ARTICLE **Communicated by Ruben Moreno-Bote**

Heterogeneous Synaptic Weighting Improves Neural Coding in the Presence of Common Noise

Pratik S. Sachdeva
pratik.sachdeva@berkeley.edu
Redwood Center for Theoretical Neuroscience and Department of Physics, University of California, Berkeley, Berkeley, CA 94720 U.S.A., and Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, U.S.A.

Jesse A. Livezey
jlivezey@lbl.gov
Redwood Center for Theoretical Neuroscience, University of California, Berkeley, Berkeley, CA 94720, U.S.A., and Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, U.S.A.

Michael R. DeWeese
deweese@berkeley.edu
Redwood Center for Theoretical Neuroscience, Department of Physics, and Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA 94720 U.S.A.

Simultaneous recordings from the cortex have revealed that neural activity is highly variable and that some variability is shared across neurons in a population. Further experimental work has demonstrated that the shared component of a neuronal population's variability is typically comparable to or larger than its private component. Meanwhile, an abundance of theoretical work has assessed the impact that shared variability has on a population code. For example, shared input noise is understood to have a detrimental impact on a neural population's coding fidelity. However, other contributions to variability, such as common noise, can also play a role in shaping correlated variability. We present a network of linear-nonlinear neurons in which we introduce a common noise input to model—for instance, variability resulting from upstream action potentials that are irrelevant to the task at hand. We show that by applying a heterogeneous set of synaptic weights to the neural inputs carrying the common noise, the network can improve its coding ability as measured by both Fisher information and Shannon mutual information, even in cases where this results in amplification of the common noise. With a broad and heterogeneous distribution of synaptic weights, a population of neurons can remove the harmful effects imposed by afferents that

Neural Computation **32, 1239–1276** (2020)
https://doi.org/10.1162/neco_a_01287

are uninformative about a stimulus. We demonstrate that some nonlinear networks benefit from weight diversification up to a certain population size, above which the drawbacks from amplified noise dominate over the benefits of diversification. We further characterize these benefits in terms of the relative strength of shared and private variability sources. Finally, we studied the asymptotic behavior of the mutual information and Fisher information analytically in our various networks as a function of population size. We find some surprising qualitative changes in the asymptotic behavior as we make seemingly minor changes in the synaptic weight distributions.

1 Introduction

Variability is a prominent feature of many neural systems: neural responses to repeated presentations of the same external stimulus typically vary from trial to trial (Shadlen & Newsome, 1998). Furthermore, neural variability often exhibits pairwise correlations, so that pairs of neurons are more (or less) likely to be co-active than they would be by chance if their fluctuations in activity to a repeated stimulus were independent. These so-called noise correlations (which we also refer to as "shared variability") have been observed throughout the cortex (Averbeck, Latham, & Pouget, 2006; Cohen & Kohn, 2011), and their presence has important implications for neural coding (Zohary, Shadlen, & Newsome, 1994; Abbott & Dayan, 1999).

If the activities of individual neurons are driven by a stimulus shared by all neurons but corrupted by noise that is independent for each neuron (so-called private variability), then the signal can be recovered by simply averaging the activity across the population (Abbott & Dayan, 1999; Ma, Beck, Latham, & Pouget, 2006). If instead some variability is shared across neurons (i.e., there are noise correlations), naively averaging the activity across the population will not necessarily recover the signal, no matter how large the population (Zohary et al., 1994). An abundance of theoretical work has explored how shared variability can be either beneficial or detrimental to the fidelity of a population code (relative to the null model of only private variability among the neurons), depending on its structure and relationship with the tuning properties of the neural population (Zohary et al., 1994; Abbott & Dayan, 1999; Yoon & Sompolinsky, 1999; Sompolinsky, Yoon, Kang, & Shamir, 2001; Averbeck & Lee, 2006; Cohen & Maunsell, 2009; Cafaro & Rieke, 2010; Ecker, Berens, Tolias, & Bethge, 2011; Moreno-Bote et al., 2014; Nogueira et al., 2020).

One general conclusion of this work highlights the importance of the geometric relationship between noise correlations and a neural population's signal correlations (Averbeck et al., 2006; Hu, Zylberberg, & Shea-Brown, 2014). To illustrate this, the mean responses of a neural population across a variety of stimuli (i.e., those responses represented by receptive fields or

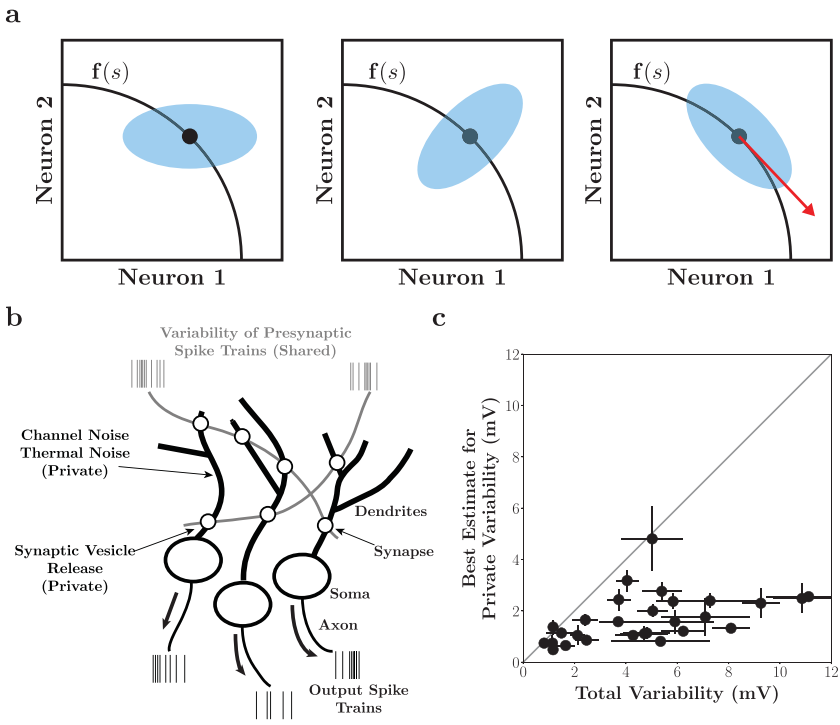

Figure 1: Private and shared variability. (a) The geometric relationship between neural activity and shared variability. Black curves denote mean responses to different stimuli. Variability for a specific stimulus (black dot) may be private (left), shared (middle), or take on the structure of differential correlations (right). The red arrow represents the tangent direction of the mean stimulus response. (b) Schematic of the types of variability that a neural population can encounter. The variability of a neural population contains both private components (e.g., synaptic vesicle release, channel noise, thermal noise) and shared components (e.g., variability of presynaptic spike trains, shared input noise). Shared variability can be induced by the variability of afferent connections (which is shared across a postsynaptic population) or inherited from the stimulus itself. Furthermore, shared variability is shaped by synaptic weighting. (c) Estimates of the private variability contributions to the total variability of neurons ($N = 28$) recorded from auditory cortex of anesthetized rats. Diagonal line indicates the identity. Figure reproduced from Deweese and Zador (2004).

