In my variance discrimination experiment I measure the neural responses for stimuli $s = s_+$ and $s = s_-$.

From these data I estimate the sensitivity of the mean, $\boldsymbol{g}'$, by:

$$g_i' = \sqrt{\frac{F_{ii}(s_+) - F_{ii}(s_-)}{s_+ - s_-}} \tag{3.30}$$

where $F_{ij}(s) = \langle \delta r_i \, \delta r_j | s \rangle$ and $\delta r_i = r_i - \langle r_i | s \rangle$. Since the response correlations conditioned on the signal s are influenced by both the noise and nuisance covariance, I use the nuisance sensitivity to estimate the nuisance covariance and remove it from the total response covariance to obtain the noise covariance Σ,

$$\Sigma_{ij} = \frac{F_{ij}(s_+) + F_{ij}(s_-) - (s_+ + s_-) g_i' g_j'}{2} \tag{3.31}$$

3.5.3.2 Optimal strategy and decision boundary

Under this neural model, I can use the following chain to construct an optimal binary choice $\hat{s}_\pm = \mathrm{sgn}\,(\hat{s} - s_0)$:

$$s_\pm \to n \to \boldsymbol{r} \to \hat{n} \to \hat{s} \to \hat{s}_\pm \tag{3.32}$$

Because the model neuronal responses are linearly tuned to orientation, the orientation can be linearly estimated from $\boldsymbol{r}$ by

$$\hat{n} = \boldsymbol{w}^\top (\boldsymbol{r} - \boldsymbol{g}_0) = \boldsymbol{w}^\top \delta \boldsymbol{r} \tag{3.33}$$

where $\delta \boldsymbol{r} = \boldsymbol{r} - \langle \boldsymbol{r} | s \rangle = \boldsymbol{r} - \boldsymbol{g}_0$. The optimal decoding weights for unbiased estimation of the orientation are

$$\boldsymbol{w}_{\mathrm{opt}} = \Sigma^{-1} \boldsymbol{g}' / J_n \tag{3.34}$$

where

$$J_n = \boldsymbol{g}'^\top \Sigma^{-1} \boldsymbol{g}' \tag{3.35}$$

is the Fisher Information about the orientation. The variance of this optimal unbiased estimate is

$$\sigma^2_{\hat{n}|n} = 1/J_n \tag{3.36}$$

Then the distribution of the estimate $\hat{n}$ given a true orientation n is $p(\hat{n}|n) = \mathcal{N}(\hat{n}|n, 1/J_n)$. I can then compute $p(\hat{n}|s)$ by marginalizing over n:

$$p(\hat{n}|s) = \int dn\, p(\hat{n}|n)\, p(n|s) \tag{3.37}$$

$$= \mathcal{N}(\hat{n}|0, s + 1/J_n) \tag{3.38}$$

The optimal unbiased estimate of the task-dependent stimulus s is then

$$\hat{s} = \hat{n}^2 - 1/J_n \tag{3.39}$$

The optimal *binary* choice based on the continuous estimate can be modeled as

$$\hat{s}_\pm = \mathrm{sgn}\,(\hat{s} - s_0) \tag{3.40}$$

where the optimal decision boundary s_0 occurs where the likelihoods of the two possible signals s_+ and s_- are equal:

$$p(\hat{s} = s_0|s_+) = p(\hat{s} = s_0|s_-) \tag{3.41}$$

Using Eq 3.39, I can also write the decision boundary in terms of the orientation variable, as $s_0 = n_0^2 - 1/J_n$, where $\hat{n} = n_0$ satisfies $p(\hat{n} = n_0|s_+) = p(\hat{n} = n_0|s_-)$. From the Gaussian distribution of the orientation estimate $p(\hat{n}|s)$ given the stimulus s, I can compute the two optimal decision boundaries $\hat{n} = n_0$:

$$n_0 = \pm\sqrt{\frac{\log\left(1/J_n + s_+\right) - \log\left(1/J_n + s_-\right)}{\left(1/J_n + s_-\right)^{-1} - \left(1/J_n + s_+\right)^{-1}}} \tag{3.42}$$

Then the optimal decision boundary s_0 for variance discrimination is

$$s_0 = n_0^2 - 1/J_n \tag{3.43}$$

$$= \frac{\log\left(1/J_n + s_+\right) - \log\left(1/J_n + s_-\right)}{\left(1/J_n + s_-\right)^{-1} - \left(1/J_n + s_+\right)^{-1}} - 1/J_n \tag{3.44}$$

3.5.3.3 Nonlinear choice correlation for optimal decoding

Since the sufficient statistics for this task are quadratic, I compute the quadratic choice correlations to test for optimal decoding. Combining Equations 3.33 and 3.39, I obtain a quadratic decoder as function of $\boldsymbol{r}$ to estimate s,

$$\hat{s} = \hat{n}^2 - 1/J_n \tag{3.45}$$

$$= \boldsymbol{w}^\top \delta\boldsymbol{r}\, \delta\boldsymbol{r}^\top \boldsymbol{w} - 1/J_n \tag{3.46}$$

$$= \boldsymbol{W}^\top \boldsymbol{R}(\boldsymbol{r}) - 1/J_n \tag{3.47}$$

where $\boldsymbol{R}(\boldsymbol{r}) = \text{vec}(\delta\boldsymbol{r}\, \delta\boldsymbol{r}^\top)$ and $\boldsymbol{W} = \text{vec}(\boldsymbol{w}\boldsymbol{w}^\top)$ are conversions of matrices to vectors.

To compute the quadratic choice correlations for the inputs to this optimal estimator, $\text{Corr}(\hat{s}, R_{jk}|s)$, I need to compute the Fisher Information in $R_{jk} = \delta r_j \delta r_k$:

$$J_{R_{jk}} = \frac{{F'_{jk}}^2}{\Gamma_{jkjk}} \tag{3.48}$$

$$= \frac{g_j'^2 g_k'^2}{(\Sigma_{jj} + s g_j'^2)(\Sigma_{kk} + s g_k'^2) - (\Sigma_{jk} + s g_j' g_k')^2} \tag{3.49}$$

where I used Equations 2.58 and 3.29 to compute $F_{jk}(s) = \langle \delta r_j \delta r_k | s \rangle$ and $\Gamma_{jkjk} = \text{Var}(\delta r_j \delta r_k | s)$.

I then estimate the Fisher Information in the population,

$$J = \frac{1}{\sigma^2_{\hat{s}|s}} = \frac{1}{\sigma^2_{\hat{n}^2|s}} = \frac{1}{2(1/J_n + s)^2} \tag{3.50}$$

$$= \frac{1}{2(1/\boldsymbol{g}'^\top \Sigma^{-1} \boldsymbol{g}' + s)^2} \tag{3.51}$$

using Equations 3.35, 3.36, and 3.47. I assume that the intrinsic uncertainty of estimating a variance from orientation dominates the smaller uncertainty of estimating orientation from neural activity, so that $J \approx 1/2s^2$.

The predicted optimal choice correlations are the square root of the ratio of Equations 3.49 and 3.51,

$$C_{R_{jk}}^{\text{opt}} \approx \frac{\sqrt{2} g'_j g'_k s}{\sqrt{(\Sigma_{jj} + s g'^2_j)(\Sigma_{kk} + s g'^2_k) - (\Sigma_{jk} + s g'_j g'_k)^2}} \tag{3.52}$$

I can then compute the predicted optimal choice correlations at the decision boundary $\text{Corr}(\hat{s}, R_{jk}|s = s_0)$ by substituting $s = s_0$ into Equation 3.52.

3.5.3.4 Choice correlation for continuous estimate and binary choice

The choice correlation for a binary choice $\hat{s}_\pm$ differs from the choice correlation for a continuous estimate $\hat{s}$ for which I developed most of the theory.

For variance discrimination conditioned on s, both $\hat{s}$ and R_{jk} in my model are quadratic functions of gaussian variables correlated by internal noise and nuisance variation. In that situation, I demonstrate the following relationship between the choice correlations of binary and continuous choices:

$$\frac{\text{Corr}(\hat{s}_\pm, R_{jk}|s = s_0)}{\text{Corr}(\hat{s}, R_{jk}|s = s_0)} = \zeta \tag{3.53}$$

For compactness, I denote the underlying gaussian variables by x and y. I then compute the change in the correlation between squares of these variables induced by a binary threshold on y^2 at the decision boundary $\theta = \langle y^2 \rangle$,

$$\zeta = \frac{\text{Corr}(x^2 \,\text{sgn}\,(y^2 - \theta))}{\text{Corr}(x^2, y^2 - \theta)} \tag{3.54}$$

and show that it does not depend on the correlation between x and y. Without loss of generality, I consider a covariance $\text{Cov}(x, y) = \left(\begin{smallmatrix} 1 & c \\ c & 1 \end{smallmatrix}\right)$. Any scaling of x will be cancelled in the correlation coefficients, and any scaling of y is absorbed by the decision boundary $\theta = \langle y^2 \rangle$ and the signum, so that $\theta = 1$.

I will use the following identities for normally distributed $y \sim p(y) = \mathcal{N}(0,1)$:

$$\left\langle y^4 \right\rangle = 3 \tag{3.55}$$

$$\left\langle y^2 \right\rangle = 1 \tag{3.56}$$

$$\left\langle \left(\operatorname{sgn}\left(y^2-1\right)\right)^2 \right\rangle = 1 \tag{3.57}$$

$$\left\langle \operatorname{sgn}\left(y^2-1\right) \right\rangle = \epsilon_1 \tag{3.58}$$

$$\left\langle y^2 \operatorname{sgn}\left(y^2-1\right) \right\rangle = \epsilon_2 + \epsilon_1 \tag{3.59}$$

where

$$\epsilon_1 = \int dy\, p(y) \operatorname{sgn}\left(y^2-1\right) = 1 - 2\operatorname{erf}(1/\sqrt{2}) \tag{3.60}$$

$$\epsilon_2 = \int dy\, p(y)\, y^2 \operatorname{sgn}\left(y^2-1\right) - \epsilon_1 = \sqrt{\frac{8}{e\pi}} \tag{3.61}$$

I first compute

$$\left\langle x^2 \operatorname{sgn}\left(y^2-1\right) \right\rangle = \tag{3.62}$$

$$= \iint x^2 \operatorname{sgn}\left(y^2-1\right) p(x|y) p(y) dx\, dy \tag{3.63}$$

$$= \int \left(\int x^2 p(x|y) dx \right) \operatorname{sgn}\left(y^2-\theta\right) p(y)\, dy \tag{3.64}$$

For a bivariate gaussian (x, y) with the assumed joint covariance, the conditional variance $\mathrm{Var}(x|y) = 1 - c^2$ does not depend on y, whereas the conditional mean depends linearly on y, as $\langle x|y \rangle = cy$, so

$$\left\langle x^2 \middle| y \right\rangle = \int x^2 p(x|y) dx = 1 - c^2 + c^2 y^2 \tag{3.65}$$

and therefore

$$\left\langle x^2 \operatorname{sgn}\left(y^2-1\right)\right\rangle = \tag{3.66}$$

$$= \int \left(1-c^2+c^2y^2\right) \operatorname{sgn}\left(y^2-1\right) p(y) \tag{3.67}$$

$$\begin{aligned} =&(1-c^2)\left\langle \operatorname{sgn}\left(y^2-1\right)\right\rangle \\ &+ c^2\left\langle y^2 \operatorname{sgn}\left(y^2-1\right)\right\rangle \end{aligned} \tag{3.68}$$

$$=(1-c^2)\epsilon_1 + c^2(\epsilon_2+\epsilon_1) \tag{3.69}$$

$$=c^2\epsilon_2+\epsilon_1 \tag{3.70}$$

I can now substitute the identities from above into the correlation coefficient:

$$C = \operatorname{Corr}(x^2, \operatorname{sgn}\left(y^2-1\right)) \tag{3.71}$$

$$= \frac{\left\langle x^2 \operatorname{sgn}\left(y^2-1\right)\right\rangle - \left\langle x^2\right\rangle \left\langle \operatorname{sgn}\left(y^2-1\right)\right\rangle}{\sqrt{\left(\left\langle x^4\right\rangle - \left\langle x^2\right\rangle^2\right)\left(\left\langle \left(\operatorname{sgn}\left(y^2-1\right)\right)^2\right\rangle - \left\langle \operatorname{sgn}(y^2-1)\right\rangle^2\right)}} \tag{3.72}$$

$$= \frac{c^2\epsilon_2+\epsilon_1-1\cdot\epsilon_1}{\sqrt{(3-1)(1-\epsilon_1^2)}} \tag{3.73}$$

$$= \frac{c^2}{\sqrt{e\pi \operatorname{erf}(1/\sqrt{2}) \operatorname{erfc}(1/\sqrt{2})}} \tag{3.74}$$

I do the same for the correlation without the signum:

$$D = \operatorname{Corr}(x^2, y^2) \tag{3.75}$$

$$= \frac{\left\langle x^2y^2\right\rangle - \left\langle x^2\right\rangle \left\langle y^2\right\rangle}{\sqrt{\left(\left\langle x^4\right\rangle - \left\langle x^2\right\rangle^2\right)\left(\left\langle y^4\right\rangle - \left\langle y^2\right\rangle^2\right)}} \tag{3.76}$$

$$= \frac{1+2c^2-1}{\sqrt{(3-1)(3-1)}} \tag{3.77}$$

$$= c^2 \tag{3.78}$$

where in the second step I used $\left\langle x^2y^2\right\rangle = 1+2c^2$.

Finally, taking the ratio $\zeta = C/D$ from Equations 3.74 and 3.78, the c^2 cancels and I obtain

$$\zeta = \left(e\,\pi\,\mathrm{erfc}\tfrac{1}{\sqrt{2}}\,\mathrm{erf}\tfrac{1}{\sqrt{2}} \right)^{-1/2} \tag{3.79}$$

$$\approx 0.73 \tag{3.80}$$

which is independent of the correlation. The same correction factor holds for cross-terms like $r_i r_j$, which can be expressed as linear combination of squares, $r_i r_j = \frac{1}{2}(r_i + r_j) - \frac{1}{2}r_i^2 - \frac{1}{2}r_j^2$. I use this correction factor ζ to scale my predicted continuous quadratic choice correlations in Figure 3.4.

3.5.4 Comparing choice correlations from internal or external noise

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

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

$$\boldsymbol{R}(\boldsymbol{r}) = \boldsymbol{F}'s + \boldsymbol{G}'n + \boldsymbol{\eta} \tag{3.81}$$

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

$$\Gamma = H + \xi \boldsymbol{G}' \boldsymbol{G}'^{\top} \tag{3.82}$$

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

$$\boldsymbol{w} = \frac{\Gamma^{-1}\boldsymbol{F}'}{J_1} \tag{3.83}$$

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

$$\hat{s} = \boldsymbol{w}^{\top}\boldsymbol{R}(\boldsymbol{r}) + b \tag{3.84}$$

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

$$c_{R_k}^{\text{sub}} = \text{Corr}(R_k, \hat{s}|s, n) \tag{3.85}$$

I 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.

I can express these choice correlations as:

$$c_{R_k}^{\text{sub}} = \frac{\text{Cov}(R_k, \hat{s}|s, n)}{\sigma_{R_k|s,n}\sigma_{\hat{s}|s,n}} \tag{3.86}$$

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

$$\text{Cov}(\boldsymbol{R}, \hat{s}|s, n) = \langle \boldsymbol{R}\hat{s}|s, n\rangle \tag{3.87}$$

$$= \langle \boldsymbol{R}\boldsymbol{R}^\top|s, n\rangle \boldsymbol{w} \tag{3.88}$$

$$= \frac{H\Gamma^{-1}\boldsymbol{F}'}{J_1} \tag{3.89}$$

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

$$\Gamma^{-1} = (H + \xi \boldsymbol{G}'\boldsymbol{G}'^\top)^{-1} \tag{3.90}$$

$$= H^{-1} - \frac{H^{-1}\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}}{1/\xi + \boldsymbol{G}'^\top H^{-1}\boldsymbol{G}'} \tag{3.91}$$

Substituting this inverse covariance into Equation 3.87, I obtain

$$\text{Cov}(\boldsymbol{R}, \hat{s}|s, n) \tag{3.92}$$

$$= \frac{1}{J_1} H(H^{-1} - \frac{H^{-1}\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}}{1/\xi + \boldsymbol{G}'H^{-1}\boldsymbol{G}'^\top})\boldsymbol{F}' \tag{3.93}$$

$$= \frac{1}{J_1}(\boldsymbol{F}' - \frac{\boldsymbol{G}'\boldsymbol{G}'^\top H^{-1}\boldsymbol{F}'}{1/\xi + \boldsymbol{G}'H^{-1}\boldsymbol{G}'^\top}) \tag{3.94}$$

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

$$\boldsymbol{J}(s, n) = \begin{bmatrix} J_{11} & J_{12} \\ J_{12} & J_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{F}'^\top H^{-1}\boldsymbol{F}' & \boldsymbol{F}'^\top H^{-1}\boldsymbol{G}' \\ \boldsymbol{G}'^\top H^{-1}\boldsymbol{F}' & \boldsymbol{G}'^\top H^{-1}\boldsymbol{G}' \end{bmatrix} \tag{3.95}$$

With these substitutions, I have

$$\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s, n) = \frac{1}{J_1}\left(\boldsymbol{F}' - \frac{J_{12}}{1/\xi + J_{22}}\boldsymbol{G}'\right) \tag{3.96}$$

The denominator of Equation 3.86 involves the variance of the sufficient statistics,

$$\sigma^2_{R_k|s,n} = H_{kk} \tag{3.97}$$

and the variance of the brain's decoder,

$$\sigma^2_{\hat{s}} = \boldsymbol{w}^\top H \boldsymbol{w} \tag{3.98}$$

$$= \boldsymbol{w}^\top (\Gamma - \xi \boldsymbol{G}'\boldsymbol{G}'^\top)\boldsymbol{w} \tag{3.99}$$

$$= \frac{1}{J_1} - \frac{J_{12}^2}{\xi J_1^2}\frac{1}{(1/\xi + J_{22})^2} \tag{3.100}$$

where I used the following results:

$$\boldsymbol{w}^\top \boldsymbol{G}'\boldsymbol{G}'^\top \boldsymbol{w} = \left(\frac{\boldsymbol{F}'\Gamma^{-1}}{J_1}\boldsymbol{G}'\right)^2 \tag{3.101}$$

$$= \frac{1}{J_1^2}\left(\boldsymbol{F}' H^{-1}\boldsymbol{G}' - \frac{\boldsymbol{F}'\Gamma^{-1}\boldsymbol{G}'\boldsymbol{G}' H^{-1}\boldsymbol{G}'}{1/\xi + \boldsymbol{G}' H^{-1}\boldsymbol{G}'}\right)^2 \tag{3.102}$$

$$= \frac{1}{J_1^2}\left(J_{12} - \frac{J_{12}J_{22}}{1/\xi + J_{22}}\right)^2 \tag{3.103}$$

$$= \frac{J_{12}^2}{\xi^2 J_1^2}\frac{1}{(1/\xi + J_{22})^2} \tag{3.104}$$

Combining the results from Equation 3.96, 3.100 and 3.97, I can compute Equation 3.86

$$c^{\mathrm{sub}}_{R_k} = \mathrm{Corr}(R_k, \hat{s}|s, n) \tag{3.105}$$

$$= \frac{\mathrm{Cov}(R_k, \hat{s}|s, n)}{\sigma_{R_k|s,n}\sigma_{\hat{s}|s,n}} \tag{3.106}$$

$$= \frac{\frac{1}{J_1}\left(F'_k - \frac{J_{12}}{1/\xi + J_{22}}G'_k\right)}{\sqrt{H_{kk}}\sigma_{\hat{s}|s,n}} \tag{3.107}$$

The optimal choice correlation when there is natural nuisance variation (Eq 3.6) is given by

$$C_{R_k}^{\text{opt}} = \sqrt{\frac{J_{1,R_k}}{J_1}} = \frac{F_k'}{\sigma_{R_k|s}\sqrt{J_1}} \tag{3.108}$$

where $J_{1,R_k} = F_k'/\sigma_{R_k|s}$ is the Fisher Information in R_k about s when there is natural nuisance variation, and $\sigma_{R_k|s} = \sqrt{H_{kk} + \xi G_k'^2}$ 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: $C_{R_k}^{\text{opt}}$ and $c_{R_k}^{\text{sub}}$. I find that the nuisance-conditioned choice correlations $c_{R_k}^{\text{sub}}$ relate to the optimal nuisance-averaged choice correlations $C_{R_k}^{\text{opt}}$ according to

$$c_{R_k}^{\text{sub}} = \beta_k C_{R_k}^{\text{opt}} - \gamma_k \tag{3.109}$$

where I have defined the following constants:

$$\beta_k = \frac{\sigma_{R_k|s}}{\sigma_{R_k|s,n}} \frac{1}{\sqrt{J_1}\sigma_{\hat{s}|s,n}} \tag{3.110}$$

$$= \sqrt{\frac{H_{kk} + \xi G_k'^2}{H_{kk}}} \frac{1}{\sqrt{J_1}\sigma_{\hat{s}|s,n}} \tag{3.111}$$

$$= \sqrt{\frac{H_{kk} + \xi G_k'^2}{H_{kk}}} \frac{1}{\sqrt{1 - \frac{J_{12}^2}{\xi J_1} \frac{1}{(1/\xi + J_{22})^2}}} \tag{3.112}$$

and

$$\gamma_k = \frac{G_k'}{\sqrt{H_{kk}}} \frac{J_{12}}{(1/\xi + J_2) J_1 \sigma_{\hat{s}|s,n}} \tag{3.113}$$

The slope β_k and offset γ_k of the relationship between these two types of choice correlations (Equation 3.109) depends on the amount of nuisance variation compared to internal noise and the suboptimality of the brain's decoding strategy. When the

signal and nuisance can be disentangled, that is, estimated nearly independently using the statistics $\boldsymbol{R}(\boldsymbol{r})$, then J_{12} is small and the choice correlations driven purely by internal fluctuations closely match the optimal choice correlations in the presence of nuisance variation (Figure 3.5A). In contrast, when nuisance variations remain partially 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 3.5B).