tuning curves) can be examined in the neural space (see Figure 1a, black curves). The correlations among the mean responses for different stimuli specify the signal correlations for a neural population (Averbeck et al., 2006). Private variability exhibits no correlational structure, and thus its

relationship with the signal correlations is determined by the mean neural activity and the individual variances (see Figure 1a, left). Shared variability, however, may reshape neural activity to lie, for example, orthogonal to the mean response curve (see Figure 1a, middle). In the case of Figure 1a, middle, neural coding is improved (relative to private variability) because the variability occupies regions of the neural space that are not traversed by the mean response curve (Montijn, Meijer, Lansink, & Pennartz, 2016). Shared variability can also harm performance, however. Recent work has identified *differential correlations*—those that are proportional to the products of the derivatives of tuning functions (see Figure 1a, right)—as particularly harmful to the performance of a population code (Moreno-Bote et al., 2014). While differential correlations are consequential, they may serve as a small contribution to a population's total shared variability, leaving "nondifferential correlations" as the dominant component of shared variability (Kohn, Coen-Cagli, Kanitscheider, & Pouget, 2016; Montijn et al., 2019; Kafashan et al., 2020).

The sources of neural variability, and their respective contributions to the private and shared components, will have a significant impact on shaping the geometry of the population's correlational structure, and therefore its coding ability (Brinkman, Weber, Rieke, & Shea-Brown, 2016). For example, private sources of variability such as channel noise or stochastic synaptic vesicle release could be averaged out by a downstream neuron receiving input from the population (Faisal, Selen, & Wolpert, 2008). However, sources of variability shared across neurons, such as the variability of presynaptic spike trains from neurons that synapse onto multiple neurons, would introduce shared variability and place different constraints on a neural code (Shadlen & Newsome, 1998; Kanitscheider, Coen-Cagli, & Pouget, 2015). In particular, differential correlations are typically induced by shared input noise (i.e., noise carried by a stimulus) or suboptimal computations (Beck, Ma, Pitkow, Latham, & Pouget, 2012; Kanitscheider et al., 2015).

Past work has examined the contributions of private and shared sources to variability in cortex (Arieli, Sterkin, Grinvald, & Aertsen, 1996; Deweese and Zador, 2004). Specifically, by partitioning subthreshold variability of a neural population into private components (synaptic, thermal, channel noise in the dendrites, and other local sources of variability) and shared components (variability induced by afferent connections), it was found that the private component of the total variability was quite small, while the shared component can be much larger (see Figures 1b and 1c). Thus, neural populations must contend with the large shared component of a neuron's variability. The incoming structure of shared variability and its subsequent shaping by the computation of a neural population is an important consideration for evaluating the strength of a neural code (Zylberberg, Pouget, Latham, & Shea-Brown, 2017).

Moreno-Bote et al. (2014) demonstrated that shared input noise is detrimental to the fidelity of a population code. Here, we instead examine

sources of shared variability, which do not necessarily result in differential correlations (they do not appear as shared input noise) and thus can be manipulated by features of neural computation such as synaptic weighting. We refer to these noise sources as "common noise" to distinguish them from the special case of shared input noise (Vidne et al., 2012; Kulkarni & Paninski, 2007). For example, a common noise source could include an upstream neuron whose action potentials are noisy in the sense that they are unimportant for computing the current stimulus. Common noise, because it is manipulated by synaptic weighting, can serve as a source of nondifferential correlations (see Figure 1a, middle), thereby having either a beneficial or a harmful impact on the strength of the population code. We aim to better elucidate the nature of this impact.

We consider a linear-nonlinear architecture (Paninski, 2004; Karklin & Simoncelli, 2011; Pillow, Paninski, Uzzell, Simoncelli, & Chichilnisky, 2005) and explore how its neural representation is affected by both a common source of variability and private noise sources affecting individual neurons independently. This simple architecture allowed us to analytically assess coding ability using both Fisher information (Abbott & Dayan, 1999; Yoon & Sompolinsky, 1999; Wilke & Eurich, 2002; Wu, Nakahara, & Amari, 2001) and Shannon mutual information. We evaluated the coding fidelity of both the linear representation and the nonlinear representation after a quadratic nonlinearity as a function of the distribution of synaptic weights that shape the shared variability within the representations (Adelson & Bergen, 1985; Emerson, Korenberg, & Citron, 1992; Sakai & Tanaka, 2000; Pagan, Simoncelli, & Rust, 2016). We find that the linear stage representation's coding fidelity improves with diverse synaptic weighting, even if the weighting amplifies the common noise in the neural circuit. Meanwhile, the nonlinear stage representation also benefits from diverse synaptic weighting in a regime where common noise may be amplified, but not too strongly. Moreover, we found that the distribution of synaptic weights that optimized the networks performance depended strongly on the relative amount of private and shared variability. In particular, the neural circuit's coding fidelity benefits from diverse synaptic weighting when shared variability is the dominant contribution to the variability. Together, our results highlight the importance of diverse synaptic weighting when a neural circuit must contend with sources of common noise.

2 Methods

The code used to conduct the analyses described in this article is publicly available on Github (https://github.com/pssachdeva/neuronoise).

2.1 Network Architecture. We consider the linear-nonlinear architecture depicted in Figure 2. The inputs to the network consist of a stimulus s along with common (gaussian) noise ξ_C. The N neurons in the network

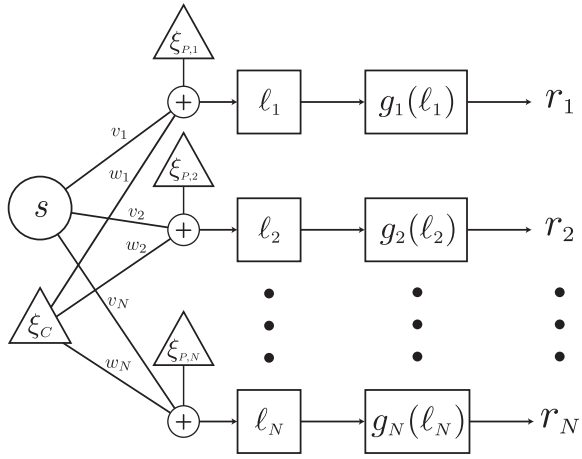

Figure 2: Linear-nonlinear network architecture. The network takes as its inputs a stimulus s and common noise ξ_C. A linear combination of these quantities is corrupted by individual private noises $\xi_{P,i}$. The output of this linear stage is then passed through a nonlinearity $g_i(\ell)$ to produce a "firing rate" r_i. The weights for the linear stage of the network, v_i and w_i, can be thought of as synaptic weighting. Importantly, the common noise is distinct from shared input noise because it is manipulated by the synaptic weighting.

take a linear combination of the inputs and are further corrupted by independent and identically distributed (i.i.d.) private gaussian noise. Thus, the output of the linear stage for the ith neuron is

$$\ell_i = v_i s + w_i \sigma_C \xi_C + \sigma_P \xi_{P,i}, \tag{2.1}$$

where $\xi_{P,i}$ is the private noise, v_i and w_i are the weights, and the common and private noise terms are scaled by positive constants σ_C and σ_P. The noisy linear combination is passed through a nonlinearity $g_i(\ell_i)$ whose output r_i can be thought of as a firing rate.

Thus, the network-wide computation is given by

$$\mathbf{r} = \mathbf{g}(\mathbf{v}s + \mathbf{w}\sigma_C \xi_C + \sigma_P \boldsymbol{\xi}_P), \tag{2.2}$$

where $\mathbf{g}(\boldsymbol{\ell})$ is an element-wise application of the network nonlinearity.

2.2 Measures of Coding Strength. In order to assess the fidelity of the population code represented by $\boldsymbol{\ell}$ or $\mathbf{r}$, we turn to the Fisher information and the Shannon mutual information (Cover & Thomas, 2012). The former has largely been used in the context of sensory decoding and correlated variability (Abbott & Dayan, 1999; Averbeck et al., 2006; Kohn et al., 2016) while

the latter has been well studied in the context of efficient coding (Attneave, 1954; Barlow, 1961; Bell & Sejnowski, 1997; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1999).

The Fisher information sets a limit by which the readout of a population code can determine the value of the stimulus. Formally, it sets a lower bound to the variance of an unbiased estimator for the stimulus. In terms of the network architecture, the Fisher information of the representation $\mathbf{r}$ (or $\boldsymbol{\ell}$) quantifies how well s can be decoded given the representation. For gaussian noise models with stimulus-independent covariance, the Fisher information is equal to the linear Fisher information (LFI):

$$I_{LFI}(s) = \frac{\partial \mathbf{f}(s)}{\partial s}^T \boldsymbol{\Sigma}^{-1}(s) \frac{\partial \mathbf{f}(s)}{\partial s}, \tag{2.3}$$

where $\mathbf{f}(s)$ and $\boldsymbol{\Sigma}(s)$ are the mean and covariance of the response (here, $\mathbf{r}$ or $\boldsymbol{\ell}$) to the stimulus s. In other cases, the LFI serves as a lower bound for the Fisher information and thus is a useful proxy when the Fisher information is challenging to calculate analytically. The estimator for I_{LFI} is the locally optimal linear estimator (Kohn et al., 2016).

The Shannon mutual information quantifies the reduction in uncertainty of one random variable given knowledge of another:

$$I[s, \mathbf{f}] = \int ds d\mathbf{f}\; p(s, \mathbf{f}) \log \left(\frac{p(s, \mathbf{f})}{p(s)p(\mathbf{f})} \right). \tag{2.4}$$

Earlier work demonstrated that the Fisher information provides a lower bound for the Shannon mutual information in the case of gaussian noise (Brunel & Nadal, 1998). However, more recent work has revealed that the relationship between the two is more nuanced, particularly in the cases where the noise model is nongaussian (Wei & Stocker, 2016). Thus, we supplement our assessment of the network's coding ability by measuring the mutual information, $I[s, \mathbf{r}]$, between the neural representation $\mathbf{r}$ and the stimulus s. As with the Fisher information, the mutual information is often intractable but fortunately can be estimated from data. Specifically, we employ the estimator developed by Kraskov and colleagues, which uses entropy estimates from k-nearest neighbor distances (Kraskov, Stögbauer, & Grassberger, 2004).

2.3 Structured Weights. The measures of coding strength are a function of the weights that shape the interaction of the stimulus and noise in the network. Thus, the choice of the synaptic weight distribution affects the calculation of these quantities. We first consider the case of structured weights in order to obtain analytical expressions for measures of coding strength.

Structured weights take on the form

$$\mathbf{w} = \left(\underbrace{1 \cdots 1}_{N/k \text{ times}} \ \underbrace{2 \cdots 2}_{N/k \text{ times}} \ \cdots \ \underbrace{k \cdots k}_{N/k \text{ times}} \right)^T . \tag{2.5}$$

Specifically, the structured weight vectors are parameterized by an integer k that divides the N weights into k homogeneous groups. The weights across the groups span the positive integers up to k. Importantly, larger k will only increase the weights in the vector. Thus, in the above scheme, increased "diversity" can be achieved only by increasing k, which will invariably result in an amplification of the signal to which the weight vector is applied. In the case that k does not evenly divide N, each group is repeated $\lceil N/k \rceil$ times, except the last group, which is only repeated $N - (N-1) \cdot \lceil N/k \rceil$ times (that is, the last group is truncated to ensure the weight vector is of size N).

Additionally, we consider cases in which k is of order N, for example, $k = N/2$. Allowing k to grow with N ensures that typical values for the weights grow with the population size. This contrasts with the case in which k is a constant, such as $k = 4$, which sets a maximum weight value independent of the population size.

2.4 Unstructured Weights. While the structured weights allow for analytical results, they possess an unrealistic distribution of synaptic weighting. Thus, we also consider the case of unstructured weights, in which the synaptic weights are drawn from some parameterized probability distribution:

$$\mathbf{v} \sim p(\mathbf{v}; \theta_{\mathbf{v}}); \ \mathbf{w} \sim p(\mathbf{w}; \theta_{\mathbf{w}}). \tag{2.6}$$

We calculate both information-theoretic quantities over many random draws from these distributions and observe how these quantities behave as some subset of the parameters θ is varied. In particular, we focus on the log-normal distribution (Iyer, Menon, Buice, Koch, & Mihalas, 2013), which has been found to describe the distribution of synaptic weights well in slice electrophysiology (Song, Sjöström, Reigl, Nelson, & Chklovskii, 2005; Sargent, Saviane, Nielsen, DiGregorio, & Silver, 2005). Specifically, the weights take on the form

$$\mathbf{w} \sim \Delta + \text{Lognormal}(\mu, \sigma), \tag{2.7}$$

where $\Delta > 0$. For a log-normal distribution, an increase in μ will increase the distribution's mean, median, and mode (see Figure 3e, inset). Thus, μ as a parameter acts similar to k for the structured weights in that increased weight diversity must be accompanied by an increase in their magnitude.