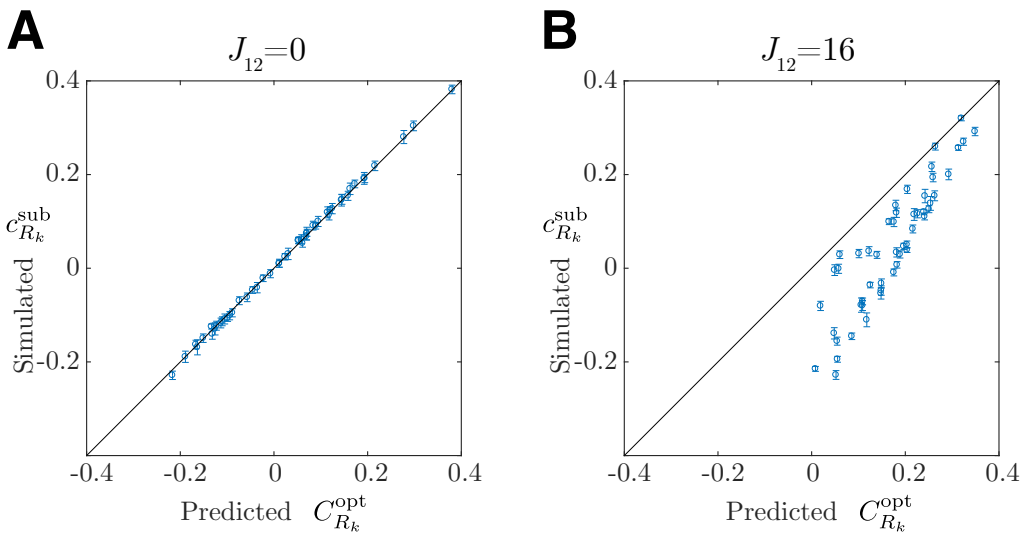


Figure 3.5 : Comparing choice correlations caused by internal and external noise. (**A**) When estimates of nuisance variables are independent of estimates of task-relevant signals, the optimal choice correlations driven by internal noise, $c^{\text{sub}}_{R_k}$, match the optimal pattern $C^{\text{opt}}_{R_k}$ expected for optimal decoding under natural nuisance variation (Equation 3.6). (**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.

For the simulations in Figure 3.5, I set the sufficient statistics to be linear $\boldsymbol{R}(\boldsymbol{r}) = \boldsymbol{r}$ for simplicity. Neural responses were generated from a Gaussian distribution with a stimulus-dependent mean and identity covariance $H = I$: $p(\boldsymbol{r}|s,n) = \mathcal{N}(\boldsymbol{F}'s + \boldsymbol{G}'n, I)$. In Figure 3.5A, $\boldsymbol{F}'$ and $\boldsymbol{G}'$ are set to be orthogonal to ensure $J_{12} =$

$\boldsymbol{F}'^{\top} H^{-1} \boldsymbol{G}' = 0$. They are picked from the eigenvector of a symmetric matrix $A^{\top} A$, where A is a matrix whose elements are generated from uniform distribution bounded by 0 and 1. In Figure 3.5B, each element in $\boldsymbol{F}'$ and $\boldsymbol{G}'$ is drawn from a uniform distribution over the interval $[0, 1]$. I simulate 10000 responses of a population with $N = 50$ neurons. The stimulus is set to 0 and the nuisance is fixed to be 1. The brain's decoder assumes a Gaussian prior over the nuisance variation with zero mean and variance $\xi = 2$. The decoding weights follow Equation 3.83, and the stimulus is estimated using Equation 3.84. Choice correlations in this fixed-nuisance experiment are computed by Equation 3.85 (vertical axis in Figure 3.5). The predicted optimal choice correlation is computed by Equation 3.108 (horizontal axis in Figure 3.5). In this setting, $\beta_k \approx 1$ when $J_{12} = 0$.

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

Chapter 4

Coarse-grained computation

秦使武安君白起击，大破赵於长平，四十馀万尽杀之。

—《史记》司马迁

State Qin commanded General Baiqi to attack state Zhao. Baiqi vanquished Zhao's army at Changping and butchered Zhao's four hundred thousand soldiers.

— 'Records of the Grand Historian', Sima Qian

Historians record the past in their books; they attempt to tell the truth and provide perspectives. There are two very different ways to record history. One is advocated by Thucydides, the father of 'scientific history' who recounts the Peloponnesian War between Sparta and Athens from fifth-century BCE to 411 BCE. He gathered enormous amounts of details and illustrated them as clearly as possible. In contrast, Sima Qian, the Grand Historian of China, leaned towards 'Simplicity and Conciseness' as his style in his 'Records of the Grand Historian.' As shown in the quote, Sima Qian very coarsely described the process of the war, focusing on the consequences across time and space, extracting the essential aspects of numerous similar wars of that period (named as the Warring States period in Chinese history). All these wars were for the same purpose: annexing other contender states and unifying the empire. One can understand them by characterizing how different states are transformed by wars, rather than by fine details of the numerous battles.

Just as historians' mission is to understand history, neuroscientists want to figure out the principles of the brain. One challenge to adopting Thucydides' method to figure out the fine detail of large-scale neural computation is that it requires a massive amount of data. More importantly, when we only want to understand the brain's algorithm [2], the fine biophysical details may not be necessary because there are many equivalent ways for the brain to implement its computations [69]: the solutions are degenerate.

The number of equivalent computations expands in the presence of uncertainty, which makes it both more difficult and less relevant to distinguish fine features of the neural representations. As described in Chapter 2, information-limiting noise in large populations of cortical neurons can cause large uncertainties; this can create large degeneracies. Just as what Sima Qian did in his study, one may understand the neural computation best at the representational level [2] — by identifying stimulus representations based on different statistics of neural populations, and then characterizing how these representations transform the information about task-relevant variables. We would particularly like to study these computations at a coarser level than single neurons.

Some types of coarse-graining are already implicit in the different models and methods of neuroscience. For example, rate coding theories [70, 71] assume that the precise timing of the spikes does not contain much information about the stimulus. Typical analyses extract representations from the spike count in 100-millisecond time windows, and build encoding and decoding model based on this firing rate. When neuroscientists measure neural activity using eCog [72] or EEG [72, 73], the measurements are spatially coarse, pooling over large populations of neurons, and the analyses must search for signals in these pooled signals. Another common technique, functional mag-

netic resonance (fMRI) [74, 75], is both spatially and temporally coarse, pooling over voxels that are $\sim 1\,\mathrm{mm}^3$ and integrating over seconds. These approaches all enforce a form of dimensionality reduction. Ideally, these reduced measurements would not lose much information about the task-relevant variables [76]. However, the way that these coarse measurement techniques reduce dimensions may not be ideal for understanding the neural encoding, since they are constrained by experimental techniques rather than coding properties, and thus may lose much of the encoded information. In some cases, when the relevant signals are sufficiently low-dimensional, arbitrarily-structured coarse-graining can be sufficient to recover the underlying signal [77]. However, given access to fine-scale measurements, the existence of intrinsic neural variability from noise and unknown internal states means that it may be advantageous to design a better, targeted coarse-graining — one that performs a dimensionality reduction that properly accounts for the way that information is encoded.

In this chapter, I will explore this coarse-graining concept in the context of redundant nonlinear population codes. I will start by proposing a new concept of equivalence classes for neural transformations. The number of equivalent computations expands in the presence of uncertainty, which makes it both more difficult and less relevant to distinguish fine features of the neural representations. This suggests that we can understand the neural transformations by picking a convenient nonlinear basis that spans interesting equivalence classes, instead of trying to reproduce the biophysical details.

As described in Chapter 2, information-limiting correlations in large populations of cortical neurons can cause large redundancies and large uncertainty about the sensory stimulus. Just as what Sima Qian did in his study, one may understand neural computation best at the representational and algorithmic levels [2], while leaving

the fine-scale implementation details to other studies where details may be more important, like pharmacology or medicine. The goal of this chapter is to develop a mathematical framework for this type of coarse-grained description of neural computations.

4.1 Equivalence classes of neural transformations

In Chapter 2, we studied information-processing in the brain with a feedforward processing chain:

$$(s, \boldsymbol{n}) \rightarrow \boldsymbol{r} \rightarrow \boldsymbol{R} \rightarrow \hat{s} \tag{4.1}$$

where s is the task-relevant stimulus, $\boldsymbol{n}$ are nuisance variables, $\boldsymbol{r}$ and $\boldsymbol{R}$ are upstream and downstream neural responses, and $\hat{s}$ is a perceptual estimate of the stimulus. This estimate may be used as a continuous behavioral output, as in a tracking task, or may drive a subsequent action, $a(\hat{s})$. In this processing chain, neural *encoding* describes the relationship between the stimulus and neural responses, $(s, \boldsymbol{n}) \rightarrow (\boldsymbol{r}, \boldsymbol{R})$. Neural *recoding* describes the transformation of the upstream responses into downstream neural responses, $\boldsymbol{r} \rightarrow \boldsymbol{R}$. Neural *decoding* describes the use of neural activity patterns to generate behavior, $(\boldsymbol{r}, \boldsymbol{R}) \rightarrow \hat{s}$.[i]

In a typical task, an agent observes a multidimensional stimulus $(s, \boldsymbol{n})$ and must act upon one particular relevant aspect of that stimulus, s, while ignoring the rest, $\boldsymbol{n}$. In the sensory inputs, the nuisance variation, $\boldsymbol{n}$, is often entangled with the task-relevant stimulus, s [4]. This entangling requires that useful recoding transformations

[i]Others sometimes use 'decoding' as a way of bounding the quality of the encoding: if you can decode something reliably from the neural data, then it must have been encoded. Here we are asking questions about the *brain's* decoding, which need not extract all of the information it encodes.

be nonlinear to allow the brain to decode the relevant stimulus from downstream neurons using simple (*e.g.* linear) operations. The recoding and the decoding can be summarized as the neural transformations between the recorded neurons and the behavioral estimate. This neural transformation tells us about the brain's algorithm. We want to characterize and infer these neural transformations using neural data recorded while the brain performs the given task.

For example, as we showed in the orientation estimation task with varying spatial phase (Section 2.4.1), the unknown nuisance variation of the spatial phase relegates the neural tuning to the quadratic (and higher-order) statistics, and thus requires the neural transformation also should be at least quadratic. This quadratic computation can be approximately implemented in different ways: squaring the neural responses directly (Figure 4.1A), or summing a set of rectifiers (ReLUs) with uniformly spaced thresholds (Figure 4.1B).

Squaring is still not equivalent to summing a set of ReLUs when the uncertainty of the estimate is extremely small: there will always be some approximation error. However, when the uncertainty of the estimate is large enough, direct squaring and the summing of ReLUs will be equivalent — distinguishing fine features of the two neural transformations will be difficult and, moreover, unnecessary.

We define an equivalence class of neural transformations as the set of all nonlinear functions of the neural responses that produce essentially the same behavior. Note that these are 'soft' equivalence classes, in the sense that these transformations are technically distinguishable with some probability, and we define the boundaries of the equivalence classes that are large enough to make this probability as small as we like. Previous work [78] similarly characterized the indistinguishability of different probability distributions, where the volume of the indistinguishable models is proportional

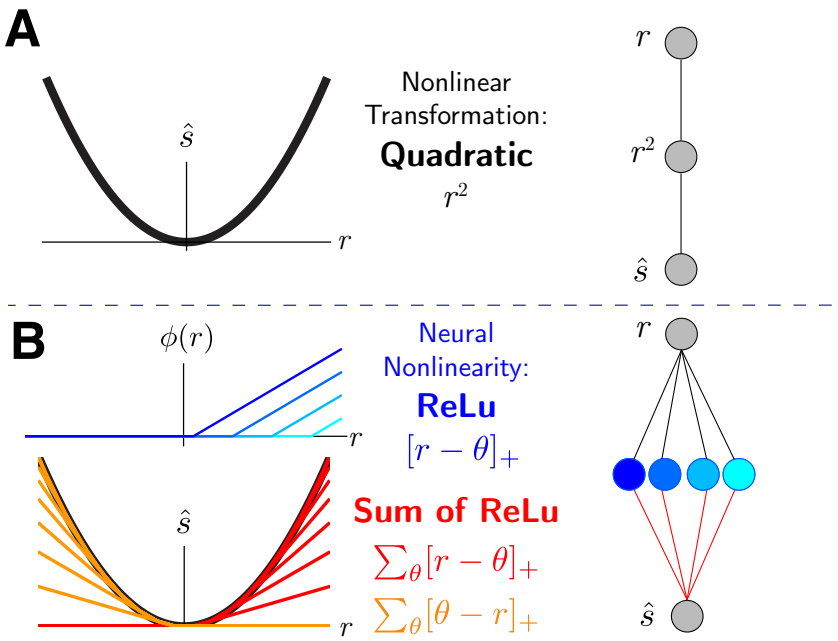


Figure 4.1 : Two ways to approximate a quadratic transformation. (**A**) A quadratic transformation can be realized by directly squaring the input. (**B**) Equivalently, a quadratic transformation can also be realized by summing a set of rectifiers $[\cdot]_+$ with uniformly spaced thresholds $\boldsymbol{\theta}$.

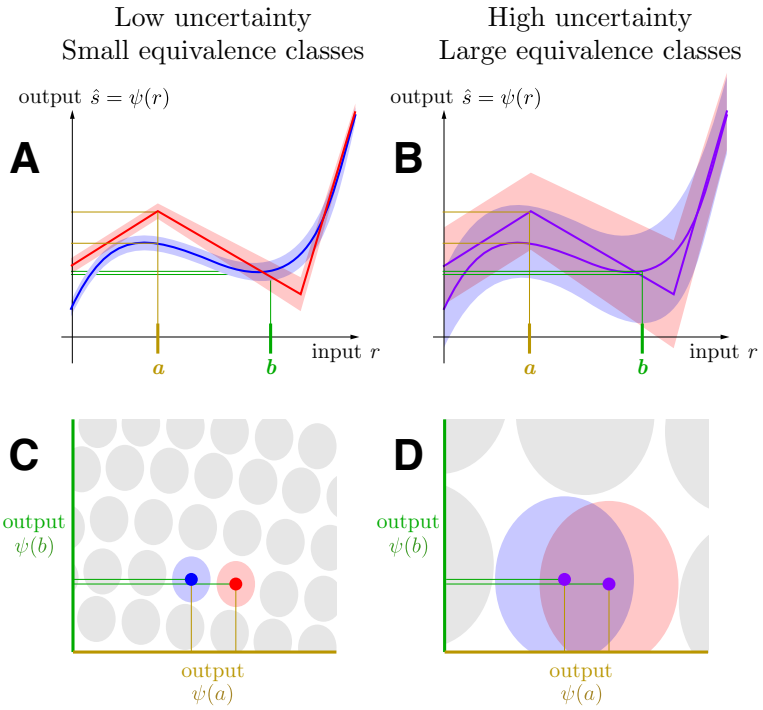


Figure 4.2 : Equivalence classes of neural transformations. The equivalence classes of neural transformations are defined as the set of all nonlinear functions of the neural responses (input) that produce essentially the same behavior (output). Two functions, a piecewise linear function (red) and a cubic function (blue) are evaluated at two input points, $r = a$ (brown) and $r = b$ (green). The shading represents the uncertainty of the output functions, such as 95% confidence intervals. (**A**) When the uncertainty of the output is low, the two functions can be reliably distinguished at $r = a$. Thus, the two functions are not in the same equivalence class. (**B**) When the uncertainty of the output is high, the two functions (purple) cannot be reliably distinguished at any input measurements. The two functions are in the same equivalence class. (**C**) When the uncertainty of the output is small, more nonlinear functions can be distinguished, indicated by the small volume of the equivalence classes. Here, we plot the output of the nonlinear functions output at two input points $r = a$ and $r = b$. The blue point is the mean output of the cubic function while the red point is the mean output of the piecewise linear function. Each ellipse denotes the region in which different points cannot be reliably distinguished, although different ellipses can be distinguished. Here there are many small equivalence classes. The difference between the red and blue mean outputs is larger than their uncertainties by a large enough margin that we can distinguish the two functions reliably, and they do not fall into the same equivalence class. (**D**) When the uncertainty of the output is large, more nonlinear functions become indistinguishable, so there are fewer equivalence classes, and each is larger. The piecewise linear function and the cubic cannot be distinguished here because the difference between the functional means is smaller than the uncertainty, so the two functions fall into the same equivalence class.

to the inverse of the determinant of Fisher information matrix. The inverse of the determinant of the Fisher information matrix is closely related to the output uncertainty [79]. Here in our definition of the equivalence class for neural transformations, the output variability determines distinguishability of different nonlinear transformations. When the uncertainty of the output is small, the size of the equivalence class is also small because small changes in the function yield distinguishable outputs (Figure 4.2A,C). However, in a redundant code where the neural population contains limited information — and thus exhibits much more output variability than a population with extensive information — then many more functions will be hard to distinguish and thus fall into the same equivalence class. Equivalence classes should then be much broader for redundant codes (Figure 4.2B,D).

4.2 Describing brain's nonlinear neural transformation

The actual neural transformations that the brain uses to estimate the sensory stimulus for complex naturalistic tasks is usually nonlinear and unknown. We can assume it can be expressed as a function of the neural responses $\hat{s} = \psi(\boldsymbol{r})$. A crucial question is then: how can we describe this nonlinear transformation? In a redundant code, the equivalence classes of neural transformations are broad. Instead of trying to reproduce the biophysical details of the actual neural transformations, we can try to pick a convenient nonlinear basis that spans interesting equivalence classes, and describe the brain's nonlinear neural transformation with this nonlinear basis.

One straightforward candidate for the nonlinear basis is the monomial basis. We can consider products of powers of neural responses, $\boldsymbol{R} = \prod_i r_i^{a_i}$, *i.e.* a Taylor series for the nonlinearities, $\boldsymbol{R} = \{1, r_i, r_i r_j, r_i r_j r_k, \cdots\}$, to approximate the actual nonlinear neural transformations (Figure 4.3A).

We could also consider other nonlinear basis functions, such as radial basis functions, random sigmoidal nonlinear neural networks, or nonlinear features learned in deep discriminative networks. As long as they can capture the equivalence class of the brain's nonlinear neural transformation, then these basis functions could be useful for understanding the essential properties of the neural transformations.