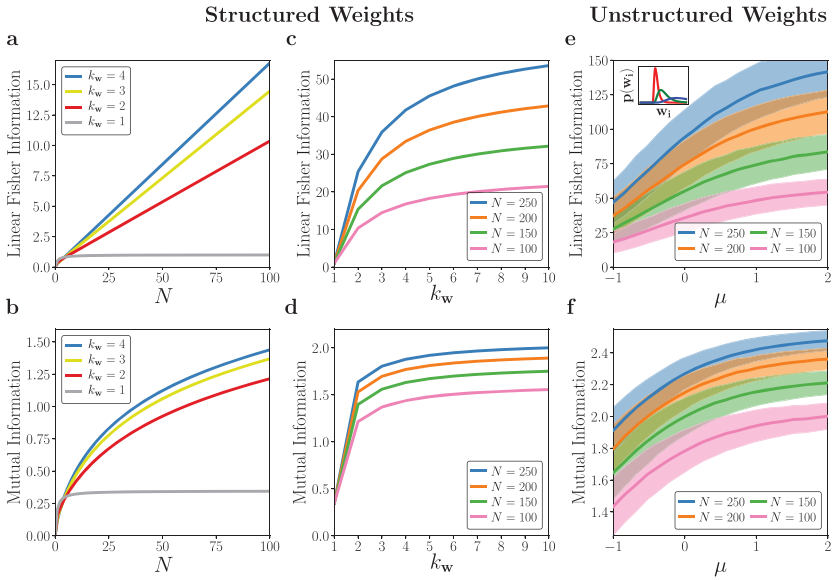

Figure 3: Network coding performance of the linear stage representation. Here, the noise variances are $\sigma_P^2 = \sigma_C^2 = 1$. Fisher information is shown on the top row and mutual information on the bottom row. (a, b) Structured weights. Linear Fisher information and mutual information are shown as a function of the population size, N, across different levels of weight heterogeneity, $k_{\mathbf{w}}$ (indicated by color). (c, d) Linear fisher information and mutual information are shown as a function of weight heterogeneity, $k_{\mathbf{w}}$, for various population sizes, N. (e, f) Unstructured weights. Linear Fisher information and mutual information are shown as a function of the mean of the log-normal distribution used to draw common noise synaptic weights. Information quantities are calculated across 1000 random drawings of weights: solid lines depict the means while the shaded region indicates one standard deviation. Inset: The distribution of weights for various choices of μ. Increasing μ shifts the distribution to the right, increasing heterogeneity.

3 Results

We consider the network's coding ability after both the linear stage ($\boldsymbol{\ell}$) and the nonlinear stage ($\mathbf{r}$). In other words, the linear stage can be considered the output of the network assuming each of the functions $g_i(\ell_i)$ is the identity. Furthermore, due to the data processing inequality, the qualitative conclusions we obtain from the linear stage should apply for any one-to-one nonlinearity.

3.1 Linear Stage. The Fisher information about the stimulus in the linear representation can be shown to be (see appendix A.1.1 for the derivation)

$$I_F(s) = \frac{1}{\sigma_P^2} \frac{\left(\sigma_P^2/\sigma_C^2\right)|\mathbf{v}|^2 + \left(|\mathbf{v}|^2|\mathbf{w}|^2 - (\mathbf{v}\cdot\mathbf{w})^2\right)}{\left(\sigma_P^2/\sigma_C^2\right) + |\mathbf{w}|^2} \tag{3.1}$$

$$= \frac{|\mathbf{v}|^2}{\sigma_P^2} \frac{\left(\sigma_P^2/\sigma_C^2\right) + |\mathbf{w}|^2 \sin^2\theta}{\left(\sigma_P^2/\sigma_C^2\right) + |\mathbf{w}|^2}, \tag{3.2}$$

which is equivalent to the linear Fisher information in this case. In equation 3.2, θ refers to the angle between $\mathbf{v}$ and $\mathbf{w}$. The mutual information can be expressed as (see appendix A.1.2 for the derivation)

$$I[s, \boldsymbol{\ell}] = \frac{1}{2}\log\left[1 + \sigma_S^2 I_F(s)\right]. \tag{3.3}$$

For the case the mutual information, we have assumed that the prior distribution for the stimulus is gaussian with zero mean and variance σ_S^2.

Examining equation 3.2 reveals that increasing the norm of $\mathbf{v}$ without changing its direction (that is, without changing θ) will increase the Fisher information, while increasing the norm of $\mathbf{w}$ without changing its direction will either decrease or maintain information (since $0 \leq \sin^2\theta \leq 1$). Additionally, if $\mathbf{v}$ and $\mathbf{w}$ become more aligned while leaving their norms unchanged, the Fisher information will decrease (since $\sin^2\theta$ will decrease). This decrease in Fisher information is consistent with the observation that alignment of $\mathbf{v}$ and $\mathbf{w}$ will produce differential correlations. If $\mathbf{v}$ and $\mathbf{w}$ are changed in a way that modulates both their norm and direction, the impact on Fisher information is less transparent.

To better understand the Fisher information, we impose a parameterized structure on the weights that allows us to increase weight diversity without decreasing the magnitude of any of the weights. This weight parameterization, which we call the structured weights, is detailed in section 2.3. We chose this parameterization for two reasons. First, we desired a scheme in which an increase in diversity must be accompanied by an amplification of common noise. We chose this behavior so that any improvement in coding ability can only be explained by the increase in diversity rather than a potential decrease in common noise. Second, we desired analytic expressions for the Fisher information as a function of population size, which is possible with this form of structured weights.

Under the structured weight parameterization, equations 3.1 and 3.3 can be explored by varying the choice of k for both $\mathbf{v}$ and $\mathbf{w}$ (we refer to them as $k_{\mathbf{v}}$ and $k_{\mathbf{w}}$, respectively). It is simplest and most informative to examine these quantities by setting $k_{\mathbf{v}} = 1$ while allowing $k_{\mathbf{w}}$ to vary, as amplifying and diversifying $\mathbf{v}$ will only increase coding ability for predictable reasons (this is indeed the case for our network) (Shamir & Sompolinsky,

2006; Ecker et al., 2011). While increasing $k_{\mathbf{w}}$ will boost the overall amount of noise added to the neural population, it also changes the direction of the noise in the higher-dimensional neural space. Thus, while we might expect that adding more noise in the system would hinder coding, the relationship between the directions of the noise and stimulus vectors in the neural space also plays a role.

We first consider how the Fisher information and mutual information are affected by the choice of $k_{\mathbf{w}}$. In the structured regime, we have

$$|\mathbf{v}|^2 = N, \tag{3.4}$$

$$\mathbf{v} \cdot \mathbf{w} = \frac{N}{k} \sum_{i=1}^{k} i = \frac{N(k+1)}{2}, \tag{3.5}$$

$$|\mathbf{w}|^2 = \frac{N}{k} \sum_{i=1}^{k} i^2 = \frac{N(k+1)(2k+1)}{6}, \tag{3.6}$$

which allows us to rewrite equation 3.1 as

$$I_F(s) = I_F = \frac{N}{2\sigma_P^2} \frac{12(\sigma_P^2/\sigma_C^2) + N(k^2-1)}{6(\sigma_P^2/\sigma_C^2) + N(2k^2+3k+1)}. \tag{3.7}$$

The form of the mutual information follows directly from plugging equation 3.7 into equation 3.3.

The analytical expressions for the structured regime reveal the asymptotic behavior of the information quantities. Neither quantity saturates as a function of the number of neurons, N, except in the case of $k_{\mathbf{w}} = 1$ (see Figures 3a and 3b). In this regime, increasing the population size of the system also enhances coding fidelity. Furthermore, both quantities are monotonically increasing functions of the common noise synaptic heterogeneity, $k_{\mathbf{w}}$ (see Figures 3c and 3d), implying that decoding is enhanced despite the fact that the amplitude of the common noise is magnified for larger $k_{\mathbf{w}}$. Our analytical results show linear and logarithmic growth for the Fisher and mutual information, respectively, as one might expect in the case of gaussian noise (Brunel & Nadal, 1998). These qualitative results hold for essentially any choice of $(\sigma_S, \sigma_P, \sigma_C)$.