4.3 Inferring fine-grained weights

For a chosen nonlinear basis, we can infer the fine-grained weights by linear regression between the functions $\boldsymbol{R}$ and the behavioral choice. As above, we take this choice to be a continuous estimate, $\hat{s}$, of the stimulus near a reference s_0. The decoding weights can then be written in terms of correlations between $\boldsymbol{R}$ and this choice:

$$\boldsymbol{w} = \mathrm{Cov}(\boldsymbol{R}|s)^{-1}\mathrm{Cov}(\boldsymbol{R}, \hat{s}|s) \tag{4.2}$$

$$= \sigma_{\hat{s}}\Gamma^{-1}H\boldsymbol{C}_{\boldsymbol{R}} \tag{4.3}$$

where $\Gamma = \mathrm{Cov}(\boldsymbol{R}|s)$ is the covariance of $\boldsymbol{R}$, $\boldsymbol{C}_{\boldsymbol{R}} = \mathrm{Corr}(\boldsymbol{R}, \hat{s}|s)$ is a vector of nonlinear choice correlations between the behavior and the basis, $\sigma_{\hat{s}}$ is the standard deviation of the estimate $\hat{s}$, and H is a diagonal matrix whose elements are the standard deviation of $\boldsymbol{R}_i$, $H_{ij} = \sqrt{\Gamma_{ii}}\delta_{ij}$. This approach to inferring the fine-grained weights is a straightforward generalization of the method in [20] to nonlinear functions of population responses.

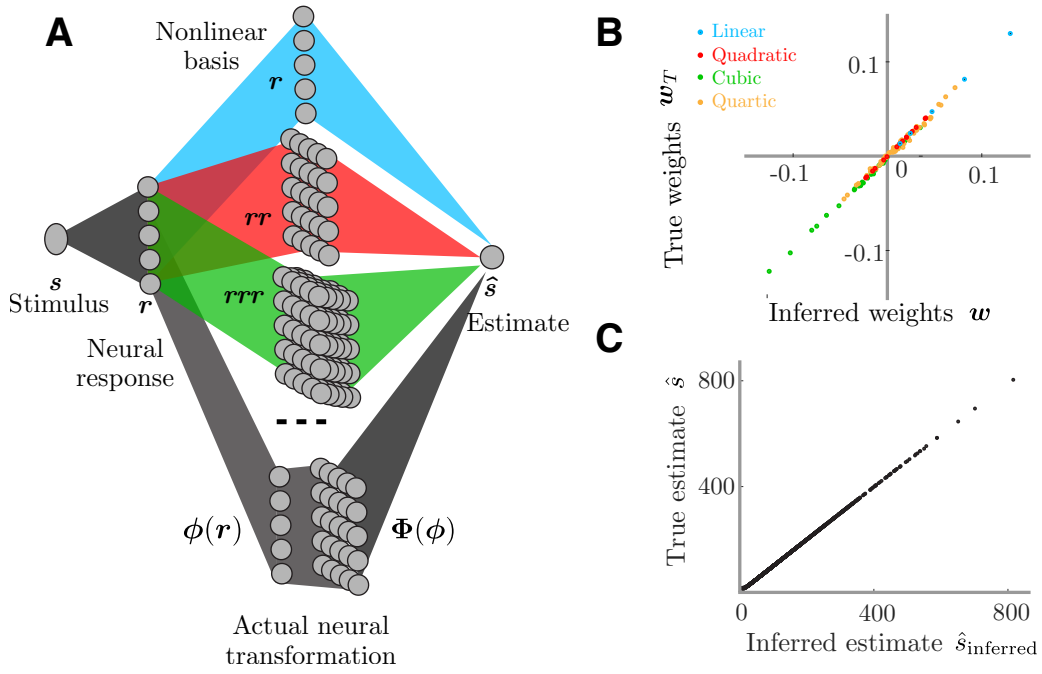


Figure 4.3 : Approximating the neural transformations with the nonlinear basis. (**A**) The actual neural transformations to formulate the decoder can be complex. We can choose a convenient basis and infer the corresponding weights. In this example, the actual neural transformation is composed of two layers: the first with $\phi_i(A\boldsymbol{r}+\boldsymbol{b})$, where ϕ_i is tanh, and the second layer is a quadratic function of ϕ_i and ϕ_j: $\Phi_k = \phi_i\phi_j$. We pick the monomial basis to infer the actual neural transformation. (**B**) Using Eq 4.3, we can infer the true weights on the monomial basis up to the fourth-order. (**C**) The inferred neural transformation successfully recovers the estimate on each trial, implying that the inferred neural transformation and the actual neural transformation fall into the same equivalence class.

In Figure 4.3, we build an example to show that we can describe the complex actual neural transformation with a monomial nonlinear basis that spans the equivalence class. Even when the chosen functions do not match, as long as the true functions are spanned by the basis functions, the inferred fined-grained weights on each monomial basis element successfully recover the coefficients of the true weights given by a Taylor series on the actual nonlinearities, ϕ, Φ (Figure 4.3B). This inferred neural transformation also successfully recovers the estimate on each trial (Figure

4.3C), implying that the inferred neural transformation and the actual neural transformation fall into the same equivalence class. Details about the simulation is in Section 4.7.1.

4.4 Redundant codes

One problem with this fine-grained approach is there are too many weights to infer from limited data. When the actual neural transformation contains significant higher-order statistics of responses, the fine-grained method for weights based on an unbiased nonlinear basis could require a number of statistics that grows exponentially with the order: if linear decoding requires N units, then quadratic decoding would need N^2, and hectic decoding (100^{th}-order polynomial) would need N^{100}. In addition to requiring many units, measuring such high-order nonlinear decoding *across* neurons — *e.g.* for cubic decoding $r_i r_j r_k$ rather than merely r_i^3 — would also require those neurons to recorded simultaneously. Many technologies now enable recordings from many neurons simultaneously. Nonetheless, there is little chance of collecting enough data to disambiguate between all of these possible contributors to avoid overfitting. Clearly there are major (perhaps insurmountable) practical challenges associated with inferring fine-grained decoding weights.

The only hope of understanding nonlinear decoding is therefore to reduce the dimensionality of the functions we wish to identify. This reduced dimensionality could arise from a strong prior over the decoder structure, such as a high degree of sparsity. It could also arise from large equivalence classes, due to substantial variability that cause many of these nominally different decoders to produce outputs that are indistinguishable up to the perceptual uncertainty.

Indeed, in a highly redundant code, we show that it is not necessary to characterize

fine-scaled neural transformations on every element in the nonlinear basis. In this section, we will first examine how the redundancy in a neural population can be induced by a cortical expansion. Then we will revisit features of redundant codes (Section 2.2.4 and extend concept of 'information-limiting correlations' to multiple neural populations, and multiple subsets of statistics derived from those populations). This extended characterization of redundant codes leads to a natural coarse-grained description of nonlinear neural transformations.

4.4.1 Redundancy induced by cortical expansion

In Section 2.2.4, we emphasized how redundancy in neural populations arises when cortical populations are much larger than the upstream sensory neuron populations that provide their inputs. The data processing inequality implies that the information in the cortical neurons cannot exceed the information in the sensory neurons. If the information in the sensory neurons is limited, the information in the cortical neurons cannot be extensive. In Figure 4.4, we simulate a cortical expansion from a smaller population of upstream neurons to a much larger downstream population. As a result, the total information content, and the information in each subset of statistics, are all limited, so the downstream population forms a redundant code.

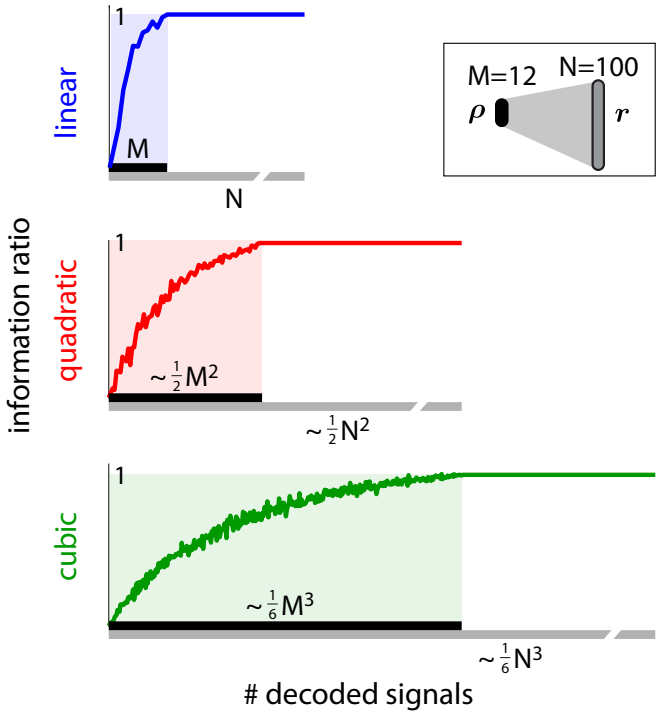


Figure 4.4 : Redundancy induced by cortical expansion. The upstream neural population $\boldsymbol{\rho}$ is generated from a cubic code whose sufficient statistics are polynomial up to third-order (Section 2.4.4 and 2.5.5). Then the upstream neural population is linearly expanded to a much larger recorded neural population $\boldsymbol{r}$ (inset). The statistics of $\boldsymbol{r}$ are separated into three subsets according to their orders. The decoder for the subset $\boldsymbol{R}_z$ is computed as $\hat{s}_z = \boldsymbol{w}_z \boldsymbol{R}_z^{\text{subset}} + c_z$, where $\boldsymbol{R}_z^{\text{subset}}$ contains increasing number of units randomly chosen from $\boldsymbol{R}_z$, $\boldsymbol{w}_z$ are the optimal weights and $z = 1, 2, 3$ denotes different subset of neural statistics whose orders are linear, quadratic and cubic. The inverse variance of the decoder is interpreted as the information in $\boldsymbol{R}_z^{\text{subset}}$. Here we show that the information in the subset saturates after decoding M^z units. The information ratio between the information in $\boldsymbol{R}_z^{\text{subset}}$ and the information in the corresponding statistics in ρ saturates to be 1. Thus there exists significant information redundancy in the downstream neural population. Details of the simulation are in Section 4.7.2.1.

4.4.2 Global and local information-limiting noises

To describe the neural population as a redundant code, previous work [1, 44] introduced 'Information-limiting correlations'. These noise fluctuations exhibit patterns across the population that are indistinguishable from those induced by the signal

itself. Consequently, the brain cannot average away this noise.

When multiple redundant neural populations contribute to a sensory percept, it is useful to consider the information-limiting noise in each population, both separately and together. We will distinguish two types of noise, global and local information-limiting noises.

'*Global* information-limiting noise' is defined as the noise that is in the direction of $\boldsymbol{F}'$: $\eta = \boldsymbol{F}'ds$, where $\boldsymbol{F}' = \partial \langle \boldsymbol{R}(\boldsymbol{r}|s) \rangle / \partial s$ is the sensitivity of the mean of all informative nonlinear statistics, ds is the noise with zero mean and variance equal to ϵ. Notice that the signal is also in this direction. Thus this noise is indistinguishable from the signal. That's why we call this noise is '*bad*'. The covariance of this noise is defined as '*Global* information-limiting correlations': $\mathrm{Cov}(\eta) = \epsilon \boldsymbol{F}' \boldsymbol{F}'^{\top}$. The noise covariance of the neural population's statistics can be decomposed into $\Gamma = \Gamma_0 + \epsilon \boldsymbol{F}' \boldsymbol{F}'^{\top}$, where ϵ is the scaling of the information-limiting component and Γ_0 is the covariance of the noise that can be averaged away with many neurons. Using Eq 2.22, we can compute the Fisher Information of the entire population as $J = \frac{1}{\epsilon + 1/J_0}$, where $J_0 = \boldsymbol{F}' \Gamma_0^{-1} \boldsymbol{F}'$ is the 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 $1/\epsilon$. When the variance of Global information-limiting noise is finite, the population information is limited even when the size of the population grows without bound. The global information-limiting correlations for linear codes are the same as the original formulation of information-limiting correlations [1].

In contrast, '*local*'[ii] information-limiting noise' is defined as the noise in the direc-

[ii] Unfortunately in this section we now have two notions of local: responses restricted to a subset of nonlinear statistics, and stimuli restricted to be near a reference. Throughout this chapter we are always addressing decoding or estimation near a reference, and 'local' will refer to the former

tion of the signal in a chosen subset z of response statistics, $\boldsymbol{R}_z$. For the fine estimation or discrimination tasks we consider, this direction is $\boldsymbol{F}'_z$, where $\boldsymbol{F}'_z = \partial \langle \boldsymbol{R}_z(\boldsymbol{r}|s) \rangle / \partial s$ is the sensitivity of the mean of informative nonlinear statistics present in the z-th subset of $\boldsymbol{R}$. This subset can be linear statistics of different populations of neuronal responses, $\boldsymbol{r}_z$, or can be subsets of nonlinear statistics of those responses, $\boldsymbol{R}_z(\boldsymbol{r})$. The noise itself takes the form $\boldsymbol{\eta}_z = \boldsymbol{F}'_z ds_z$ where we assume scalar fluctuations along the signal direction to be $ds_z \sim \mathcal{N}(0, \epsilon_{zz})$, with variance ϵ_{zz}. Locally within individual subsets z, this noise looks like information-limiting noise, but at least some of it could potentially be filtered out by appropriately combining signals from multiple subsets of responses. The covariance of this noise is defines '*Local* information-limiting correlations': $\mathrm{Cov}(\eta_z) = \epsilon_z \boldsymbol{F}'_z \boldsymbol{F}'^{\top}_z$.

The distinction between the local and global information-limiting noise at the population level is analogous to different types of correlations at the single-neuron level. For example, we can examine the result for a pair of idealized neurons, x and y [19]. Imagine the activity of neuron x is given by $x = s + n$, where s is signal and n is standard Gaussian noise. The information in neuron x is then 1. Imagine that neuron y carries the same stimulus-related signal and the same noise on every trial, except that the noise is multiplied by a factor of 2, $y = s + 2n$. The information in neuron y is then $1/4$. The noise n is locally information limiting for x and y separately. However, if we decode x and y simultaneously, we can filter out the local noise by combining the two neurons: $2x - y = s$. Then we can get infinite information about the signal, so this noise is not globally information-limiting. If we had instead $y = s + n$, then we could not distinguish noise from signal even when we decode them simultaneously, and it would indeed be (globally) information-limiting.

---

notion, specific to one population.

This example generalizes to multiple subsets of neural statistics $\boldsymbol{R}_z$, instead of single neurons x and y. The relevant signal and local information-limiting noise lies along the direction of $\boldsymbol{F}'_z$.

To make this concept concrete, we consider an example using a *linear* population code with two populations [19, 80]. Each population, $\boldsymbol{r}_z$, has mean responses tuned to the variable of interest, and this mean completely captures all information about the signal ($p(\boldsymbol{r}_z|s)$ is in the exponential family with linear sufficient statistics). The signal can be locally estimated from $\boldsymbol{r}_z$ linearly, $\hat{s}_z = \boldsymbol{w}_z^\top \boldsymbol{r}_z + c_z$. When the covariance of the population contains local information-limiting correlations $\epsilon_z \boldsymbol{f}'_z \boldsymbol{f}'^\top_z$, the local estimate variance $\sigma^2_{\hat{s}_z}$ is lower-bounded by ϵ_z even with optimal decoding. However, when we decode globally, we might be able to exceed this bound, depending on how these information-limiting fluctuations are correlated between populations $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. For example, if some part of the local information-limiting noise in $\boldsymbol{r}_1$ is injected by $\boldsymbol{r}_2$ (such as in a hierarchical sampling model for inference [65, 81, 82]), then a decoder that reads out from both could subtract the known variability from $\boldsymbol{r}_1$ [83]. This would then remove some of these *local* information-limiting fluctuations. If all of the local information-fluctuations can be removed in this manner, such that there are no global information-limiting correlations, then the information content would be determined by the remaining noise Γ_0, which could allow the information to grow extensively with population size. (I include the case with no global information-limiting noise for completeness and to clarify the nature of these correlations, but view it as an unlikely scenario in the brain since information should be limited by the smaller sensory population.)

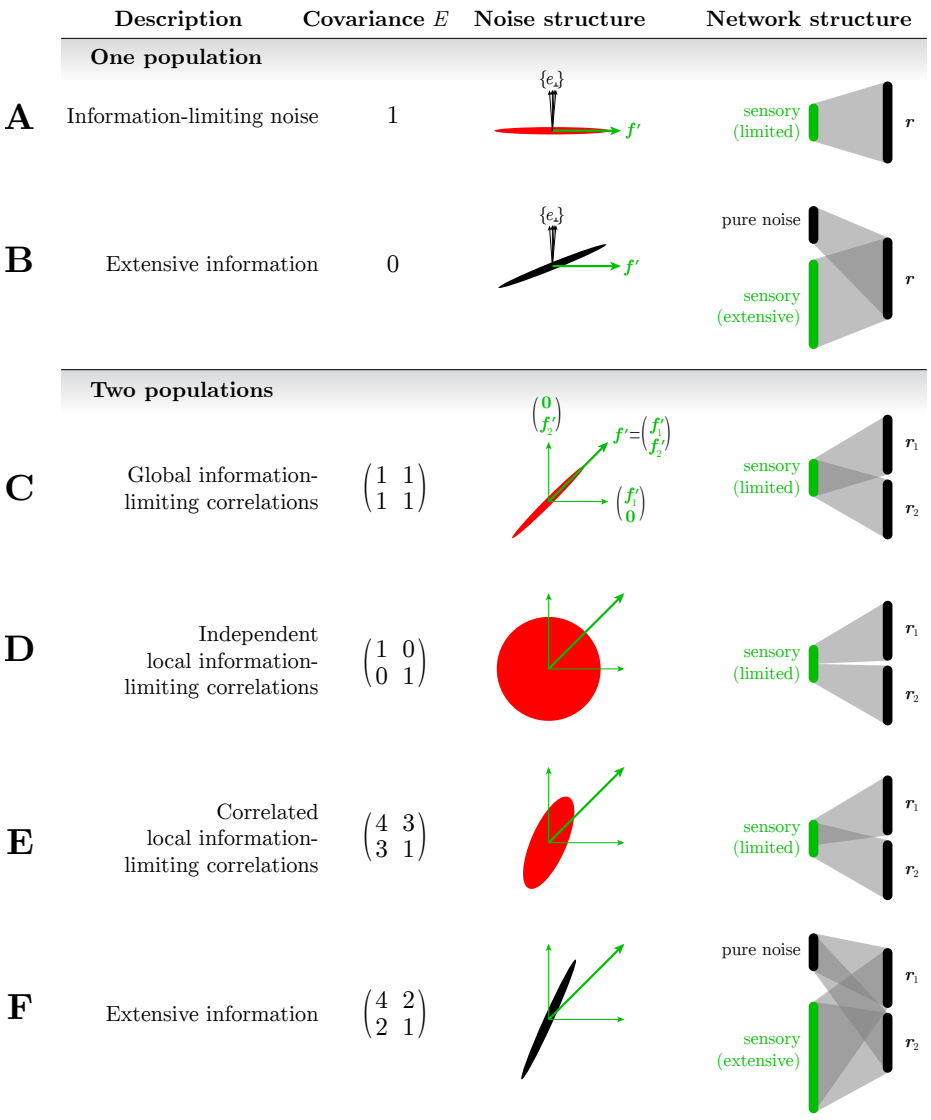


Figure 4.5 : Properties of local and global information-limiting correlations in one population and two populations. The four columns describe different situations, example information-limiting scale matrices E, visualizations of corresponding covariance ellipses (red) relative to the signal direction(s) $\boldsymbol{f}'$ (green), and a schematic of a feedforward network architecture that induces this noise structure in the downstream neurons.

A: In **one population** [1], information-limiting noise varies along the $\boldsymbol{f}'$ direction, while the variability in orthogonal directions ($\{e_\perp\}$, black axes) may be narrow or wide.