In the case of $k_{\mathbf{w}} = 1$, the signal and common noise are aligned perfectly in the neural representation. Thus, the common noise becomes equivalent in form to shared input noise. As a consequence, we observe the saturation of both Fisher information and mutual information as a function of the neural population. This saturation implies the existence of differential correlations, consistent with the observation that information-limiting correlations occur under the presence of shared input noise (Kanitscheider et al., 2015).

The structured weight distribution we described allows us to derive analytical results, but the limitation to only a fixed number of discrete synaptic

weight values is not realistic for biological networks. Thus, we use unstructured weights, described in section 2.4, in which the synaptic weights are drawn from a log-normal distribution. In this case, we estimate the linear Fisher information and the mutual information over many random draws according to $w_i \sim \Delta + \text{Lognormal}(\mu, \sigma^2)$. We are primarily concerned with varying μ, as an increase in this quantity uniformly increases the mean, median, and mode of the log-normal distribution (see Figure 3e, inset), akin to increasing $k_{\mathbf{w}}$ for the structured weights.

Our numerical analysis demonstrates that increasing μ increases the average Fisher information and average mutual information across population sizes (see Figures 3e and 3f: bold lines). In addition, the benefits of larger weight diversity are felt more strongly by larger populations (see Figures 3e and 3f: different colors).

In the structured weight regime, our analytical results show that weight heterogeneity can ameliorate the harmful effects of *additional* information-limiting correlations induced by common noise mimicking shared input noise. They do not imply that weight heterogeneity prevents differential correlations, as the common noise in this model is manipulated by synaptic weighting, in contrast with true shared input noise. For unstructured weights, we once again observe that larger heterogeneity affords the network improved coding performance, despite the increased noise in the system. Together, these results show that linear networks could manipulate common noise to prevent it from causing induced differential correlations. However, neural circuits, which must perform other computations that may dictate the structure of the weights on the common noise inputs, can still achieve good decoding performance provided that the circuits' synaptic weights are heterogeneous.

3.2 Quadratic Nonlinearity. We next consider the performance of the network after a quadratic nonlinearity $g_i(x) = x^2$ for all neurons i. This nonlinearity has been used in a neural network model to perform quadratic discriminant analysis (Pagan et al., 2016) and as a transfer function in complex cell models (Adelson & Bergen, 1985; Emerson et al., 1992; Sakai & Tanaka, 2000). Furthermore, we chose this nonlinearity because we were able to calculate the linear Fisher information analytically (as an approximation to the Fisher information); see appendix A.3 for a numerical analysis with an exponential nonlinearity. However, the mutual information is apparently not analytically tractable; we performed a numerical approximation using simulated data.

3.2.1 Linear Fisher Information. An analytic expression of the linear Fisher information is calculated in appendix A.1.3. Its analytic form is too complicated to be restated here, but we will examine it numerically for both the structured and unstructured weights. The qualitative behavior of the Fisher information depends on the magnitude of the common variability (σ_C) and

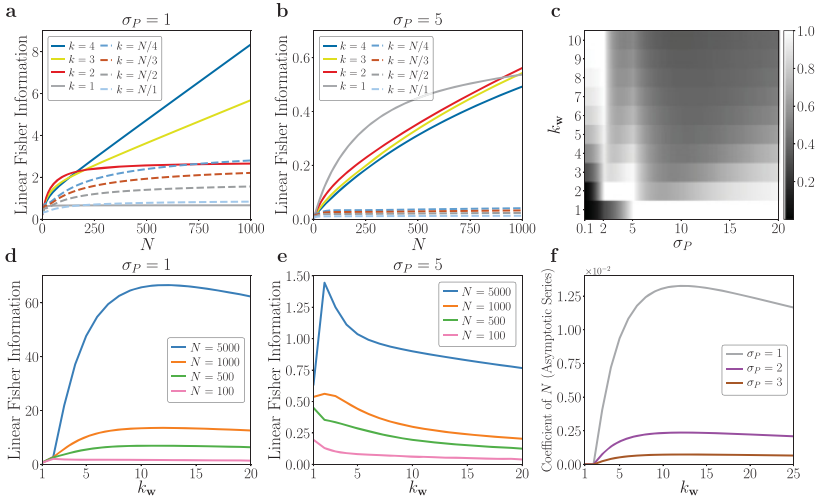

Figure 4: Linear Fisher information after quadratic nonlinearity in a network with structured weights. (a) Fisher information as a function of population size when $\sigma_P = \sigma_C = 1$, that is, private and common noise have equal variances. Solid lines denote constant k while dashed lines denote k scaling with population size. (b) Same as panel a, but for a network where private variance dominates ($\sigma_P = 5, \sigma_C = 1$). (c) Normalized Fisher information. For a choice of σ_P, the Fisher information is calculated for a variety of $k_{\mathbf{w}}$ (y-axis) and divided by the maximum Fisher information (across the $k_{\mathbf{w}}$, for the choice of σ_P). For a given σ_P, the normalized Fisher information is equal to one at the value of $k_{\mathbf{w}}$, which maximizes decoding performance. (d) Behavior of the Fisher information as a function of synaptic weight heterogeneity for various population sizes ($\sigma_P = \sigma_C = 1$). (e) Same as panel d, but for networks where private variance dominates ($\sigma_P = 5, \sigma_C = 1$). (f) The coefficient of the linear term in the asymptotic series of the Fisher information at different levels of private variability. At $k_{\mathbf{w}} = 1, 2$, the coefficient of N is exactly zero.

private variability (σ_P) in a more complicated fashion than the linear stage, which depends on these variables primarily through their ratio σ_C/σ_P. Thus, we separately consider how common and private variability affect coding efficacy under various synaptic weight structures.

As before, we first consider the structured weights with $k_{\mathbf{v}}$ set to 1 while only varying $k_{\mathbf{w}}$. We start with the special case where $\sigma_P = \sigma_C = 1$ (i.e., equal private and common noise variance). Here, the Fisher information saturates for both $k_{\mathbf{w}} = 1$ and $k_{\mathbf{w}} = 2$, but increases without bound for larger $k_{\mathbf{w}}$ (see Figure 4a). We can also consider the case where the structured weight heterogeneity grows in magnitude with the population size (i.e., $k_{\mathbf{w}}$ is a

function of N). In this scenario, the Fisher information is much smaller and saturates (see Figure 4a, dashed lines).

The information saturation (or growth) for various $k_{\mathbf{w}}$ can be understood in terms of the geometry of the covariance describing the neural population's variability. Information saturation occurs if the principal eigenvector(s) of the covariance align closely (but not necessarily exactly) with the differential correlation direction, $\mathbf{f}'$, while the remaining eigenvectors quickly become orthogonal to $\mathbf{f}'$ as population size increases (Moreno-Bote et al., 2014; see appendix A.2 for more details). When $k_{\mathbf{w}} = 1$, the common noise aligns perfectly with the stimulus, and so the principal eigenvector of the covariance aligns exactly with $\mathbf{f}'$ (as in Figure 1a, right). When $k_{\mathbf{w}} > 1$, the principal eigenvector aligns closely, but not exactly, with the differential correlation direction. However, when $k_{\mathbf{w}} = 2$, the remaining eigenvectors become orthogonal quickly enough for information to saturate. This does not occur when $k_{\mathbf{w}} > 2$. The case of $k_{\mathbf{w}} \sim O(N)$, meanwhile, is slightly different. Here, the variances of the covariance matrix scale with population size, so that the neurons simply exhibit too much variance for any meaningful decoding to occur. However, we believe that it is unreasonable to expect that the synaptic weights of a neural circuit scale with the population size, making this scenario biologically implausible.