B: If variability along the signal direction is narrow (black), then the information grows extensively with population size. This fact can be seen from Eq 2.79 in Section 2.5.6: $J = \frac{1}{1/J_0+\epsilon}$. When the variability of the global information-limiting correlations

ϵ is negligible, the Fisher Information will be dominated by J_0, which grows extensively with population size.
C: With **two populations**, each population has its own local signal direction, $\boldsymbol{f}_1'$ and $\boldsymbol{f}_2'$; the global signal lies in the direction $(\boldsymbol{f}_1', \boldsymbol{f}_2')$. Noise in $(\boldsymbol{f}_1', \boldsymbol{f}_2')$ direction (red ellipse) is global information-limiting noise, and it can arise when both populations receive input from the same pool of sensory neurons.
D: Two populations receive input from two completely distinct sensory pools, making their local information-limiting noise to be independent.
E: If the neurons receive some input from distinct sensory pools, then their local information-limiting noise is correlated but not identical.
F: Local information-limiting noise is not globally limiting (black ellipse). Such noise must arise from a low-dimensional *non-sensory* source because it does not fluctuate in the same pattern as the signal itself.

4.4.3 Redundant linear codes with multiple populations

More generally, we can decompose the covariance of the entire neural population as

$$\mathrm{Cov}(\boldsymbol{r}|s) = \Sigma = \Sigma_0 + UEU^\top \tag{4.4}$$

where Σ_0 is the information-extensive part and $UEU^\top$ contains both all *local* information-limiting components, $\epsilon_{zz}\boldsymbol{f}'_z\boldsymbol{f}'^\top_z$, and components from their *co*variances, $\epsilon_{zz'}\boldsymbol{f}'_z\boldsymbol{f}'^\top_{z'}$. Specifically, U is a block diagonal matrix whose diagonal blocks contain the neural populations' sensitivities $\boldsymbol{f}'_z = \partial_s \langle \boldsymbol{r}_z|s\rangle$,

$$U = \begin{pmatrix} \boldsymbol{f}'_1 & & \\ & \ddots & \\ & & \boldsymbol{f}'_{N_z} \end{pmatrix} \tag{4.5}$$

E is an $N_z \times N_z$ matrix reflecting the covariance of the local information-limiting correlations (we name it the 'information-limiting scale matrix'),

$$E = \begin{pmatrix} \epsilon_{11} & \cdots & \epsilon_{1N_z} \\ \vdots & \epsilon_{zz'} & \vdots \\ \epsilon_{N_z1} & \cdots & \epsilon_{N_zN_z} \end{pmatrix} \tag{4.6}$$

In Figure 4.5, we give the examples showing the properties of local and global information-limiting correlations in one population and two populations. The four columns describe different situations, example information-limiting scale matrices E, visualizations of corresponding covariance ellipses (red) relative to the signal direction(s) $\boldsymbol{f}'$ (green), and a schematic of a feedforward network architecture that induces this noise structure in the downstream neurons.

In one population [1], information-limiting noise varies along the $\boldsymbol{f}'$ direction, while the variability in orthogonal directions ($\{e_\perp\}$, black axes) may be narrow or

wide (Figure 4.5A). If variability along the signal direction is narrow (black), the information grows extensively with population size (Figure 4.5B). This fact can be seen from Eq 2.79 in Section 2.5.6: $J = \frac{1}{1/J_0+\epsilon}$. When the variability of the global information-limiting correlations ϵ is negligible, the Fisher Information will be dominated by J_0, which grows extensively with population size.

In two populations, each population has its own local signal direction, $\boldsymbol{f}_1'$ and $\boldsymbol{f}_2'$; the global signal lies in the direction $(\boldsymbol{f}_1', \boldsymbol{f}_2')$. Noise in that direction (red ellipse) is global information-limiting noise, and it can arise when both neurons receive input from the same pool of sensory neurons (Figure 4.5C). Figure 4.5D illustrates the two populations receive input from two completely distinct sensory pools, thus their local information-limiting noise is independent. Figure 4.5E shows a scenario in which some (not *all*) neurons receive input from distinct sensory pools, so their local information-limiting noise is correlated but not identical. In these two examples, the linear span defined by $\{(\alpha_1\boldsymbol{f}_1', \alpha_2\boldsymbol{f}_2'), (\beta_1\boldsymbol{f}_1', \beta_2\boldsymbol{f}_2')\}$ contains global information-limiting noise along $(\boldsymbol{f}_1', \boldsymbol{f}_2')$, so the total information is still limited.

Figure 4.5E shows the situation where the local information-limiting noise is not globally limiting (black ellipse). Such noise must arise from a low-dimensional *non-sensory* source because it does not fluctuate in the same pattern as the signal itself. The last example (Figure 4.5F) is not biological plausible because the sensory population is not likely to contain extensive information and the cortex neural population has no reason to inherit signal and noise separately from two upstream sources.

Note that this decomposition of total noise covariance into a locally information-limiting part and the rest depends on a sensible partitioning of population $\boldsymbol{r}$ into subsets $\boldsymbol{r}_z$. This may be based on having distinct neural populations, for example in different brain areas. Because neural populations in different brain areas might

inherit distinct information from different sensory inputs, we can then partition them into different subsets.

In the following derivation, we will compute the total information in the entire population when the information-limiting scale matrix E is full rank (Figure 4.5D,E). We will show that the total information will be determined by the inverse of E when each population size is large.

As shown in Section 2.5.1.1, the total Fisher information can be written as $J = \boldsymbol{f}'^{\top}\Sigma^{-1}\boldsymbol{f}'$ (Eq 2.22). Since $U\mathbf{1} = \boldsymbol{f}'$ where $\mathbf{1}$ is a vector of all 1's, we can use Equation 4.4 and the above definitions to rewrite the Fisher Information as

$$J = \mathbf{1}^{\top}U^{\top}\Sigma^{-1}U\mathbf{1} = \mathbf{1}^{\top}J_M\mathbf{1} \tag{4.7}$$

Here the Information matrix is defined as $J_M = U^{\top}\Sigma^{-1}U$ and reflects the Fisher Information content in each neural population, as well its correlations across those populations.[iii]

Applying the Woodbury lemma, we invert Σ as

$$\Sigma^{-1} = \Sigma_0^{-1} - \Sigma_0^{-1}U^{\top}(E^{-1} + U\Sigma_0^{-1}U^{\top})^{-1}U\Sigma_0^{-1} \tag{4.8}$$

[iii] Note that this is not the usual Fisher Information Matrix, which reflects the correlated Fisher Information about multiple distinct variables,

$$J_{zz'} = \left\langle \partial_{s_z} \log p(\boldsymbol{r}|\boldsymbol{s})\, \partial_{s_{z'}} \log p(\boldsymbol{r}|\boldsymbol{s}) \right\rangle_{p(\boldsymbol{r}|\boldsymbol{s})}$$

Instead, here this is the information content that multiple measurements have about a single variable,

$$J_{Mzz'} = \langle \partial_s \log p(\boldsymbol{r}_z|s)\, \partial_s \log p(\boldsymbol{r}_{z'}|s) \rangle_{p(\boldsymbol{r}|s)}$$

The difference is where the subscripts z, z' are placed: in the multivariate stimulus $\boldsymbol{s}$ or in the informative signals $\boldsymbol{r}$. For exponential family distributions with linear sufficient statistics, $J_{Mzz'} = \boldsymbol{f}'_z\Sigma^{-1}\boldsymbol{f}'_{z'}$.

We can then express the Information matrix as

$$J_M = U\Sigma^{-1}U^\top \tag{4.9}$$

$$= U\Sigma_0^{-1}U^\top - U\Sigma_0^{-1}U^\top(E^{-1} + U\Sigma_0^{-1}U^\top)^{-1}U\Sigma_0^{-1}U^\top \tag{4.10}$$

Defining $J_{M0} = U\Sigma_0^{-1}U^\top$ to be the information matrix arising from the information-extensive part of the covariance, we obtain

$$J_M = J_{M0} - J_{M0}(E^{-1} + J_{M0})^{-1}J_{M0} \tag{4.11}$$

$$= J_{M0}(I - (E^{-1} + J_{M0})^{-1}J_{M0}) \tag{4.12}$$

$$= J_{M0}((E^{-1} + J_{M0})^{-1}(E^{-1} + J_{M0}) - (E^{-1} + J_{M0})^{-1}J_{M0}) \tag{4.13}$$

$$= J_{M0}(E^{-1} + J_{M0})^{-1}E^{-1} \tag{4.14}$$

$$= (J_{M0}^{-1} + E)^{-1} \tag{4.15}$$

In the limit of extensive population size, elements in J_{M0} will also be extensive [5, 6]. Meanwhile E is finite. Thus E will be much larger than J_{M0}^{-1} along all dimensions,

$$E \gg J_{M0}^{-1} \tag{4.16}$$

Then we can have $J_M \approx E^{-1}$. The total information can be approximated as

$$J \approx \mathbf{1}^\top E^{-1}\mathbf{1} \tag{4.17}$$

From the above derivation, we can see that the total information can still be limited when there are local information-limiting correlations $UEU^\top$ decomposed in the covariance of the entire population. This is because at least *some* global information-limiting correlations is generically contained in $UEU^\top$.

4.4.4 Redundant codes with multiple nonlinear statistics

Information-limiting correlations were originally defined in linear population codes [1]. In Section 2.2.4, we extended this concept to nonlinear population codes. For

nonlinear population codes, the covariance of the sufficient statistics can be locally decomposed into an extensive information part Γ_0 and a global information-limiting part $\epsilon \boldsymbol{F}' \boldsymbol{F}'^{\top}$. The global information-limiting noise bounds the variance of unbiased estimate $\sigma_{\hat{s}}^2$ to no smaller than ϵ even with optimal decoding.

In Section 4.4.3, we described redundant linear codes in multiple neural populations. Here we extend this idea to the nonlinear population codes, describing redundant codes with multiple nonlinear statistics. This can be realized by decomposing the covariance of the nonlinear statistics as

$$\mathrm{Cov}(\boldsymbol{R}(\boldsymbol{r})|s) = \Gamma = \Gamma_0 + UEU^{\top} \tag{4.18}$$

where Γ_0 is the extensive information part and $UEU^{\top}$ contains both local information-limiting components, $\epsilon_{zz} \boldsymbol{F}'_z \boldsymbol{F}'^{\top}_z$, and components from their *co*variances, $\epsilon_{zz'} \boldsymbol{F}'_z \boldsymbol{F}'^{\top}_{z'}$. Specifically, U is a block diagonal matrix whose diagonal blocks contain the nonlinear basis' sensitivities $\boldsymbol{F}'_z = \partial_s \langle \boldsymbol{R}_z(\boldsymbol{r})|s \rangle$,

$$U = \begin{pmatrix} \boldsymbol{F}'_1 & & \\ & \ddots & \\ & & \boldsymbol{F}'_{N_z} \end{pmatrix} \tag{4.19}$$

E is the information-limiting scale matrix reflecting the covariance of information-limiting noise in each statistic,

$$E = \begin{pmatrix} \epsilon_{11} & \cdots & \epsilon_{1N_z} \\ \vdots & \epsilon_{zz'} & \vdots \\ \epsilon_{N_z 1} & \cdots & \epsilon_{N_z N_z} \end{pmatrix} \tag{4.20}$$

The expressions here for the redundant codes for multiple nonlinear statistics is very similar to the expressions for the the redundant linear codes for multiple populations. The redundant codes for multiple nonlinear statistics is a more general case because

we can consider the multiple subsets of nonlinear statistics to serve as linear statistics in multiple derived populations. Then everything will be reduced to the scenario described in Section 4.4.3.

Similar to the derivation from Eq 4.7 to Eq 4.17 in Section 4.4.3, we can approximate the total information in the redundant codes of different statistics as $J = \mathbf{1}^\top E^{-1}\mathbf{1}$ when $E \gg J_{M0}^{-1}$, where $J_{M0} = U\Gamma_0^{-1}U^\top$.

The decomposition of the noise covariance into an information-limiting part and an extensive part is generic when there is an expansion from the sensory periphery to the cortex. On the other hand, a useful decomposition of the noise covariance into locally information-limiting parts depends on a sensible partitioning of $\boldsymbol{R}(\boldsymbol{r})$ into subsets $\boldsymbol{R}_z(\boldsymbol{r})$. This is natural when the populations of neurons are from different brain areas, and may be helpful when neuronal feature selectivities are clustered. Future work will focus on good ways of partitioning the cortical populations, perhaps by appealing to the statistical structure of task variables. The present work focuses on how to evaluate the brain's decoding for a given partition.

4.5 Coarse-grained description of brain's neural transformation

In the extended redundant codes described in Section 4.4.3 and 4.4.4, the information in each subset of statistics $\boldsymbol{R}_z$ is limited. This ensures that many local decoders $\boldsymbol{W}_z$ give the same output $\hat{\boldsymbol{s}}_z$ up to the intrinsic uncertainty of the code, which place these fine-grained decoders in the same equivalence class for that subset of statistics. This suggests that we might safely reduce the dimensionality of each subset of the redundant statistics, extracting only their estimate of the stimulus. Then we could

describe the brain's computations by how it combines many such partial estimates $\hat{\boldsymbol{s}} = (\hat{s}_1, \ldots, \hat{s}_{N_z})$, into a scalar final estimate $\hat{s}$.

This is equivalent to the following distributed two-step decoding scheme to describe the brain's neural transformation:

$$s \xrightarrow[\text{code}]{\text{redundant}} \boldsymbol{r} \xrightarrow{\text{nonlinearity}} \boldsymbol{R}(\boldsymbol{r}) \xrightarrow{\text{grouping}} \{\boldsymbol{R}_z\} \xrightarrow[\text{decoding}]{\text{fine-grained}} \{\hat{s}_z\} \xrightarrow[\text{decoding}]{\text{coarse-grained}} \hat{s} \tag{4.21}$$

This scheme can give us a useful description of the neural transformations when the recorded neurons can be described as a redundant code, with redundancy within and across multiple populations and multiple subsets of statistics.

To describe these neural transformations, we first compute a basis for possible nonlinear transformations using functions $\boldsymbol{R}$ based on the recorded neurons. For *optimal* decoding, this nonlinear basis should span the sufficient statistics of the encoding model. However, if the brain is not optimal, for accurate reconstruction of the actual suboptimal neural decoder, the brain's neural transformation should lie within the span of the chosen nonlinear basis. Ideally, we would like to preserve all the information contained in the recorded neuronal responses $\boldsymbol{r}$ so that we can find out how the brain's decoder uses them. Then we partition the nonlinear basis into different subsets $\boldsymbol{R}_z(\boldsymbol{r})$. As discussed in Section 4.4.3 and Section 4.4.4, natural groupings should be based on prior knowledge both about the brain's anatomy (e.g. distinct populations) and about the neural encoding model (which determines which statistics of neural responses are tuned to the stimulus and which subset is information-limited). With a sensible partitioning that matches the encoding model, many elements of each subset will contain same information because they inherit it from the same upstream information source. We then decode each subset $\boldsymbol{R}_z(\boldsymbol{r})$ separately to generate unbiased estimates $\hat{s}_z$ that each contain part of the information in the full population. As in Section 4.3, in this step we are directly processing the many

statistics of neural responses (*e.g.* $\{r_{zj}r_{zj'}\}_{j,j' \in N_z}$ for the N_z neurons in population z), so we name this step 'fine-grained decoding'. With this fine-grained decoding step, we desire to squeeze out the redundancy while still preserving all of the information about the relevant stimulus in these unbiased partial estimates. As we will see below in Sections 4.5.1, this places some restrictions on the correlations between these partial estimates. We will then combine these partial estimates $\hat{\boldsymbol{s}}$ into one global estimate $\hat{s}$. In this *coarse-grained decoding* step, instead of dealing with large number of statistics $\boldsymbol{R}$ and corresponding weights $\boldsymbol{W}_z$, we will describe a smaller number of weights a_z, just one for each $\hat{s}_z$. This enables us to express putative perceptual estimates more simply as $\hat{s} \approx \boldsymbol{a}^\top \hat{\boldsymbol{s}}$, even when the transformation is implemented in the brain by a potentially intricate set of neural transformations $\hat{s} = \psi(\boldsymbol{r})$ on all sensory responses.

In the following sections, I will illustrate the two-step decoding scheme in detail. I will first illustrate the fine-grained decoding and specify the condition under which the fine-grained decoding step is a sufficient dimensionality reduction [76]. Second, I will illustrate how to infer the coarse-grained decoding weights using linear regression between behavioral choices and the chosen dimensionality reduction. Third, I will show results of a simulation that demonstrate the applicability of the two-step decoding scheme. Fourth, I will give an example showing the two-step decoding scheme is not guaranteed to be valid when the chosen grouping of neurons is not matched to the encoding model. Last, I will describe general conditions when the two-step decoding scheme is suboptimal under a non-matched grouping. These sections illustrate the value and limitations of the coarse-grained description of neural computation.

4.5.1 Fine-grained decoding of subsets of statistics

When the size of a cortical population is much larger than the number of its sensory inputs, the neural population can be described as a redundant code with redundancy in multiple populations and multiple subsets of statistics. The covariance of the recorded neural population's nonlinear statistics can then be decomposed as Eq 4.18: $\mathrm{Cov}(\boldsymbol{R}(\boldsymbol{r})|s) = \Gamma = \Gamma_0 + UEU^\top$. We assume the grouping in the two-step decoding matches the grouping in the redundant neural encoding model. Specifically, each subset collects the neural statistics that contains local information-limiting noise with the same information-limiting variance.[iv] Then in the fine-grained decoding step, we decode each subset $\boldsymbol{R}_z(\boldsymbol{r})$ separately to generate unbiased estimates $\hat{s}_z$ that each contain part of the information in the full population:

$$\hat{s}_z = \boldsymbol{w}_z^\top \boldsymbol{R}_z(\boldsymbol{r}) + c_z \tag{4.22}$$

where $\boldsymbol{w}_z$ is the corresponding vector of decoding weights. We construct a vector of these partial estimates, including all the estimates constructed from the subsets, $\hat{\boldsymbol{s}} = [\hat{s}_1, \cdots, \hat{s}_{N_z}]^\top$. This fine-grained decoding is a way of doing dimensionality reduction [84–87]. The dimensionality reduction is said to be sufficient [76] when the information in the partial estimate vector, $J_{\hat{\boldsymbol{s}}}$, is the same as the total information in the neural population, J:

$$J_{\hat{\boldsymbol{s}}} = J \tag{4.23}$$

[iv] In principle, this will be difficult to infer from data since the information-limiting noise may give only a small additional variance for each neuron, even when it has a large collective effect. However, here we make the strong assumption that we know the encoding model and can group these responses appropriately. Future work will aim to find good ways to group these responses based only on experimental data.

Assuming the distributions of the unbiased estimates are still in the exponential family, we can use the results in Eq 4.7 to compute $J_{\hat{\boldsymbol{s}}}$ as

$$J_{\hat{\boldsymbol{s}}} = \boldsymbol{F}_{\hat{\boldsymbol{s}}}'^{\top} \mathrm{Cov}(\hat{\boldsymbol{s}}|s)^{-1} \boldsymbol{F}_{\hat{\boldsymbol{s}}}' \tag{4.24}$$

where $\boldsymbol{F}_{\hat{\boldsymbol{s}}}' = \partial \langle \hat{\boldsymbol{s}}|s \rangle / \partial s$ are the partial estimates' sensitivities. Because these partial estimates $\hat{s}_z$ are unbiased by construction, the sensitivities of each estimate to changes in the stimulus will all be ones, $\boldsymbol{F}_{\hat{\boldsymbol{s}}}' = (1, \cdots, 1) = \mathbf{1}$. Thus the information in the partial estimate vector can be expressed as

$$J_{\hat{\boldsymbol{s}}} = \mathbf{1}^{\top} \mathrm{Cov}(\hat{\boldsymbol{s}}|s)^{-1} \mathbf{1} \tag{4.25}$$

We desire the dimensionality reduction from nonlinear statistics $\boldsymbol{R}$ to the partial estimate vector $\hat{\boldsymbol{s}}$ to be sufficient, because then we can study the brain's computation based on the partial estimate vector, instead of dealing with an extensive number of nonlinear statistics of responses.

As illustrated in Section 4.4.3 and 4.4.4, in the limit of having a large number of nonlinear statistics that are tuned to the stimulus based on a smaller number of sensory inputs, the total information of a extended redundant code is approximately equal to $J = \mathbf{1}^{\top} E^{-1} \mathbf{1}$ (Eq 4.17). Comparing it with Eq 4.25, we can rewrite the condition for sufficiency of the fine-grained decoding step as

$$\mathrm{Cov}(\hat{\boldsymbol{s}}|s) = E \tag{4.26}$$

The optimality condition in Eq 4.26 should require two things. One is the grouping matches the one in the redundant neural encoding model (Eq 4.18). Specifically, each subset collects the neural statistics that contains local information-limiting noise with the same information-limiting variance. In Section 4.5.4, we will show an example where non-matched grouping shall cause the two-step decoding scheme to be invalid.

Another requirement is that the fine-grained decoding weights $\boldsymbol{w}_z$ cannot be too suboptimal: they should reduce the noise $\boldsymbol{w}_z^\top \Gamma_{0zz} \boldsymbol{w}_z$ to a level below the variance of local information-limiting noise ϵ_{zz}, since otherwise the partial estimate vector will not extract most of the available information. For a highly redundant code, this is not a difficult criterion to meet, so we assume the brain will do so.

4.5.2 Coarse-grained decoding to reveal the essence of neural computation

In the coarse-grained decoding step, the behavioral estimate $\hat{s}$ can be expressed as a linear combination of the partial estimates,

$$\hat{s} = \boldsymbol{a}^\top \hat{\boldsymbol{s}} \tag{4.27}$$

where $\boldsymbol{a}$ is a vector of scaling factors. The scaling factors can tell us how the brain uses the information in different subsets of $\boldsymbol{R}_z$. After the dimensionality reduction accomplished in the fine-grained decoding step, there will be no degenerate description for the scaling factors: Different scaling factors will create differentiable decoders, making their neural transformations not in the same equivalence class.

This suggests that we should focus on inferring the coarse-grained decoding scaling factors in the interest of better understanding the actual neural transformation used by the brain. In fact, we can infer the scaling factors by regressing the partial estimate vectors $\hat{\boldsymbol{s}}$ against the actual estimate of the brain. When the optimality condition of the fine-grained decoding (Eq 4.26) is satisfied, the regression error will be small. The inferred neural transformation based on this two-step decoding scheme should be in the same equivalence class as the actual neural transformation.

The optimal scaling factor vector can be computed by decoding the partial esti-

mate vector optimally,

$$\boldsymbol{a}_{\text{opt}} = \frac{\text{Cov}(\hat{\boldsymbol{s}}|s)^{-1}\boldsymbol{F}'^{\top}_{\hat{\boldsymbol{s}}}}{\boldsymbol{F}'_{\hat{\boldsymbol{s}}}\text{Cov}(\hat{\boldsymbol{s}}|s)^{-1}\boldsymbol{F}'^{\top}_{\hat{\boldsymbol{s}}}} \tag{4.28}$$

where $\boldsymbol{F}'_{\hat{\boldsymbol{s}}} = \partial \langle \hat{\boldsymbol{s}} \rangle / \partial s$ is the sensitivity of the partial estimate vector. Because these partial estimates $\hat{s}_z$ are unbiased, the sensitivities of each estimate to changes in the stimulus will all be ones, $\boldsymbol{F}'_{\hat{\boldsymbol{s}}} = (1, \cdots, 1) = \mathbf{1}$. When the optimality condition of the fine-grained decoding (Eq 4.26) satisfies, the optimal scaling factor vector can be expressed as

$$\boldsymbol{a}_{\text{opt}} = \frac{E^{-1}\mathbf{1}^{\top}}{J} \tag{4.29}$$

where J is the total information in the extended redundant code (Eq 4.17).

The inferred scaling factors $\boldsymbol{a}$ from the regression can then be compared with the optimal scaling factor $\boldsymbol{a}_{\text{opt}}$. From this comparison, we can discover how the brain combines the information from different subsets of the nonlinear statistics when it's performing the task.

Another way to check the efficiency of the brain's coarse-grained computation is using the nonlinear choice correlation test proposed in Chapter 3. In the fine-grained decoding step, the optimality condition (Eq 4.26) guarantees that the partial estimate vector preserve all the information in the entire neural population. Thus, we can use the proportionality between nonlinear choice correlation and those predicted from optimality within one subset to check how the brain's decoder use the information in that subset. The following derivation will reveal the dependence between the proportionality $\boldsymbol{\alpha}$ and the inferred scaling factors $\boldsymbol{a}$.

The choice correlation between the brain's actual estimate and the nonlinear basis

unit R_{zk} (the kth element of subset z) can be expressed as

$$C_{zk} = \text{Corr}(\hat{s}, R_{zk}) = \frac{(\Gamma \boldsymbol{W})_{zk}}{\sqrt{\Gamma_{zkzk} \boldsymbol{W}^T \Gamma \boldsymbol{W}}} \tag{4.30}$$

where $\boldsymbol{W}$ are the equivalent weights that formulate brain's actual estimate based on the nonlinear basis, Γ is the covariance of the nonlinear basis and Γ_{zkzk} is the variance of R_{zk}.

Combining the expression from the two-step decoding scheme(Eq 4.22 and 4.27), the overall weights $\boldsymbol{W}$ can be wrote as

$$\boldsymbol{W} = [a_1 \boldsymbol{w}_1^\top, \cdots, a_{N_z} \boldsymbol{w}_{N_z}^\top] \tag{4.31}$$

Substituting this expression into 4.30 and assuming the covariance Γ is dominated by information limiting part, $\Gamma \approx UEU^T$, we can approximate the choice correlation as

$$C_{zk} \approx \frac{(UEU^T \boldsymbol{W})_{zk}}{\sqrt{\Gamma_{zkzk} \boldsymbol{W}^T UEU^T \boldsymbol{W}}} \tag{4.32}$$

$$= \frac{(UE\boldsymbol{a}^T)_{zk}}{\sqrt{\Gamma_{zkzk} \boldsymbol{a} E \boldsymbol{a}^T}} \tag{4.33}$$

$$\approx \frac{(E\boldsymbol{a}^T)_i}{\boldsymbol{a}E\boldsymbol{a}^\top} \sqrt{\frac{J_{zk}}{J}} \tag{4.34}$$

$$= \alpha_z \sqrt{\frac{J_{zk}}{J}} \tag{4.35}$$

where $J_{zk} = \frac{F'_{ik}}{\sqrt{\Gamma_{zkzk}}}$ is the Fisher Information contained in $R_{zk}(\boldsymbol{r})$, and $J = \frac{1}{\boldsymbol{W}\Gamma\boldsymbol{W}} \approx \frac{1}{\boldsymbol{a}E\boldsymbol{a}^\top}$ is the information that brain's decoder has extracted. Here we used the following result, $\boldsymbol{W}^\top U = \boldsymbol{a}$, which is because $\boldsymbol{w}_z^\top \boldsymbol{F}'_z = 1$ (the fine-grained weights are used to create unbiased estimates).

Equation 4.35 predicts that, for extended redundant codes, the nonlinear choice correlations should be proportional to the information content of each statistic R_{zk}, with proportionality $\boldsymbol{\alpha}$ that depends on the scaling factor $\boldsymbol{a}$ and E. Substituting the expression for the optimal scaling factor (Eq 4.29) into the expression for the

proportionality, we can get $\boldsymbol{\alpha} = \mathbf{1}$. Thus when the proportionality are all equal to one, the scaling factor is optimal, otherwise it's suboptimal. This conclusion matches the one we derived in Section 3.2.4 (Eq 3.7), where we use the proportionality to indicate the efficiency of the brain's total decoding efficiency. Here the optimality condition of the fine-grained decoding allows us to use the proportionality $\boldsymbol{\alpha}$ to evaluate the optimality of the coarse-grained decoding, which reveals the essence of the brain's computation.

4.5.3 Applicability of distributed two-step decoding scheme in a cubic redundant code

In Figure 4.6A, we illustrate the distributed two-step decoding scheme as an efficient way to analyze the essential property of the brain's actual neural computation. We use the following simulation to demonstrate that applicability of this scheme. First, neural responses are simulated using the extended redundant cubic codes (Section 4.7.2.3). We assume the brain used the two-step decoding scheme as the actual neural transformation to estimate the stimulus with optimal fine-grained decoding and suboptimal coarse-grained decoding with the scaling factor $\boldsymbol{a}_{\text{sub}} = [0.2105, 1.5789, -0.7895]$. To analyze the brain's actual neural transformation, we apply the distributed two-step decoding scheme here. In the distributed two-step decoding scheme, we first compute the polynomial basis up third-order: $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j, r_i r_j r_k\}$. Then we group the basis elements according to their order, and create one unbiased estimate $\hat{s}_z$ from each subset. Since elements of these subsets are redundant, we can decode near-optimally even with only a subset of size M^z chosen randomly from the statistics $\boldsymbol{R}_z$ (Figure 4.4). This can be confirmed by finding that the measured covariance of the partial estimate vector, $\text{Cov}(\hat{\boldsymbol{s}}|s)$, is approximately equal to the information-limiting scale

matrix, E. In the simulations, the information-limiting scale matrix is set to be

$$E = \begin{pmatrix} 1 & -0.45 & 0.37 \\ -0.45 & 1 & -0.39 \\ 0.37 & -0.39 & 1 \end{pmatrix} \tag{4.36}$$

The measured covariance of the partial estimate vector is

$$\mathrm{Cov}(\hat{\boldsymbol{s}}|s) = \begin{pmatrix} 0.02 & 0.01 & -0.02 \\ 0.01 & 0.05 & 0.09 \\ -0.02 & 0.09 & -0.05 \end{pmatrix} + E \approx E \tag{4.37}$$

This shows that the optimality condition of the fine-grained decoding is satisfied (Eq 4.26), revealing the partial estimate vector has extracted all the information in the recorded neurons.

In the coarse-grained decoding step, we infer the scaling factors $\boldsymbol{a}$ by regressing the partial estimate vector $\hat{\boldsymbol{s}}$ against the true estimate $\hat{s}$. In the simulations, the inferred scaling factors are $\boldsymbol{a} = [0.2140 \pm 0.0091, 1.5630 \pm 0.0456, -0.9423 \pm 0.1016]$. In Figure 4.6C, we plotted this inferred scaling factors against the brain's true scaling factors $\boldsymbol{a}_{\mathrm{sub}}$ and found that they matched with each other, suggesting that we have successfully infer the brain's neural transformation by using the distributed two-step decoding scheme.

In Figure 4.6B, we plot the measured nonlinear choice correlation against the predicted nonlinear choice correlation, with different color denoting the choice correlation for different subsets. Within each subset, the measured choice correlation is proportional to the optimal prediction, $\sqrt{J_{zk}/J}$, which is consistent with locally optimal decoding of each subset. However, the Figure also clearly shows that the proportionalities $\boldsymbol{\alpha}_z$ for subset z are not equal to one, and differ between different subsets. This demonstrates that the brain's decoder does not optimally combine the

partial estimates from different subsets, which matches the properties of our simulated model for the brain's decoder.

This simulation verifies the applicability of the two-step decoding scheme. More details about the simulation can be found in Section 4.7.3.

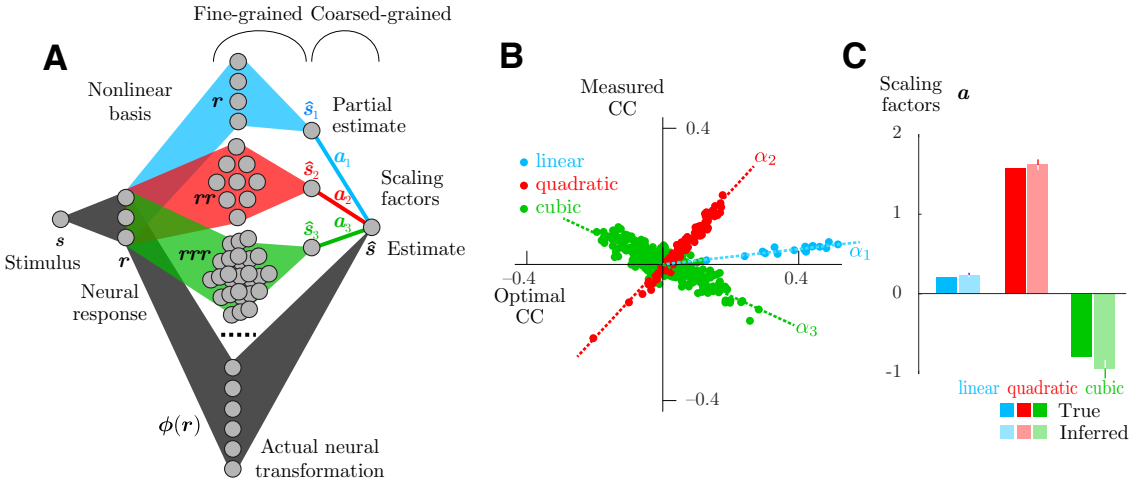


Figure 4.6 : **A**: In a redundant code, the actual neural transformations (grey) can be enacted in other equivalent ways (color). Since many fine details do not matter, it is valuable to characterize these equivalent nonlinearities by the scaling factors $\boldsymbol{a}$ of different subset of the nonlinear basis (*e.g.*, polynomials). **B**: Simulations for this scheme show that, as predicted by Eq. 4.35, different subsets produce different slopes when plotting measured versus optimal nonlinear choice correlations. **C**: Inferred scaling factors of nonlinear basis match the true scaling factors used by the brain, demonstrating the applicability of the two-step decoding scheme.

4.5.4 The optimality of two-step decoding scheme under non-matched grouping

In the distributed two-step decoding scheme, we make a big assumption that the grouping of the nonlinear basis matches the grouping in the neural encoding model. Here we will analyze the optimality of the two-step decoding scheme in an example where the grouping is not matched.

We consider a task which stimulates multiple sensory populations and requires the brain to integrate the information from these sensory inputs. Example for such a task can be auditory-visual spatial localization [88–93]: The brain needs to report the location of an object, and there will be visual and auditory cues about the location.

Imaging we have recordings of two cortex neural populations, $\boldsymbol{r}_z$, $z = 1, 2$, each inherited distinct information about the variable of interest from different sensory populations. Each population, $\boldsymbol{r}_z$ has linear and quadratic statistics tuned to the variable of interest, and these statistics completely capture all information about the signal.

In Section 4.7.2.2, we show that a redundant code with redundancy in linear and quadratic statistics can be created by embedding the information-limiting noise $p(d\boldsymbol{s}_z) = p([ds_{z1}, ds_{z2}]) = \mathcal{N}(0, E_z)$ in linear and quadratic statistics of population z. The encoding probability is then

$$p(\boldsymbol{r}_z|s, d\boldsymbol{s}_z) = N(\boldsymbol{f}_z(s + ds_{z1}), \Sigma_z(s + ds_{z2})) \tag{4.38}$$

where the covariance of the embedded noise is

$$E_z = \begin{pmatrix} \epsilon_{zxx} & \epsilon_{zxy} \\ \epsilon_{zxy} & \epsilon_{zyy} \end{pmatrix}. \tag{4.39}$$

Here the embedded information-limiting noises in the two populations are independent to each other, thus the correlations between the embedded noise in different populations are zero:

$$\mathrm{Cov}([ds_{11}, ds_{12}, ds_{21}, ds_{22}]|s) =: E_{\mathrm{tot}} = \begin{pmatrix} E_1 & \mathbf{0} \\ \mathbf{0} & E_2 \end{pmatrix}. \tag{4.40}$$

This setting describes the circumstance where the two neural populations inherit distinct information from completely different sensory populations.

In such an extended redundant code, the total information can be approximated as

$$J = \mathbf{1}^\top E_{\text{tot}}^{-1}\mathbf{1} = \mathbf{1}^\top E_1^{-1}\mathbf{1} + \mathbf{1}^\top E_2^{-1}\mathbf{1} \tag{4.41}$$

This approximation achieves when there are extensive number of neurons in both populations (Eq 4.16).

In the following derivations, we will show that the two-step decoding scheme is valid when when the grouping matches the one in the encoding model. Otherwise, the two-step decoding scheme is not guaranteed to be valid.

If we have both the knowledge of the connectivity (the route by which the information is inherited from the sensory populations by the cortical populations) and knowledge of the encoding model of these neural populations (which nonlinear statistics are tuned and which subset is information-limited), then we will partition the statistics according to their orders and their regions (Figure 4.7A): $\boldsymbol{R}_{z1}(\boldsymbol{r}_z) = \boldsymbol{r}_z$ is the linear statistics of population z, $\boldsymbol{R}_{z2}(\boldsymbol{r}_z) = \text{vect}(\delta\boldsymbol{r}_z\delta\boldsymbol{r}_z^\top)$ is the quadratic statistics of population z, where $\delta\boldsymbol{r}_z = \boldsymbol{r}_z - \langle\boldsymbol{r}_z|s\rangle$. Here we centralized $\boldsymbol{r}_z$ to get rid of the linear information when we compute the quadratic statistics. This grouping matches the grouping in the encoding model. We can then formulate the partial estimate for each subset: $\hat{s}_{zj} = \boldsymbol{w}_{zj}\boldsymbol{R}_{zj} + c_{zj}$, where z denotes different populations and j denotes different statistics based on their orders. Assume these partial estimates extracted all the information in each subset, the optimality condition of the fine-grained decoding (Eq 4.26) will be satisfied: $\text{Cov}(\hat{\boldsymbol{s}}|s) = \text{Cov}([\hat{s}_{11}, \hat{s}_{12}, \hat{s}_{21}, \hat{s}_{22}]|s) = E_{\text{tot}}$. Thus the partial estimate vector preserves all the information in both populations, *i.e.*, the dimensionality reduction from $\boldsymbol{R}$ to $\hat{\boldsymbol{s}}$ is sufficient [76]. The distributed two-step decoding scheme is then valid given this partition.

However, if we lack the knowledge of the brain's connectivity, blindly using the

distributed two-step decoding scheme may be insufficient. The two-step decoding scheme under this scenario is illustrated in Figure 4.7B. We will first mistakenly mix the two populations, $\boldsymbol{r} = [\boldsymbol{r}_1, \boldsymbol{r}_2]$ and merely partition the nonlinear statistics according to the orders:

$$\boldsymbol{R}_{\cdot 1}(\boldsymbol{r}) = \boldsymbol{r} \tag{4.42}$$

$$\boldsymbol{R}_{\cdot 2}(\boldsymbol{r}) = \text{vect}(\delta\boldsymbol{r}\delta\boldsymbol{r}^\top) \tag{4.43}$$

$$\boldsymbol{R}(\boldsymbol{r}) = [\boldsymbol{R}_{\cdot 1}, \boldsymbol{R}_{\cdot 2}] \tag{4.44}$$

This grouping doesn't match the one in the encoding model. The optimality condition in Eq 4.26 is not satisfied.

Based on this not-matched grouping, we first do fine-grained decoding on $\boldsymbol{R}_{\cdot 1}$:

$$\hat{s}_{\cdot 1} = \boldsymbol{w}_{\cdot 1}\boldsymbol{R}_{\cdot 1} \tag{4.45}$$

When the partial estimate $\hat{s}_{\cdot 1}$ extracted all the information in the linear statistics of both populations, we can express it as the weighted sum of $\hat{s}_{11}$ and $\hat{s}_{21}$:

$$\hat{s}_{\cdot 1} = b_{11}\hat{s}_{11} + b_{21}\hat{s}_{21} \tag{4.46}$$

where $\hat{s}_{z1}$ is the partial estimate that has extracted the information in the linear statistics of population z. The optimal scaling is

$$b_{11} = \frac{\epsilon_{1xx}^{-1}}{\epsilon_{1xx}^{-1} + \epsilon_{2xx}^{-1}} \tag{4.47}$$

$$b_{21} = \frac{\epsilon_{2xx}^{-1}}{\epsilon_{1xx}^{-1} + \epsilon_{2xx}^{-1}} \tag{4.48}$$

where $1/\epsilon_{zxx}$ is the variance of $\hat{s}_{z1}$, $z = 1, 2$. Considering the statistics $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{21}$ are conditionally independent, decoding them separately and combining the decoders is equivalent to decoding all of them with one step.