When private variability dominates, we observe qualitatively different finite network behavior ($\sigma_P = 5$; see Figure 4b). For $N = 1000$, both $k_{\mathbf{w}} = 1$ and $k_{\mathbf{w}} = 2$ exhibit better performance relative to larger values of $k_{\mathbf{w}}$ (by contrast, the case with $k_{\mathbf{w}} \sim O(N)$ quickly saturates). We note that, unsurprisingly, the increase in private variability has decreased the Fisher information for all cases we considered compared to $\sigma_P = 1$ (compare the scales of Figures 4a and 4b). Our main interest, however, is identifying effective synaptic weighting strategies given some amount of private and common variability.

The introduction of the squared nonlinearity produces qualitatively different behavior at the finite network level. In contrast with Figure 3, increased heterogeneity does not automatically imply improved decoding. In fact, there is a regime in which increased heterogeneity improves Fisher information, beyond which we see a reduction in decoding performance (see Figure 4d). If the private variability is increased, this regime shrinks or becomes nonexistent, depending on the population size (see Figure 4e). Furthermore, entering this regime for higher private variability requires smaller $k_{\mathbf{w}}$ (i.e., less weight heterogeneity).

The results shown in Figures 4d and 4e imply that there exists an interesting relationship among the network's decoding ability, its private variability, and its synaptic weight heterogeneity $k_{\mathbf{w}}$. To explore this further, we examine the behavior of the Fisher information at a fixed population size ($N = 1000$) as a function of both σ_P and $k_{\mathbf{w}}$ (see Figure 4c). To account for the fact that an increase in private variability will always decrease the Fisher information, we calculate the *normalized* Fisher information: for a given

choice of σ_P, each Fisher information is divided by the maximum across a range of $k_{\mathbf{w}}$ values. Thus, a normalized Fisher information allows us to determine what level of synaptic weight heterogeneity maximizes coding fidelity, given a particular level of private variability σ_P.

Figure 4c highlights three interesting regimes. When the private variability is small, the network benefits from larger weight heterogeneity on the common noise. But as the neurons become noisier, the "Goldilocks zone" in which the network can leverage larger noise weights becomes constrained. When the private variability is large, the network achieves superior coding fidelity by having less heterogeneous weights, despite the threat of induced differential correlations from the common noise. Between these regimes, there are transitions for which many choices of $k_{\mathbf{w}}$ result in equally good decoding performance.

It is important to point out that Figures 4a to 4e capture only finite network behavior. Therefore, we extended our analysis by validating the asymptotic behavior of the Fisher information as a function of the private noise by examining its asymptotic series at infinity (see Figure 4f). For $k_{\mathbf{v}} = 1, 2$, the coefficient of the linear term is zero for any choice of σ_P, implying that the Fisher information always saturates. In addition, when the common noise weights increase with population size (i.e., $k_{\mathbf{w}} \sim O(N)$), the asymptotic series is always sublinear (not shown in Figure 4f). Thus, there are multiple cases in which the structure of synaptic weighting can induce differential correlations in the presence of common noise. Increased heterogeneity allows the network to escape these induced differential correlations and achieve linear asymptotic growth. If $k_{\mathbf{w}}$ becomes too large, however, the linear asymptotic growth begins to decrease. Once $k_{\mathbf{w}}$ scales as the population size, differential correlations are once again significant.

Next, we reproduce the analysis with unstructured weights. As before, we draw 1000 samples of common noise weights from a shifted log-normal distribution with varying μ. The behavior of the average (linear) Fisher information is qualitatively similar to that of the structured weights (see Figure 5). There exists a regime for which larger weight heterogeneity improves the decoding performance, beyond which coding fidelity decreases (see Figure 5a). If the private noise variance dominates, this regime begins to disappear for smaller networks (see Figure 5b). Thus, with very noisy neurons, the coding fidelity of the network is improved when the synaptic weights are less heterogeneous (and therefore smaller).

To summarize these results, we once again plot the normalized Fisher information (this time, normalized across choices of μ and averaged over 1000 samples from the log-normal distribution) for a range of private variabilities (see Figure 5c). The heat map exhibits a similar transition at a specific level of private variability. At this transition, a wide range of μ's provide the network with similar decoding ability. For smaller σ_P, we see behavior comparable to Figure 5a, where there exists a regime of improved Fisher information. Beyond the transition, the network performs better with less

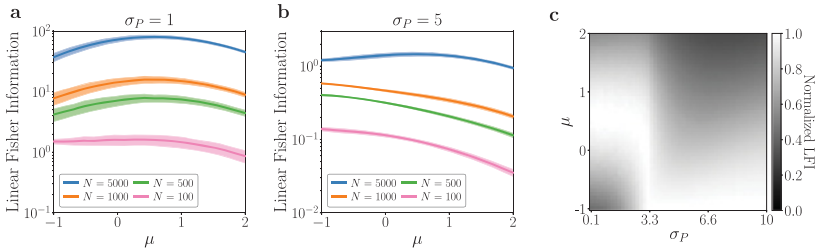

Figure 5: Linear Fisher information after quadratic nonlinearity, unstructured weights. In contrast to Figure 4, panels a and b are plotted on a log scale. (a) Linear Fisher information as a function of the mean, μ, of the log-normal distribution used to draw the common noise synaptic weights. Solid lines denote means, while shaded regions denote one standard deviation across the 1000 drawings of weights from the log-normal distribution. (b) Same as panel a but for networks in which private variability dominates ($\sigma_P = 5$, $\sigma_C = 1$). (c) Normalized linear Fisher information. Same plot as Figure 4c, but the average Fisher information across the 1000 samples is normalized across μ (akin to normalizing across $k_{\mathbf{w}}$).

diverse synaptic weighting, though it becomes less stringent as σ_P increases. The behavior exhibited by this heat map is similar to Figure 4c but contains fewer uniquely identifiable regions. This may imply that the additional regions in Figure 4c are an artifact of the structured weights.

The amount of the common noise will also affect how the network behaves and what levels of synaptic weight heterogeneity are optimal. For example, consider a network with private noise variability set to $\sigma_P = 1$. When common noise is small, the Fisher information is comparable among various choices of synaptic weight diversity (see Figure 6a). When the common noise dominates, however, the network benefits strongly from diverse weighting (see Figure 4b), though it is punished less severely for having $k_{\mathbf{w}}$ scale with N (see Figure 6b, dashed lines; compare to Figure 4b). These observations are true at finite population size. As before, the Fisher information saturates for $k_{\mathbf{w}} = 1, 2$ and $k_{\mathbf{w}} \sim O(N)$, no matter the choice of common noise variance.

We calculated the normalized Fisher information across a range of common noise strengths to determine the optimal synaptic weight distribution. The results for structured weights and unstructured weights are shown in Figures 6c and 6d, respectively. While they strongly resemble Figures 4c and 5c, they exhibit opposite qualitative behavior. As before, there are three identifiable regions in Figure 6c, each divided by abrupt transitions where many choices of $k_{\mathbf{w}}$ are equally good for decoding. For small common noise, the coding fidelity is improved with less heterogeneous weights, but as the common noise increases, the network enters the Goldilocks regions.

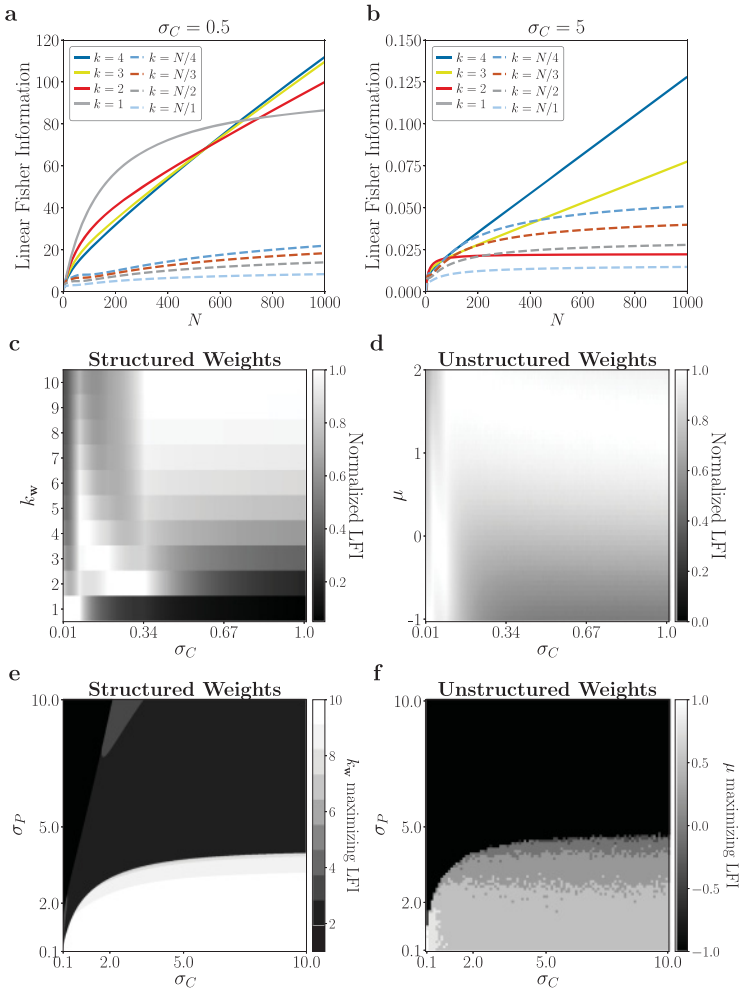

Figure 6: The relationship among common noise, private noise, and synaptic weight heterogeneity. (a, b) Fisher information as a function of population size, N, when common noise contribution is drowned out by private noise (a) and common noise dominates ($\sigma_P = 1$) (b). Solid lines indicate constant $k_{\mathbf{w}}$, while dashed lines refer to $k_{\mathbf{w}}$ that scales with N. (c, d) Normalized Fisher information as a function of common noise for structured weights (c) and unstructured weights (d). For unstructured weights, each Fisher information is calculated by averaging over 1000 networks with their common noise weights drawn from the respective distribution. (e) The value of $k_{\mathbf{w}}$ that maximizes the network's Fisher information for a given choice of σ_P and σ_C. The maximum is taken over $k_{\mathbf{w}} \in [1, 10]$. (f) The value of μ that maximizes the average Fisher information over 1000 draws for a given choice of σ_P and σ_C.

After another abrupt transition near $\sigma_C \approx 0.34$, the network performance is greatly improved by heterogeneous weights.

Thus, common noise and private noise seem to have opposite impacts on the optimal choice of synaptic weight heterogeneity. When private noise dominates, the Fisher information is maximized under a set of homogeneous weights, since coding ability is harmed by amplification of common noise. When common noise dominates, the network coding is improved under diverse weighting: this prevents additional differential correlations and helps the network cope with the punishing effects on coding due to the amplified noise.

How should we choose the synaptic weight distribution within the extremes of private or common noise dominating? We assess the behavior of the Fisher information as both σ_P and σ_C are varied over a wide range. For the structured weights, we calculate the choice of $k_{\mathbf{w}}$ that maximized the network's Fisher information (within the range $k_{\mathbf{w}} \in [1, 10]$; see Figure 6e). For the unstructured weights, we calculate the choice of μ that maximizes the network's average Fisher information over 1000 drawings of $\mathbf{w}$ from the log-normal distribution specified by μ (see Figure 6f).

Figures 6e and 6f reveal that the network is highly sensitive to the values of σ_P and σ_C. Figure 6e exhibits a bandlike structure and abrupt transitions in the value of $k_{\mathbf{w}}$, which maximizes Fisher information. This bandlike structure would most likely continue to form for smaller σ_P if we allowed $k_{\mathbf{w}} > 10$. One might expect that the bandlike structure is due to the artificial structure in the weights; however, we see that Figure 6f also exhibits these types of bands. Note that the regime of interest for us is when private variability is a smaller contribution to the total variability than the common variability. When this is the case, Figures 6e and 6f imply that a population of neurons will be best served by having a diverse set of synaptic weights, even if the weights amplify irrelevant signals.

Together, these results highlight how the introduction of the nonlinearity in the network reveals an intricate relationship among the amount of shared variability, private variability, and the optimal synaptic weight heterogeneity. Our observations that the network benefits from increased synaptic weight heterogeneity in the presence of common noise are predicated on the size of the network (see Figures 4a and 4b and 6a and 6b) and the amount of private and shared variability (see Figures 4c, 6c, and 6d). In particular, when shared variability is the more significant contribution to the overall variability, the coding performance of the network benefits from increased heterogeneity, whether the weights are structured or unstructured (see Figures 6e and 6f). This implies that in contrast to the linear network, there exist regimes where increasing the synaptic weight heterogeneity beyond a point will harm coding ability (see Figures 4d and 4e and 5a and 5b), demonstrating that there is a trade-off between the benefits of synaptic weight heterogeneity and the amplification of common noise it may introduce.