Likewise, the partial estimate that has extract all the information in quadratic statistics $\boldsymbol{R}_{\cdot 2}$ is

$$\hat{s}_{\cdot 2} = \boldsymbol{w}_{\cdot 2}\boldsymbol{R}_{\cdot 2} = b_{12}\hat{s}_{12} + b_{22}\hat{s}_{22} \tag{4.49}$$

$$= \frac{\epsilon_{1yy}^{-1}}{\epsilon_{1yy}^{-1} + \epsilon_{2yy}^{-1}}\hat{s}_{12} + \frac{\epsilon_{2yy}^{-1}}{\epsilon_{1yy}^{-1} + \epsilon_{2yy}^{-1}}\hat{s}_{22} \tag{4.50}$$

Even though these two partial estimates extract all the information in their subset $\boldsymbol{R}_{\cdot i}$, the dimensionality reduction from $\boldsymbol{R}(\boldsymbol{r})$ to $[\hat{s}_{\cdot 1}, \hat{s}_{\cdot 2}]$ is still not guaranteed to be sufficient. This can be seen by computing the information contains in the partial estimate vector,

$$J_{\text{sub}} = \mathbf{1}^{\top} E_{\text{sub}}^{-1} \mathbf{1} \tag{4.51}$$

where

$$E_{\text{sub}} = \text{Cov}([\hat{s}_{\cdot 1}, \hat{s}_{\cdot 2}]|s) \tag{4.52}$$

$$= \begin{pmatrix} (\epsilon_{1xx}^{-1} + \epsilon_{2xx}^{-1})^{-1} & b_{11}b_{12}\epsilon_{1xy} + b_{21}b_{22}\epsilon_{2xy} \\ b_{11}b_{12}\epsilon_{1xy} + b_{21}b_{22}\epsilon_{2xy} & (\epsilon_{1yy}^{-1} + \epsilon_{2yy}^{-1})^{-1} \end{pmatrix} \tag{4.53}$$

Here we use the following results,

$$\text{Cov}(\hat{s}_{\cdot 1}, \hat{s}_{\cdot 2}|s) = b_{11}b_{12}\text{Cov}(\hat{s}_{11}, \hat{s}_{12}|s) + b_{21}b_{22}\text{Cov}(\hat{s}_{21}, \hat{s}_{22}|s) \tag{4.54}$$

$$= b_{11}b_{12}\epsilon_{1xy} + b_{21}b_{22}\epsilon_{2xy} \tag{4.55}$$

$$\text{Var}(\hat{s}_{\cdot 1}|s) = (\epsilon_{1xx}^{-1} + \epsilon_{2xx}^{-1})^{-1} \tag{4.56}$$

$$\text{Var}(\hat{s}_{\cdot 2}|s) = (\epsilon_{1yy}^{-1} + \epsilon_{2yy}^{-1})^{-1} \tag{4.57}$$

Comparing the information in the partial estimate vector $[\hat{s}_{\cdot 1}, \hat{s}_{\cdot 2}]$ and the total information given by the encoding model (Eq 4.41), we can find

$$J_{\text{sub}} \leq J \tag{4.58}$$

Equality is achieved when $\epsilon_{1xy} = \epsilon_{2xy}$.

To evaluate the suboptimality of the two-step decoding scheme based on this not-matched grouping, we set the diagonal elements in E_{tot} to be 1 and compute the ratio between the information difference and the total information given by the encoding model, $\frac{J-J_{\text{sub}}}{J}$, as a function of ϵ_{1xy} and ϵ_{2xy}. This ratio can be expressed as

$$\frac{J - J_{\text{sub}}}{J} = \left(\frac{\epsilon_{1xy} - \epsilon_{2xy}}{\epsilon_{1xy} + \epsilon_{2xy} + 2} \right)^2 \tag{4.59}$$

In Figure 4.7C, we plot this ratio as the function of ϵ_{1xy} and ϵ_{2xy}. We found that the suboptimality will become more severe when the difference between ϵ_{1xy} and ϵ_{2xy} is larger. The difference between ϵ_{1xy} and ϵ_{2xy} describes the degree of similarity between the underlying local information-limiting correlations structure of the two populations. When we formulate a subset by over-coarsely grouping the statistics of two very different neural populations, the corresponding partial estimate extracted from this subset is not sufficient to describe the relationship between the statistics derived from two very different neural populations. We will need more than one partial estimates for each subset.

With this example, we show that the two-step decoding is not guaranteed to be valid when the grouping is not matched. This is because the over-coarsely grouping will make the dimensionality reduction from the nonlinear basis of responses to the partial estimates to be insufficient. Then the inferred coarse-grained computation based on these partial estimates can't accurately capture the behavior of the full one-step, fine-grained decoder of a redundant nonlinear population code.

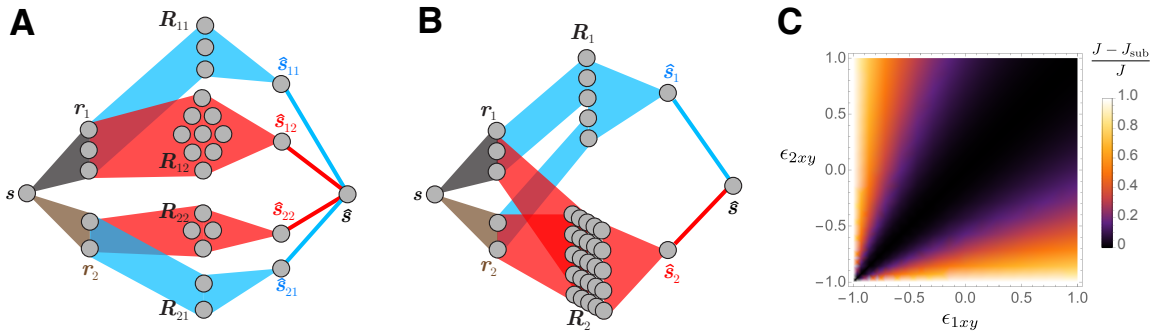


Figure 4.7 : The optimality of two-step decoding scheme when the grouping doesn't match the one in the redundant neural encoding model. The encoding model in the this example includes two neural populations $\boldsymbol{r}_1$ and $\boldsymbol{r}_2$. Each population contains distinct pieces of information about the stimulus. We assume they can all be described by redundant quadratic codes: the information about the stimulus is in the linear and quadratic sufficient statistics; the information is redundant within each subset of statistics where the grouping is according to the order. (**A**): We first compute the linear (blue) and quadratic (red) statistics for each population and group them according to regions and orders. This grouping matches the one in the encoding model. The optimality condition in Eq 4.26 can be satisfied, which makes the dimensionality reduction in the fine-grained decoding step to be sufficient. Thus the two-step decoding scheme is valid under this matched grouping. (**B**): We don't use the knowledge of brain's connectivity and mistakenly mixed the two populations. We compute linear(blue) and quadratic(red)) statistics and group them merely according to orders. This grouping is over-coarsely, which doesn't match the one in the encoding model. The information contained in the partial estimates is not guaranteed to be equal to the total information in the two populations (Eq 4.58). Thus the two-step decoding scheme is not guaranteed to be valid under this over-coarsely grouping. (**C**): We set the diagonal elements in E_{tot} to be 1 and plot the ratio between the information difference and the total information $\frac{J-J_{\text{sub}}}{J}$ as function of ϵ_{1xy} and ϵ_{2xy}. When $\epsilon_{1xy} = \epsilon_{2xy}$, the information difference is zero, making the two-step decoding scheme to be valid. When the value of ϵ_{1xy} and ϵ_{2xy} are more distinct, the information difference will become larger compared to the total information. The suboptimality of the two-step decoding scheme under the over-coarsely grouping becomes more severe.

4.5.5 When does the two-step decoding scheme fail?

In the previous section, we described an example where the two-step decoding scheme is suboptimal under a non-matched grouping. Here we describe the general conditions when the two-step decoding scheme is a suboptimal strategy, and the assumptions under which this coarse-graining fails to capture the structure of a suboptimal decoder. In brief, we decompose noise in each subset into a locally information-limiting part and other, removable correlations. The former is preserved after a fine-grained decoding step, while the latter is filtered out. Only if the removable noise components in one subset are correlated with the information-limiting components in another subset will the coarse-grained decoder lose crucial information that the brain might use.

Recall the structure of information-limiting noises, namely noise that is indistinguishable from a change in the signal. A crucial property of information-limiting noises (local and global) is their uniqueness: for any given population and any given signal, only noise directed along the vector $\boldsymbol{F}'_z$ will be relevant for optimal decoding near a reference stimulus [1]. We can therefore uniquely decompose the noise covariance into these two parts,

$$\Gamma_{zz} = \frac{\boldsymbol{F}'_z \boldsymbol{F}'^{\top}_z}{J_z} + \Gamma_{zz\perp} \tag{4.60}$$

The local information-limiting component is not generally orthogonal to the remainder; that *is* true however after whitening the noise according to $\Gamma_{zz}^{-1/2}\boldsymbol{R}_z$, where the information-limiting component has unit variance along the transformed signal direction, $\hat{\boldsymbol{e}}_{z0} = \Gamma_{zz}^{-1/2}\boldsymbol{F}'_z/\sqrt{J_z}$, generating covariance $\hat{\boldsymbol{e}}_{z0}\hat{\boldsymbol{e}}_{z0}^{\top} = \Gamma_{zz}^{-1/2}\boldsymbol{F}'_z\boldsymbol{F}'^{\top}_z\Gamma_{zz}^{-\top/2}/J_z$. The remaining whitened noise covariance is an isotropic $N_z - 1$-dimensional identity matrix in the orthogonal subspace, $I_{z\perp}$.

It can be convenient to express the noise in terms of the stochastic variables ds_z

and η_z,

$$\boldsymbol{R}_z = \frac{\boldsymbol{F}'_z}{\sqrt{J_z}} ds_z + \sum_{k_z}^{N_z-1} \Gamma_{zz}^{1/2} \hat{\boldsymbol{e}}_{zk_z} \eta_{zk_z} \tag{4.61}$$

where without loss of generality we have eliminated the mean responses. The above noise fluctuations are independent samples from a standard normal distribution $\mathcal{N}(0,1)$, $\langle ds_z \, \eta_{zk_z} \rangle = 0$,[v] and occupy orthogonal directions in the whitened space, $\hat{\boldsymbol{e}}_{z0}$ and $\hat{\boldsymbol{e}}_{zk_z}$ respectively. Unlike an eigendecomposition, however, their directions are not orthogonal in the neural space $\boldsymbol{R}_z$, $\boldsymbol{F}_z'^{\top} \hat{\boldsymbol{e}}_{zk_z} \neq 0$. This decomposition will be useful for examining the consequences of improper coarse-graining.

For a optimal linear estimate of one subset of nonlinear statistics, we have $\boldsymbol{w}_z^{\text{opt}} = \Gamma_{zz}^{-1} \boldsymbol{F}'_z / J_z$. This decoder removes all noise *except* for the information-limiting noise because

$$\boldsymbol{w}_z^{\text{opt}\top} \Gamma_{zz}^{1/2} \hat{\boldsymbol{e}}_{zk_z} = \frac{\boldsymbol{F}_z'^{\top} \Gamma^{-1}}{J_z} \Gamma_{zz}^{1/2} \hat{\boldsymbol{e}}_{zk_z} = \frac{1}{\sqrt{J_z}} \frac{\boldsymbol{F}_z'^{\top} \Gamma_{zz}^{-1/2}}{\sqrt{J_z}} \hat{\boldsymbol{e}}_{zk_z} = \frac{1}{\sqrt{J_z}} \hat{\boldsymbol{e}}_{z0}^{\top} \hat{\boldsymbol{e}}_{zk_z} = 0 \tag{4.62}$$

and therefore

$$\hat{s}_z = \boldsymbol{w}_z^{\text{opt}\top} \boldsymbol{R}_z = \frac{ds_z}{\sqrt{J_z}} \tag{4.63}$$

which has variance $1/J_z$.

A second crucial aspect of information-limiting correlations is that the noise standard deviation in this direction scales with the population size, just like the signal amplitude. We may neglect noise whose amplitude scales more slowly, since for large populations these fluctuations will be negligible (*i.e.* permit extensive information). We refer to noise along one eigenvector whose variance grows with population size as $O(N_z^2)$ as 'big noise'.

---

[v] For convenience in decomposing noise into information-limiting parts and the rest, here we have chosen to rescale the noise terms ds_z and η_{zk_z} to have unit variance. This differs from to Chapters 2 and 3 where ds_z was assumed to have variance ϵ_z.

Both global and local information-limiting noise are big in this sense, but there may be some other big directions as well. An optimal decoder of a subset of statistics will not permit these directions to influence the output. Importantly, we assume that the populations are redundant enough that this is easy for the brain to avoid by decent fine-grained decoding, at least to a level that the additional variance caused by suboptimality [19,94] competes with the intrinsic information limit in each population. Instead, we assume that the brain's primary challenge is in weighing information from different subsets of statistics.

However, even if this hypothesis is true, any coarse-grained description we identify should not inadvertently allow such big noise to corrupt our inferences about the brain. We have identified a circumstance for which this is a risk, namely when noise is big but not locally information-limiting in one subset of statistics, but *is* correlated with information-limiting noise in another area.

While the noise terms ds_z and η_{zk_z} are independent *within* a subset, in this decomposition they are not independent *across* subsets. Notably, the covariance of the local information-limiting noises in subsets x and y is

$$\langle ds_x\, ds_y \rangle = \sqrt{J_x J_y}\epsilon_{xy} \tag{4.64}$$

where ϵ_{xy} is the corresponding element in Information-limiting scale matrix E. The remaining noise η_z in each group may also be correlated across groups, $\langle \eta_{xk_x} \eta_{yk_y} \rangle \neq 0$. However, according to Equation 4.62, after optimally decoding each redundant subset, none of this noise passes to $\hat{\boldsymbol{s}}$.

The more interesting case is when there are correlations between local information-limiting noise in one group and non-information-limiting big noise in another group: $\langle ds_x\, \eta_{yk_y} \rangle \neq 0$. Upon optimal fine-grained decoding of one redundant subset, the η_{yk_y} noise is lost, and with it, the opportunity to remove the correlated part of ds_x. This

local decoding is equivalent to a loss of a component pointing away from the signal direction, and consequently this reduces the angles between the eigenvectors and the signal direction. The total information content can be written as [1]

$$J = \|\boldsymbol{F}'\|^2 \sum_k \frac{\cos^2\theta_k}{\sigma_k^2} \tag{4.65}$$

where

$$\cos^2\theta_k = \frac{(\boldsymbol{v}_k^\top \boldsymbol{F}')^2}{\|\boldsymbol{F}'\|^2} \tag{4.66}$$

is the angle between the signal direction $\boldsymbol{F}'$ and the kth eigenvector $\boldsymbol{v}_k$ of the covariance Γ, with associated eigenvalue σ_k^2. Note that the sum over components $\sum_k \cos^2\theta_k = 1$ is equal to the same value for all covariances, so as one angle decreases, another must increase. Since these angles θ_k between the signal direction and the directions with large noise variance decrease when the additional big noise component is removed by coarse-graining, it follows that the Fisher Information decreases upon coarse-graining of noise with this structure.

Although our approach of coarse-grained decoding is indeed intended to describe suboptimal neural calculations, our goal is to capture the mild suboptimal scaling of different aspects of redundant codes, rather than the much more perplexing suboptimality of ignoring noise the brain must create itself. In this section we showed how this effort of inferring the scaling of different subsets of statistics can fail if the brain injects large variance noise that masquerades as information-limiting correlations locally. Even in these circumstances, we may consider expanding our coarse-grained description to summarize additional *uninformative* dimensions whose noise is used by the brain [83].

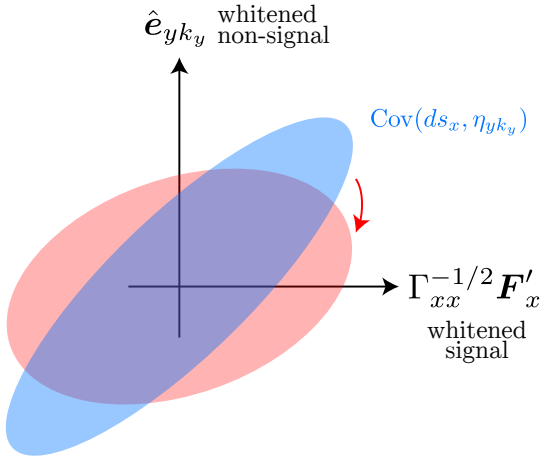


Figure 4.8 : The correlation between the local removable noise in subset y and the local information-limiting noise in another subset x will cause the fine-grained decoding to be an insufficient dimensionality reduction. Here we plot the covariance between local removable noise in subset y, η_{yk_y}, and local information-limiting noise in another subset x, ds_x, in the locally whitened space. The horizontal axis is the direction of both the whitened local signal and local information-limiting noise in subset x. The vertical axis is the direction of k_y-th local removable noise in subset y. We assume they are correlated, $\left\langle ds_x \eta_{yk_y} \right\rangle \neq 0$. The fine-grained decoding in subset y will essentially eliminate the η_{yk_y} noise: it will give the same covariance of $\hat{\boldsymbol{s}}$ regardless of this correlation between preserved information-limiting noise and the filtered big noise η_{yk_y}. This change in correlation reduces the angle, θ_k, between the signal and the eigenvector with large noise variance. The total information will then be more determined by the eigenvector with large noise variance, which decreases the Fisher Information available after the fine-grained decoding step.

4.6 Discussion

In this chapter, I demonstrated the conceptual and practical advantage of modeling computation after dimensionality reduction based on task variables, rather than trying to model individual neural nonlinearities. When the information in the neural responses is redundant, the uncertainty of the brain's estimate cannot be arbitrarily small even for large populations. We saw that, as a consequence, multiple neural transformations can equivalently describe the relationship between the behavioral estimate and the recorded neural responses. As long as we understand how the information is transformed, the detailed patterns of the neural nonlinearities are of lower interest. Thus in a nonlinear redundant code, we are freer to pick a convenient nonlinear basis that approximates the actual neural transformation up to an equivalence relation given by the intrinsic uncertainty, and then apply the distributed two-step decoding scheme to infer a coarse-grained description of the brain's computations. I demonstrated through simulations that this scheme can be an efficient way to analyze the essential property of the brain's computation.

The best condition to apply this two-step decoding scheme is to tasks for which we have some knowledge about the brain's anatomy and encoding model. Ideal tasks should stimulate multiple sensory cortical neural populations with a modest number of nonlinearly entangled nuisance variables influencing each sensory cortical neural population, and require the brain to perform multisensory integration to estimate the task-relevant stimulus. Different nuisance variations will cause the neural representations entangling in different ways between different sensory cortical neural populations. Disentangling them requires different nonlinear operations. We can then make sensible partitioning of the neurons' nonlinear statistics according to where the neurons are recorded and what nonlinear statistics can disentangle neural represen-

tations of the relevant stimulus. Applying the two-step decoding scheme to analyze the neural recordings in these tasks, we can understand how the brain performs multisensory integration with different statistics of multiple neural populations. There are some interesting examples require the brain to perform multisensory integration, including auditory-visual spatial localization [88–93], visual-proprioceptive localization [95] or multisensory visual/vestibular heading task [96–98]. Unfortunately, these experiments are designed without the explicit use of nuisance variables. An important future direction of this work is to test the proposed scheme in multisensory estimation tasks with different nuisance variations affecting sensory inputs.

A more conventional approach to dimensionality reduction is to use PCA on the responses themselves [99], rather than to optimize the decoder as in our method. This is equivalent to using the total correlations rather than the noise correlations. In general, we expect most signal directions to strongly drive neuronal variability, so this dimensionality reduction technique should capture the *signal* well. However, this approach could also be highly suboptimal because it could also capture more *noise* than necessary, because noise correlations that could be filtered out might have a large projection onto the signal space, and instead of removing that noise, signal-based dimensionality reduction would needlessly preserve it.

4.7 Methods

4.7.1 Nonlinear transformations with a monomial basis

Here we build an example to show that we can describe a neural transformation using a monomial nonlinear basis that approximates the actual neural nonlinearities (Section 4.3).

In this example, the actual neural transformation estimates the signal using a two-layer nonlinear neural network receiving input from the upstream neurons $\boldsymbol{r}$. The first layer is a simple nonlinear function $\boldsymbol{\phi}(A\boldsymbol{r}+\boldsymbol{b})$, where A is a mixing matrix, $\boldsymbol{b}$ is a bias term, and $\boldsymbol{\phi}$ is an element-wise nonlinear function. The second layer is a quadratic function of the units in the first layer

$$\Phi_{(ij)}(\boldsymbol{\phi}) = \phi_i(A\boldsymbol{r}+\boldsymbol{b})\phi_j(A\boldsymbol{r}+\boldsymbol{b}) \tag{4.67}$$

The brain's true estimate is a linear weighted sum of the second layer units,

$$\hat{s} = \boldsymbol{W}^\top \boldsymbol{\Phi}(\boldsymbol{\phi}) \tag{4.68}$$

where $\boldsymbol{W}$ is the vector of weights.

Instead of reproducing the details of the actual neural transformation, here we use the monomial basis, $\boldsymbol{R}(\boldsymbol{r}) = \{1, r_i, r_i r_j, r_i r_j r_k, \cdots\}$, to approximate the actual neural transformation. We can use linear regression between the monomial basis $\boldsymbol{R}$ and the true estimates $\hat{s}$ to infer the weights, $\boldsymbol{w}$ (Eq 4.3). The inferred estimate can then be expressed as $\hat{s}_{\text{inferred}} = \boldsymbol{w} \cdot \boldsymbol{R}$. A successful approximation will allow the inferred neural transformation to be in the same equivalence class as the actual neural transformation.

In our simulations, the size of the upstream neural population $\boldsymbol{r}$ is 5. These neural responses are generated from a normal distribution with the mean to be zero and the variance to be 0.1. The nonlinear function in the first layer of the neural network is the hyperbolic tangent, $\phi(x) = \tanh x$. The elements in the mixing matrix A and bias $\boldsymbol{b}$ are generated from a standard normal distribution. The elements in the weights at the last layer, $\boldsymbol{W}$, are randomly generated from a normal distribution with zero mean and variance of 0.1. We infer the weights of the monomial basis up to the fourth order.

In Figure 4.3B, we compare the inferred weights $\boldsymbol{w}$ with the equivalent weights $\boldsymbol{w}_T$ of the true transformation expressed in the monomial basis. The true weights of the actual neural transformation, $\boldsymbol{w}_T = [\boldsymbol{w}_{T0}, \boldsymbol{w}_{T1}, \boldsymbol{w}_{T2}, \boldsymbol{w}_{T3}, \boldsymbol{w}_{T4}]$, are computed from a Taylor expansion on $\hat{s}(\boldsymbol{r})$ at $\boldsymbol{r} = 0$: $\boldsymbol{w}_{Tz} = \frac{\nabla^z \hat{s}}{z!}$ for $z = 0, 1, 2, 3, 4$.

In Figure 4.3C, we show that the inferred estimate $\hat{s}_{\text{inferred}}$ is approximately equal to the true estimate $\hat{s}$, placing the inferred and actual neural transformations in the same equivalence class.

4.7.2 Redundant codes

4.7.2.1 Redundancy induced by cortical expansion

Here we build a simulation to demonstrate the claim in Section 4.4.1: The information redundancy in the neural population can be induced by cortical expansion.

The upstream neural responses $\boldsymbol{\rho}$ are first generated from the cubic codes described in Section 2.4.4 and Section 2.5.5. In these cubic codes, the sufficient statistics in $p(\boldsymbol{\rho}|s)$ are polynomials up to third-order, $\boldsymbol{T}(\boldsymbol{\rho}) = \{\rho_i, \rho_i\rho_j, \rho_i\rho_j\rho_k\}$. We estimate the stimulus by optimally decoding subsets of statistics of $\boldsymbol{\rho}$, $\hat{s}_{\boldsymbol{T}_z} = \boldsymbol{w}_z \boldsymbol{T}_z(\boldsymbol{\rho}) + b_z$, where the subsets are partitioned according to their orders, $\boldsymbol{T}_1 = \{\rho_i\}$, $\boldsymbol{T}_2 = \{\rho_i\rho_j\}$ and $\boldsymbol{T}_3 = \{\rho_i\rho_j\rho_k\}$. The inverse variance of the decoders is interpreted as the information in each subset $\boldsymbol{T}_z$: $J_{\boldsymbol{T}_z} \approx 1/\sigma^2_{\hat{s}_{\boldsymbol{T}_z}}$. In this simulation, the size of the upstream population is $M = 12$ and we generate 10000 trials.

Then we expand the upstream population $\boldsymbol{\rho}$ to the downstream population $\boldsymbol{r}$ by $\boldsymbol{r} = A\boldsymbol{\rho}$, where A is 100×12 matrix whose elements are generated from a standard normal distribution. Thus the size of the downstream population $\boldsymbol{r}$ will be $N = 100$. This expansion mimics the information processing of the cortical expansion from sensory neurons to cortex neurons.

Because the expansion chosen here is a linear transformation, the sufficient statistics for $p(\boldsymbol{r}|s)$ are still polynomials up to third-order, $\boldsymbol{R}(\boldsymbol{r}) = \{r_i, r_i r_j, r_i r_j r_k\}$. We then compute the polynomial nonlinearities of $\boldsymbol{r}$ and then separate them into three subsets according to their orders, $\boldsymbol{R}_1 = \{r_i\}$, $\boldsymbol{R}_2 = \{r_i r_j\}$ and $\boldsymbol{R}_3 = \{r_i r_j r_k\}$.

To show the information is redundant in the linear, quadratic, and cubic statistics, we can construct estimates from random subsets of these statistics and evaluate the information decoded optimally from these estimates according to $\hat{s}_z^{\text{subset}} = \boldsymbol{w}_z \boldsymbol{R}_z^{\text{subset}} + c_z$, where $\boldsymbol{R}_z^{\text{subset}}$ contains units randomly chosen from $\boldsymbol{R}_z$, and where $\boldsymbol{w}_z$ are the optimal weights: $\boldsymbol{w}_z = \text{Cov}(\boldsymbol{R}_z^{\text{subset}}|s)^{-1}\text{Cov}(\boldsymbol{R}_z^{\text{subset}}, s|s)$. The inverse variance of the decoder is interpreted as the decoded information in $\boldsymbol{R}_z^{\text{subset}}$, $J_{\boldsymbol{R}_z^{\text{subset}}} = 1/\sigma^2_{\hat{s}_z^{\text{subset}}}$.

In Figure 4.4, we plot the information ratio $J_{\boldsymbol{R}_z^{\text{subset}}}/J_{\boldsymbol{T}_z}$ against the number of the decoded units. From the simulation, we find that the information ratio saturates at 1 after decoding M^z units. Considering that there are many more N^z units in the $\boldsymbol{R}_z$, there must be multiple units carrying identical information, making the downstream population a redundant code. Thus, we have shown that the cortical expansion can cause information redundancy in the nonlinear statistics of a larger downstream population.

4.7.2.2 Redundant codes with quadratic statistics

Here we create a redundant code where the information is redundant in linear and quadratic statistics of the responses. The redundancy in the subset of the nonlinear statistics is introduced by embedding local information-limiting noise. We will use this redundant quadratic codes in the example in Section 4.5.4.

In Section 2.4.3 and 2.5.4, we created a quadratic code in which the distribution of neural responses is described by the exponential family with polynomials up to

second-order:

$$\boldsymbol{R}_1 = \boldsymbol{r} \tag{4.69}$$

$$\boldsymbol{R}_2 = \delta \boldsymbol{r}^{\otimes 2} \quad \delta \boldsymbol{r} = \boldsymbol{r} - \langle \boldsymbol{r} \rangle_{p(\boldsymbol{r}|s)} \tag{4.70}$$

where we centralize $\boldsymbol{r}$ to get rid of the linear information when computing the quadratic statistics.[vi] A familiar example is the Gaussian distribution with stimulus-dependent mean $\boldsymbol{f}(s)$ and covariance $\Sigma(s)$: $p(\boldsymbol{r}|s) = N(\boldsymbol{f}(s), \Sigma(s))$.

Here we generalize this quadratic code to a redundant code by embedding the Gaussian noise $d\boldsymbol{s} = [ds_1, ds_2] \sim N(\boldsymbol{0}, E)$ separately into linear and quadratic statistics. The encoding probability is then

$$p(\boldsymbol{r}|s, d\boldsymbol{s}) = N(\boldsymbol{f}(s + ds_1), \Sigma(s + ds_2)) \tag{4.71}$$

where the covariance of the embedded noise is

$$E = \begin{pmatrix} \epsilon_{xx} & \epsilon_{xy} \\ \epsilon_{xy} & \epsilon_{yy} \end{pmatrix}. \tag{4.72}$$

As we have derived in Section 2.5.6, we can apply the law of total covariance to decompose the covariance of nonlinear statistics $\boldsymbol{R}(\boldsymbol{r}) = [\boldsymbol{R}_1, \boldsymbol{R}_2]$ conditioned on the stimulus into two parts:

$$\Gamma = \mathrm{Cov}_{\boldsymbol{r},ds}(\boldsymbol{R}(\boldsymbol{r})|s) \tag{4.73}$$

$$= \langle \mathrm{Cov}_{\boldsymbol{r}}(\boldsymbol{R}(\boldsymbol{r})|s, ds) \rangle_{ds} + \mathrm{Cov}_{ds} \langle \boldsymbol{R}(\boldsymbol{r})|s, ds \rangle_{\boldsymbol{r}} \tag{4.74}$$

$$\approx \Gamma_0 + UEU^\top \tag{4.75}$$

where Γ_0 is the covariance of $\boldsymbol{R}$ in the absence of information-limiting correlations and $UEU^\top$ is the local information-limiting correlations; U is the block diagonal matrix

[vi] $\boldsymbol{x}^{\otimes 2} = \mathrm{vect}(\boldsymbol{x} \otimes \boldsymbol{x}) = \{x_i x_j\}$.

whose diagonal blocks contain the nonlinear statistics' sensitivities $\boldsymbol{F}_1' = \partial_s \langle \boldsymbol{R}_1 | s \rangle$ and $\boldsymbol{F}_2' = \partial_s \langle \boldsymbol{R}_2 | s \rangle$

$$U = \begin{pmatrix} \boldsymbol{F}_1' & \mathbf{0} \\ \mathbf{0} & \boldsymbol{F}_2' \end{pmatrix} \tag{4.76}$$

Finally, E is the information-limiting scale matrix described by the covariance of the embedded noise $d\boldsymbol{s}$.

Notice that the expression in Eq 4.75 matches the decomposition in Eq 4.18, suggesting that we have embedded a low-rank structure $UEU^\top$ in the neural nonlinear statistics' covariance. When there is an extensive number of nonlinear statistics that are tuned to the stimulus, we find that $E \gg J_{M0}^{-1}$ (Eq 4.16) where $J_{M0} = U\Gamma_0^{-1}U^\top$. Then the total information in this population will be $J = \mathbf{1}^\top E^{-1}\mathbf{1}$ (Eq 4.17). This result suggests that we have built a redundant code in which the information is redundant in the quadratic statistics of the responses.

4.7.2.3 Redundant codes with cubic statistics

Here we create a redundant code where the information is redundant in linear, quadratic and cubic statistics of the responses. Unlike in Section 4.7.2.1, where redundancy was induced by a cortical expansion, here the redundancy in the subset of the nonlinear statistics is enhanced by directly adding local information-limiting noise as well as by a moderate cortical expansion. We apply this model to generate recorded neural responses used in the simulations that demonstrate the applicability of the distributed coarse-grained decoding scheme (Section 4.5.3).

In Section 2.4.4 and 2.5.5, we created a cubic code in which the distribution of neural responses is described by the exponential family with polynomials up to third-order. We approximate a three-neuron cubic code first using purely cubic

components, $p(\boldsymbol{z}|s) = \frac{1}{Z}\exp\left(-\|\boldsymbol{z}\|^4 + \gamma(s)\, z_i z_j z_k\right)$. For simplicity, we then consider pure cubic codes with non-overlapping cliques of three variables, $p(\boldsymbol{z}|s) = \prod_\alpha p(\boldsymbol{z}_\alpha|s) = \prod_\alpha p(z_{\alpha_1}, z_{\alpha_2}, z_{\alpha_3}|s)$. Finally, we apply a stimulus-dependent affine transformation on the pure cubic code $\boldsymbol{z}$ to include linear and quadratic statistics: $\boldsymbol{\rho} = \boldsymbol{f}(s) + \Sigma^{1/2}(s)\boldsymbol{z}(s)$, where $\boldsymbol{f}(s)$ and $\Sigma(s)$ describes the desired signal-dependent mean and covariance.

Here we generalize this cubic code to a redundant one by adding Gaussian noises directly to the signal variable s. These perturbations are distinct but correlated, and are given by $d\boldsymbol{s} = [ds_1, ds_2, ds_3] \sim \mathcal{N}(\boldsymbol{0}, E)$, and are chosen to affect the linear, quadratic and cubic statistics respectively. Then the resulting responses can be expressed as

$$\boldsymbol{\rho} = \boldsymbol{f}(s + ds_1) + \Sigma^{1/2}(s + ds_2)\boldsymbol{z}(s + ds_3) \tag{4.77}$$

To further enhance the redundancy in the neural population, we expand the upstream neural population $\boldsymbol{\rho}$ with M units to the recorded neural population $\boldsymbol{r}$ with $N \gg M$ units:

$$\boldsymbol{r} = A\boldsymbol{\rho} \tag{4.78}$$

where A is a $N \times M$ expansion matrix. This expansion mimics the information processing from the sensory neurons to the cortex neurons.

In order to represent distinct information in different subset of the statistics, we compute the orthogonal polynomial statistics of $\boldsymbol{r}$:

$$\boldsymbol{R}_1 = \boldsymbol{r} \tag{4.79}$$

$$\boldsymbol{R}_2 = \delta\boldsymbol{r}^{\otimes 2} \quad \delta\boldsymbol{r} = \boldsymbol{r} - \langle\boldsymbol{r}\rangle_{p(\boldsymbol{r}|s)} \tag{4.80}$$

$$\boldsymbol{R}_3 = \boldsymbol{y}^{\otimes 3} \quad \boldsymbol{y} = \Gamma^{-1}_{p(\boldsymbol{r}|s)}\delta\boldsymbol{r} \tag{4.81}$$

where we centralize $\boldsymbol{r}$ to get rid of the linear information when we compute the

quadratic and cubic statistics,[vii] and we whiten $\delta\boldsymbol{r}$ with the conditional covariance of $\boldsymbol{r}$ to get eliminate the quadratic information when computing the cubic statistics.

As we have derived in Section 2.5.6, we can apply the law of total covariance to approximately decompose the covariance of nonlinear statistics $\boldsymbol{R}(\boldsymbol{r}) = [\boldsymbol{R}_1, \boldsymbol{R}_2, \boldsymbol{R}_3]$ conditioned on the stimulus into two parts

$$\Gamma = \mathrm{Cov}(\boldsymbol{R}(\boldsymbol{r})|s) \approx \Gamma_0 + UEU^\top \tag{4.82}$$

where Γ_0 is the information-extensive part and $UEU^\top$ is the local information-limiting correlations. In the local information-limiting part, $U_{\boldsymbol{T}0}$ is a block diagonal matrix whose diagonal blocks contain the nonlinear basis' sensitivities $\boldsymbol{F}_1^{'} = \partial_s \langle \boldsymbol{F}_1(\boldsymbol{r})|s\rangle$, $\boldsymbol{F}_2^{'} = \partial_s \langle \boldsymbol{R}_2(\boldsymbol{r})|s\rangle$ and $\boldsymbol{F}_3^{'} = \partial_s \langle \boldsymbol{R}_3(\boldsymbol{r})|s\rangle$,

$$U = \begin{pmatrix} \boldsymbol{F}_1^{'} & & \\ & \boldsymbol{F}_2^{'} & \\ & & \boldsymbol{F}_3^{'} \end{pmatrix} \tag{4.83}$$

$E = \mathrm{Cov}(d\boldsymbol{s})$ is again the information-limiting scale matrix described by the covariance of the embedded noise $d\boldsymbol{s}$.

Notice that the expression in Eq 4.82 matches the decomposition in Eq 4.18, suggesting that we have embedded a low-rank structure $UEU^\top$ in the neural nonlinear statistics' covariance. The cortical expansion ensures that there are a large number of nonlinear statistics $\boldsymbol{R}(\boldsymbol{r})$ that are tuned to the stimulus. This ensures the approximation: $E \gg J_{M0}^{-1}$ (Eq 4.16) where $J_{M0} = U\Gamma_0^{-1}U^\top$. Then the total information is $J = \mathbf{1}^\top E^{-1}\mathbf{1}$ (Eq 4.17). This suggests that we have built a redundant code in which the information is redundant in linear, quadratic and cubic statistics of the responses.

[vii] $\boldsymbol{x}^{\otimes 3} = \mathrm{vect}(\boldsymbol{x} \otimes \boldsymbol{x} \otimes \boldsymbol{x}) = \{x_i x_j x_k\}$.

4.7.3 Inferring the brain's neural transformation with distributed two-step decoding scheme in a redundant cubic code

Here we describe the simulation used to demonstrate the applicability of distributed two-step decoding scheme (Section 4.5.3).

In this simulation, the upstream neural responses $\boldsymbol{\rho}$ are generated from the redundant cubic codes described in Section 4.7.2.3, where the size of the upstream neural population is $M = 12$ and the size of the recorded neural population $\boldsymbol{r}$ is $N = 100$. We used 10000 trials. The elements in the expansion matrix A were generated from a standard normal distribution. The embedded local information limiting noise was normally distributed with zero mean and covariance

$$E = \begin{pmatrix} 1 & -0.45 & 0.37 \\ -0.45 & 1 & -0.39 \\ 0.37 & -0.39 & 1 \end{pmatrix} \tag{4.84}$$

We simulate the brain's decoder using a two-step decoding scheme, in which the nonlinear functions of the neural population are effectively grouped according to their orders,

$$\boldsymbol{R}_1 = \boldsymbol{r} \tag{4.85}$$

$$\boldsymbol{R}_2 = \delta\boldsymbol{r}^{\otimes 2} \quad \delta\boldsymbol{r} = \boldsymbol{r} - \langle \boldsymbol{r} \rangle_{p(\boldsymbol{r}|s)} \tag{4.86}$$

$$\boldsymbol{R}_3 = \boldsymbol{y}^{\otimes 3} \quad \boldsymbol{y} = \Gamma^{-1}_{p(\boldsymbol{r}|s)} \delta\boldsymbol{r} \tag{4.87}$$

The simulated brain's fine-grained decoding step was optimal, yielding $\hat{s}_z^{\text{brain}} = \boldsymbol{w}_z^{\text{opt}} \boldsymbol{R}_z$. The brain's final estimate was based on this coarse-grained representation according to $\hat{s} = \boldsymbol{a}_{\text{sub}} \hat{\boldsymbol{s}}^{\text{brain}}$, where the suboptimal scaling factors are $\boldsymbol{a}_{\text{sub}} = [0.2105, 1.5789, -0.7895]$.

To analyze the brain's actual neural transformation, we apply the distributed two-step decoding scheme here. We first compute the polynomial basis up to third-order.

Then we group the nonlinear basis according to their order (Eq 4.87). In the fine-grained decoding step, we formulate unbiased estimate $\hat{s}_z$ by decoding subsets of the subset $\boldsymbol{R}_z$:

$$\hat{s}_z = \boldsymbol{w}_z \boldsymbol{R}_z^{\text{subset}} \tag{4.88}$$

where $\boldsymbol{R}_z^{\text{subset}}$ contains M^z units which are randomly chosen from $\boldsymbol{R}_z$, where M is the size of the upstream neural population. $\boldsymbol{w}_z$ is the optimal weights based on the chosen units. From the simulation results in Figure 4.4, these partial estimates can extract all the information in each subset even though they did not decode all the units. This can be confirmed by checking the optimality condition of the fine-grained decoding step (4.26): the measured covariance of the partial estimate vector is approximately equal to the information-limiting scale matrix in the encoding model,

$$\text{Cov}(\hat{\boldsymbol{s}}|s) = \begin{pmatrix} 1.02 & -0.44 & 0.35 \\ -0.44 & 1.05 & -0.30 \\ 0.35 & -0.30 & 0.95 \end{pmatrix} \approx E \tag{4.89}$$

In the coarse-grained decoding step, we infer the scaling factors $\boldsymbol{a}$ by regressing the partial estimate vector $\hat{\boldsymbol{s}}$ against the true estimate $\hat{s}$. In Figure 4.6C, we showed the inferred scaling factors $\boldsymbol{a}$ match the brain's true scaling factors $\boldsymbol{a}_{\text{sub}}$, suggesting that we have successfully inferred the brain's neural transformation by using the distributed two-step decoding scheme.

Chapter 5

Conclusion

Despite the clear importance of computation that is both nonlinear and distributed, and evidence for nonlinear coding in the cortex [49–51], most neuroscience applications of population coding concepts have assumed linear codes and linear readouts [10, 19, 26, 100, 101]. The few that directly address nonlinear population codes either have an impossibly large amount of encoded information [6, 46], or investigate abstract properties unrelated to structured tasks [102].

In this thesis, I contribute a general mathematical framework in which distributed nonlinear computation can be understood and analyzed. My statistical perspective on feedforward nonlinear coding in the presence of nuisance variables provides a useful framework for thinking about neural computation. I provide a remarkably simple test to determine if downstream nonlinear computation decodes all that is encoded. We don't expect that the brain will be optimal in all cases, so for suboptimal computation I propose a coarse-grained description of the neural transformation — modeling computation after a dimensionality reduction based on task variables, rather than trying to model the biophysical details of actual neural transformations.

My 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 neural systems with the nonlinear tasks for which they are required. Some interesting examples where the theory could have practical relevance include motion detection using photoreceptors [62], visual search with distractors (XOR-type tasks) [47, 63], sound localization in early auditory processing before the inferior colliculus [64], or context switching in higher-level cortex [58].

The theory focuses on the encoding and decoding of a single task-relevant variable. Future work can extend this to the neural coding of multiple task-relevant variables, both static and time-dependent. Statistics of neural activities should encode these task-relevant variables individually as well as jointly and even the interactions of these variables. The brain can perform interesting computations— e.g., probabilistic inference — with these neural activities [103, 104]. Then we can apply the theory to identify brain's neural computation.

Our applications included nonlinear tasks, but fairly simple ones. For more complex tasks such as image classification, blindly applying my methods to neural recordings early in the processing chain is unlikely to be useful. This is because the early sensory representation of the target (e.g. image class) will be extremely entangled. Finding a simple nonlinear basis that can disentangle these representations will be very challenging. Instead, a more promising application would be based on downstream neural responses that are tuned to some intermediate features (e.g. object parts, such as eyes in a facial recognition task [105]). We can then apply my method to identify the neural computations proceeding from these intermediate neural responses that can eliminate a smaller number of nuisance variations (such as poses or

facial expressions) [25]. On the other hand, it would be worthwhile to apply my theory to analyze if the computations in multilayer feedforward neural networks are trained to solve a task with modest number of nonlinearly entangled nuisance variables.

My theory of coarse-grained computation has focused on understanding nonlinear neural decoding, which relates neural activity to behavioral output. However, neural computation passes through many stages from sensory input to action, and it is important to understand the mid-level transformations. It would be interesting to generalize my approach so it can apply to recoding as well. Here the aim would be to characterize the transformation between one coarse-grained representation in one brain area to a second coarse-grained representation in another area.

This work is currently limited to feedforward processing, which unquestionably oversimplifies cortical processing. If we want to understand the dynamic neural processes that mediate fluid natural behaviors, we need to consider the brain as a recurrent network. One approach might be to unroll the recurrent network and treat it like a feedforward neural network. Then we can apply the theory to study the relationship between the stimulus, the neurons and the behavior at different times. This unrolled network will have commensurately more variables and more data, which requires a better design of the coarse-graining method that can properly account for the information about the dynamics of the encoded stimulus. A second approach is to directly model the recurrence, and look at equilibrium properties or transient dynamics [80].

After we have identified the neural computations of the brain using my theory, an important next step is to compare the inferred neural computations with the computations predicted by hypothesized algorithms for the same task. When the neural computations match the predictions, we can gain support which algorithm the

brain is using. Based on our proposed measurements we can try to design brain-inspired algorithms according to our inferred neural computations. Since the brain is currently the smartest and most flexible computational device known, if we can learn its transformations and tricks it could be beneficial for artificial intelligence applications aiming to solve real-world problems.

One potential application of neuroscience is to repair or even augment our brains. As neural technology develops for reading and writing neural signals, we need to interpret the data we read and choose what signals to write. These choices must reflect the kinds of neural representations and transformations the brain uses. My thesis provides methods to understand the neural code that are targeted toward the distributed and nonlinear computations that are core properties of the brain. My work should therefore help us better understand, and perhaps even improve, neural computation.

Bibliography

[1] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature neuroscience*, 17(10):1410–1417, 2014.

[2] David Marr. *Vision: A Computational Investigation Into*. WH Freeman, 1982.

[3] Hermann von Helmholtz and James Powell Cocke Southall. *Treatise on physiological optics*, volume 3. Courier Corporation, 2005.

[4] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[5] Maoz Shamir and Haim Sompolinsky. Nonlinear population codes. *Neural computation*, 16(6):1105–1136, 2004.

[6] Alexander S Ecker, Philipp Berens, Andreas S Tolias, and Matthias Bethge. The effect of noise correlations in populations of diversely tuned neurons. *Journal of Neuroscience*, 31(40):14272–14283, 2011.

[7] Qianli Yang and Xaq Pitkow. Robust nonlinear neural codes. In *APS March Meeting Abstracts*, 2015.

[8] Qianli Yang and Xaq Pitkow. Robust nonlinear neural codes. *Cosyne abstract*, 2015.

[9] Qianli Yang. *Nonlinear neural codes*. PhD thesis, Rice University, 2015.

[10] Kenneth H Britten, William T Newsome, Michael N Shadlen, Simona Celebrini, and J Anthony Movshon. A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Visual neuroscience*, 13(1):87–100, 1996.

[11] Michael N Shadlen, Kenneth H Britten, William T Newsome, and J Anthony Movshon. A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, 16(4):1486–1510, 1996.

[12] Jonathan V Dodd, Kristine Krug, Bruce G Cumming, and Andrew J Parker. Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area mt. *Journal of Neuroscience*, 21(13):4809–4821, 2001.

[13] Ian Krajbich, Carrie Armel, and Antonio Rangel. Visual fixations and the computation and comparison of value in simple choice. *Nature neuroscience*, 13(10):1292, 2010.

[14] Victor de Lafuente and Ranulfo Romo. Neuronal correlates of subjective sensory experience. *Nature neuroscience*, 8(12):1698, 2005.

[15] Stefan Treue and Julio C Martinez Trujillo. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature*, 399(6736):575, 1999.

[16] Jamie D Roitman and Michael N Shadlen. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of neuroscience*, 22(21):9475–9489, 2002.

[17] Yong Gu, Dora E Angelaki, and Gregory C DeAngelis. Neural correlates of multisensory cue integration in macaque mstd. *Nature neuroscience*, 11(10):1201,

2008.

[18] Gopathy Purushothaman and David C Bradley. Neural population code for fine perceptual decisions in area mt. *Nature neuroscience*, 8(1):99, 2005.

[19] Xaq Pitkow, Sheng Liu, Dora E Angelaki, Gregory C DeAngelis, and Alexandre Pouget. How can single sensory neurons predict behavior? *Neuron*, 87(2):411–423, 2015.

[20] Ralf M Haefner, Sebastian Gerwinn, Jakob H Macke, and Matthias Bethge. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nature neuroscience*, 16(2):235–242, 2013.

[21] Qianli Yang and Xaq Pitkow. Essential nonlinear properties in neural decoding. *Cosyne abstract*, 2017.

[22] Edward H Adelson and James R Bergen. Spatiotemporal energy models for the perception of motion. *Josa a*, 2(2):284–299, 1985.

[23] Nicole C Rust and James J DiCarlo. Selectivity and tolerance (âinvarianceâ) both increase as visual information propagates from cortical area v4 to it. *Journal of Neuroscience*, 30(39):12978–12995, 2010.

[24] Marino Pagan, Luke S Urban, Margot P Wohl, and Nicole C Rust. Signals in inferotemporal and perirhinal cortex suggest an untangling of visual target information. *Nature neuroscience*, 16(8):1132, 2013.

[25] Ethan M Meyers, Mia Borzello, Winrich A Freiwald, and Doris Tsao. Intelligent information loss: the coding of facial identity, head pose, and non-face informa-

tion in the macaque face patch system. *Journal of Neuroscience*, 35(18):7069–7081, 2015.

[26] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[27] Alexander S Ecker, Philipp Berens, Georgios A Keliris, Matthias Bethge, Nikos K Logothetis, and Andreas S Tolias. Decorrelated neuronal firing in cortical microcircuits. *science*, 327(5965):584–587, 2010.

[28] Alexander S Ecker, Philipp Berens, R James Cotton, Manivannan Subramaniyan, George H Denfield, Cathryn R Cadwell, Stelios M Smirnakis, Matthias Bethge, and Andreas S Tolias. State dependence of noise correlations in macaque primary visual cortex. *Neuron*, 82(1):235–248, 2014.

[29] George H Denfield, Alexander S Ecker, Tori J Shinn, Matthias Bethge, and Andreas S Tolias. Attentional fluctuations induce shared variability in macaque primary visual cortex. *bioRxiv*, page 189282, 2017.

[30] Alexander S Ecker, George H Denfield, Matthias Bethge, and Andreas S Tolias. On the structure of neuronal population activity under fluctuations in attentional state. *Journal of Neuroscience*, 36(5):1775–1789, 2016.

[31] Mehrdad Jazayeri and J Anthony Movshon. Optimal representation of sensory information by neural populations. *Nature neuroscience*, 9(5):690–696, 2006.

[32] Adrian Gopnik Bondy and Bruce G Cumming. Feedback dynamics determine

the structure of spike-count correlation in visual cortex. *bioRxiv*, page 086256, 2016.

[33] Bruno B Averbeck, Peter E Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358, 2006.

[34] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811, 2011.

[35] Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual review of neuroscience*, 39, 2016.

[36] Larry F Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural computation*, 11(1):91–101, 1999.

[37] Marlene R Cohen and John HR Maunsell. Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594, 2009.

[38] Marlene R Cohen and William T Newsome. Estimates of the contribution of single neurons to perception depend on timescale and noise correlation. *Journal of Neuroscience*, 29(20):6635–6648, 2009.

[39] Timothy J Gawne and Barry J Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience*, 13(7):2758–2771, 1993.

[40] Haefner Ralf and Matthias Bethge. Evaluating neuronal codes for inference using fisher information. In *Advances in neural information processing systems*, pages 1993–2001, 2010.

[41] Johannes Burge and Priyank Jaini. Accuracy maximization analysis for sensory-perceptual tasks: Computational improvements, filter robustness, and coding advantages for scaled additive noise. *PLoS computational biology*, 13(2):e1005281, 2017.

[42] Jeffrey Beck, Vikranth R Bejjanki, and Alexandre Pouget. Insights from a simple expression for linear fisher information in a recurrently connected population of spiking neurons. *Neural computation*, 23(6):1484–1502, 2011.

[43] MA Paradiso. A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological cybernetics*, 58(1):35–49, 1988.

[44] Ehud Zohary, Michael N Shadlen, and William T Newsome. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, 370(6485):140–143, 1994.

[45] Haim Sompolinsky, Hyoungsoo Yoon, Kukjin Kang, and Maoz Shamir. Population coding in neuronal systems with correlated noise. *Physical Review E*, 64(5):051904, 2001.

[46] Maoz Shamir and Haim Sompolinsky. Implications of neuronal diversity on population coding. *Neural Computation*, 18(8):1951–1986, 2006.

[47] Marino Pagan, Eero P Simoncelli, and Nicole C Rust. Neural quadratic dis-

criminant analysis: Nonlinear decoding with v1-like computation. *Neural computation*, 28(11):2291–2319, 2016.

[48] Diego A Gutnisky and Valentin Dragoi. Adaptive coding of visual information in neural populations. *Nature*, 452(7184):220, 2008.

[49] Adam Kohn and Matthew A Smith. Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *The Journal of neuroscience*, 25(14):3661–3673, 2005.

[50] Bruno B Averbeck and Daeyeol Lee. Effects of noise correlations on information encoding and decoding. *Journal of Neurophysiology*, 95(6):3633–3644, 2006.

[51] Ifije E Ohiorhenuan, Ferenc Mechler, Keith P Purpura, Anita M Schmid, Qin Hu, and Jonathan D Victor. Sparse coding and high-order correlations in fine-scale cortical networks. *Nature*, 466(7306):617, 2010.

[52] Adrián Ponce-Alvarez, Alexander Thiele, Thomas D Albright, Gene R Stoner, and Gustavo Deco. Stimulus-dependent variability and noise correlations in cortical mt neurons. *Proceedings of the National Academy of Sciences*, 110(32):13162–13167, 2013.

[53] Mattia Rigotti, Omri Barak, Melissa R Warden, Xiao-Jing Wang, Nathaniel D Daw, Earl K Miller, and Stefano Fusi. The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590, 2013.

[54] Marino Pagan and Nicole C Rust. Dynamic target match signals in perirhinal cortex can be explained by instantaneous computations that act on dynamic input from inferotemporal cortex. *The Journal of Neuroscience*, 34(33):11067–11084, 2014.

[55] Marlene R Cohen and William T Newsome. Context-dependent changes in functional circuitry in visual area mt. *Neuron*, 60(1):162–173, 2008.

[56] Richard D Lange and Ralf M Haefner. Inferring the brain9s internal model from sensory responses in a probabilistic inference framework. *bioRxiv*, page 081661, 2016.

[57] Valerio Mante, David Sussillo, Krishna V Shenoy, and William T Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78, 2013.

[58] A Saez, M Rigotti, S Ostojic, S Fusi, and CD Salzman. Abstract context representations in primate amygdala and prefrontal cortex. *Neuron*, 87(4):869–881, 2015.

[59] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17:1410–1417, 2014.

[60] Hendrikje Nienborg and Bruce G Cumming. Psychophysically measured task strategy for disparity discrimination is reflected in v2 neurons. *Nature neuroscience*, 10(12):1608, 2007.

[61] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[62] Tomaso Poggio and Christof Koch. Synapses that compute motion. *Scientific American*, 256(5):46–53, 1987.

[63] Wei Ji Ma, Vidhya Navalpakkam, Jeffrey M Beck, Ronald Van Den Berg, and Alexandre Pouget. Behavior and neural basis of near-optimal visual search. *Nature neuroscience*, 14(6):783, 2011.

[64] Kevin A Davis, Ramnarayan Ramachandran, and Bradford J May. Auditory processing of spectral cues for sound localization in the inferior colliculus. *Journal of the Association for Research in Otolaryngology*, 4(2):148–163, 2003.

[65] Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

[66] Incheol Kang and John HR Maunsell. Potential confounds in estimating trial-to-trial correlations between neuronal response and behavior using choice probabilities. *Journal of neurophysiology*, 108(12):3403–3415, 2012.

[67] David M Green and John A Swets. *Signal detection theory and psychophysics*. John Wiley, 1966.

[68] Andreas S Tolias, Alexander S Ecker, Athanassios G Siapas, Andreas Hoenselaar, Georgios A Keliris, and Nikos K Logothetis. Recording chronically from the same neurons in awake, behaving primates. *Journal of neurophysiology*, 98(6):3780–3790, 2007.

[69] Eve Marder and Adam L Taylor. Multiple models to capture the variability in biological neurons and networks. *Nature neuroscience*, 14(2):133, 2011.

[70] Rufin Van Rullen and Simon J Thorpe. Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. *Neural computation*, 13(6):1255–1283, 2001.

[71] Jacques Gautrais and Simon Thorpe. Rate coding versus temporal order coding: a theoretical approach. *Biosystems*, 48(1-3):57–65, 1998.

[72] Muhammad Naeem, Clemens Brunner, and Gert Pfurtscheller. Dimensionality reduction and channel selection of motor imagery electroencephalographic data. *Computational intelligence and neuroscience*, 2009, 2009.

[73] Abdulhamit Subasi and M Ismail Gursoy. Eeg signal classification using pca, ica, lda and support vector machines. *Expert Systems with Applications*, 37(12):8659–8666, 2010.

[74] Karl J Friston, Christopher D Frith, Richard SJ Frackowiak, and Robert Turner. Characterizing dynamic brain responses with fmri: a multivariate approach. *Neuroimage*, 2(2):166–172, 1995.

[75] Hui Shen, Lubin Wang, Yadong Liu, and Dewen Hu. Discriminative analysis of resting-state functional connectivity patterns of schizophrenia using low dimensional embedding of fmri. *Neuroimage*, 49(4):3110–3121, 2010.

[76] Kofi P Adragni and R Dennis Cook. Sufficient dimension reduction and prediction in regression. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4385–4405, 2009.

[77] Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

[78] Vijay Balasubramanian. Mdl, bayesian inference, and the geometry of the space of probability distributions. *Advances in minimum description length: Theory and applications*, pages 81–98, 2005.

[79] Harald Cramér. *Mathematical methods of statistics*, volume 9. Princeton university press, 1999.

[80] Kaushik Lakshminarasimhan, Alexandre Pouget, Gregory DeAngelis, Dora Angelaki, and Xaq Pitkow. Inferring decoding strategies for multiple correlated neural populations. *bioRxiv*, page 108019, 2017.

[81] Patrik O Hoyer and Aapo Hyvärinen. Interpreting neural response variability as monte carlo sampling of the posterior. In *Advances in neural information processing systems*, pages 293–300, 2003.

[82] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[83] Joel Zylberberg. Untuned but not irrelevant: A role for untuned neurons in sensory information coding. *bioRxiv*, page 134379, 2017.

[84] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[85] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[86] Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995, 2008.

[87] Garrett B Stanley, Fei F Li, and Yang Dan. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *Journal of Neuroscience*, 19(18):8036–8042, 1999.

[88] Konrad P Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS one*, 2(9):e943, 2007.

[89] Mark T Wallace, GE Roberson, W David Hairston, Barry E Stein, J William Vaughan, and Jim A Schirillo. Unifying multisensory signals across time and space. *Experimental Brain Research*, 158(2):252–258, 2004.

[90] Gregg H Recanzone. Auditory influences on visual temporal rate perception. *Journal of neurophysiology*, 89(2):1078–1093, 2003.

[91] Willard R Thurlow and Charles E Jack. Certain determinants of the "ventriloquism effect". *Perceptual and motor skills*, 36(3_suppl):1171–1184, 1973.

[92] David H Warren, Robert B Welch, and Timothy J McCarthy. The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics*, 30(6):557–564, 1981.

[93] Daniel A Slutsky and Gregg H Recanzone. Temporal and spatial dependency of the ventriloquism effect. *Neuroreport*, 12(1):7–10, 2001.

[94] J. M. Beck, W. J. Ma, X. Pitkow, P. E. Latham, and A. Pouget. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–9, 2012.

[95] Samuel J Sober and Philip N Sabes. Flexible strategies for sensory integration during motor planning. *Nature neuroscience*, 8(4):490, 2005.

[96] Aihua Chen, Gregory C DeAngelis, and Dora E Angelaki. Functional specializations of the ventral intraparietal area for multisensory heading discrimination. *Journal of Neuroscience*, 33(8):3567–3581, 2013.

[97] Yong Gu, Sheng Liu, Christopher R Fetsch, Yun Yang, Sam Fok, Adhira Sunkara, Gregory C DeAngelis, and Dora E Angelaki. Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron*, 71(4):750–761, 2011.

[98] Dora E Angelaki, Yong Gu, and Gregory C DeAngelis. Multisensory integration: psychophysics, neurophysiology, and computation. *Current opinion in neurobiology*, 19(4):452–458, 2009.

[99] John P Cunningham and M Yu Byron. Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, 17(11):1500, 2014.

[100] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432, 2006.

[101] Arnulf BA Graf, Adam Kohn, Mehrdad Jazayeri, and J Anthony Movshon. Decoding the activity of neuronal populations in macaque primary visual cortex. *Nature neuroscience*, 14(2):239, 2011.

[102] Baktash Babadi and Haim Sompolinsky. Sparseness and expansion in sensory representations. *Neuron*, 83(5):1213–1226, 2014.

[103] Rajkumar Vasudeva Raju and Xaq Pitkow. Inference by reparameterization in neural population codes. In *Advances in Neural Information Processing Systems*, pages 2029–2037, 2016.

[104] Xaq Pitkow and Dora E Angelaki. Inference in the brain: Statistics flowing in redundant population codes. *Neuron*, 94(5):943–953, 2017.

[105] Le Chang and Doris Y Tsao. The code for facial identity in the primate brain. *Cell*, 169(6):1013–1028, 2017.