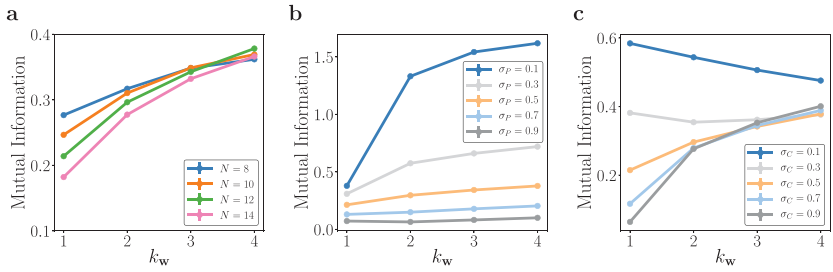

Figure 7: Mutual information computed by applying the KSG estimator on data simulated from the network with quadratic nonlinearity and structured weights. The estimates consist of averages over 100 data sets, each containing 100,000 samples. Standard error bars are smaller than the size of the markers. (a) Mutual information as a function of common noise weight heterogeneity for various population sizes N. We consider smaller N than in the case of Fisher information as computation time becomes prohibitive for larger dimensionalities. Here, $\sigma_P = \sigma_C = 0.5$. (b) The behavior of mutual information for various choices of σ_P, while $\sigma_C = 0.5$. (c) The behavior of mutual information for various choices of σ_C, while $\sigma_P = 0.5$.

3.2.2 Mutual Information. When the network possesses a quadratic nonlinearity, the mutual information $I[s, \mathbf{r}]$ is far less tractable than for the linear case. Therefore, we computed the mutual information numerically on data simulated from the network, using an estimator built on k-nearest neighbor statistics (Kraskov et al., 2004). We refer to this estimator as the KSG estimator.

We applied the KSG estimator to 100 unique data sets, each containing 100,000 samples drawn from the linear-nonlinear network. We then estimated the mutual information within each of the 100 data sets. The computational bottleneck for the KSG estimator lies in finding nearest neighbors in a *kd*-tree, which becomes prohibitive for large dimensions ($\sim$20), so we considered much smaller population sizes than in the case of Fisher information. Furthermore, the KSG estimator encountered difficulties when samples became too noisy, so we limited our analysis to smaller values of (σ_P, σ_C). Due to these constraints, we are only able to probe the finite network behavior of the mutual information.

Our results for the structured weights are shown in Figure 7. When utilizing estimators of mutual information from data, caution should be taken before comparing across different dimensions due to bias in the KSG estimator (Gao, Ver Steeg, & Galstyan, 2015). Thus, we restrict our observations to within a specified population size. First, we evaluated the mutual information for various population sizes ($N = 8, 10, 12, 14$) in the case where

$\sigma_C = \sigma_P = 0.5$. Observe that, as before, the mutual information increases with larger weight heterogeneity ($k_{\mathbf{w}}$; see Figure 7a). The improvement in information occurs for all four population sizes.

Decreasing the private variability increases mutual information (see Figure 7b). However, the network sees a greater increase in information with diverse weighting when σ_P is small. This is consistent with the small σ_P regime highlighted in Figure 4c: the smaller the private variability, the more the network benefits from larger synaptic weight heterogeneity. Similarly, decreasing the common variability increases mutual information (see Figure 7c). If the common variability is small enough (e.g., $\sigma_C = 1$), then larger $k_{\mathbf{w}}$ harms the encoding. Thus, when the common noise is small enough, the amplification of noise that results when $k_{\mathbf{w}}$ is increased harms the network's encoding. It is only when the common variability becomes the dominant contribution to the variability that the diversification provided by larger $k_{\mathbf{w}}$ improves the mutual information.

As for the unstructured weights, we calculated the mutual information $I[s, \mathbf{r}]$ over 100 synaptic weight distributions drawn from the aforementioned log-normal distribution. For each synaptic weight distribution, we applied the KSG estimator to 100 unique data sets, each consisting of 10,000 samples. Thus, the mutual information estimate for a given network was computed by averaging over the individual estimates across the 100 data sets. With this procedure, we explored how the mutual information behaves as a function of the private noise variability, common noise variability, and mean of the log-normal distribution.

Similar to the normalized Fisher information, we present the normalized mutual information as a function of the private and common variances (see Figure 8). For a given σ_P or σ_C, the mutual information is calculated across a range of $\mu \in [-1, 1]$. The normalized mutual information is obtained by dividing each individual mutual information by the maximum value across the μ. Thus, for a given σ_P, the value of μ whose normalized mutual information is 1 specifies the log-normal distribution that maximizes the network's encoding performance. As private variability increases, the network benefits more greatly benefits diverse weighting (larger μ; see Figure 8a). As common variability increases, the network once again prefers more diverse weighting. If the common variability is small enough, however, the network is better suited to homogeneous weights (see Figure 8b). Therefore, the analysis using the unstructured weights largely corroborates our findings for the structured weights shown in Figure 7.

Thus, these results highlight that there exist regimes where neural coding, as measured by the Shannon mutual information, benefits from increased synaptic weight heterogeneity. Furthermore, similar to the case of the linear Fisher information, the improvement in coding occurs more significantly when shared variability is large relative to private variability.

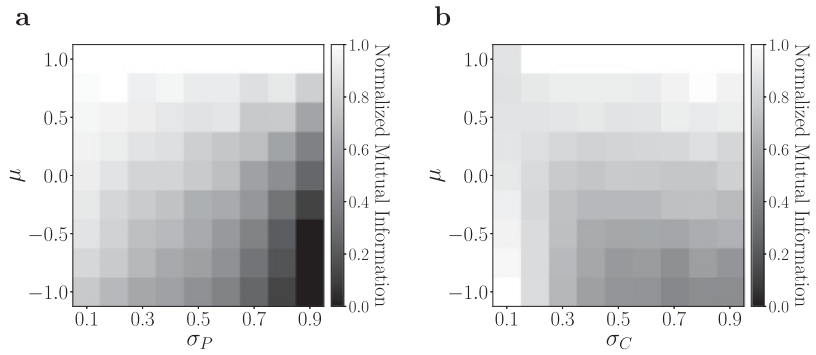

Figure 8: Normalized mutual information for common and private variability. For a given μ, 100 networks were created by drawing common noise weights $\mathbf{w}$ from the corresponding log-normal distribution. The mutual information shown is the average across the 100 networks. For a specified network, the mutual information was calculated by averaging KSG estimates over 100 simulated data sets, each containing 10,000 samples. Finally, for a choice of (σ_P, σ_C), mutual information is normalized to the maximum across values of μ. (a) Normalized mutual information as a function of μ and private variability ($\sigma_C = 0.5$). (b) Normalized mutual information as a function of μ and common variability ($\sigma_P = 0.5$).

4 Discussion

We have demonstrated in a simple model of neural activity that if synaptic weighting of common noise inputs is broad and heterogeneous, coding fidelity is actually improved despite inadvertent amplification of common noise inputs. We showed that for squaring nonlinearities, there exists a regime of heterogeneous weights for which coding fidelity is maximized. We also found that the relationship between the magnitude of private and shared variability is vital for determining the ideal amount of synaptic heterogeneity. In neural circuits where shared variability is dominant, as has been reported in some parts of the cortex (Deweese & Zador, 2004), larger weight heterogeneity results in better coding performance (see Figure 6e).

Why are we afforded improved neural coding under increased synaptic weight heterogeneity? An increase in heterogeneity, as we have defined it, ensures that the common noise is magnified in the network. At the same time, however, the structure of the correlated variability induced by the common noise is altered by increased heterogeneity. Previous work has demonstrated that the relationship between signal correlations and noise correlations is important in assessing decoding ability; for example, the sign rule states that noise correlations are beneficial if they are of opposite sign

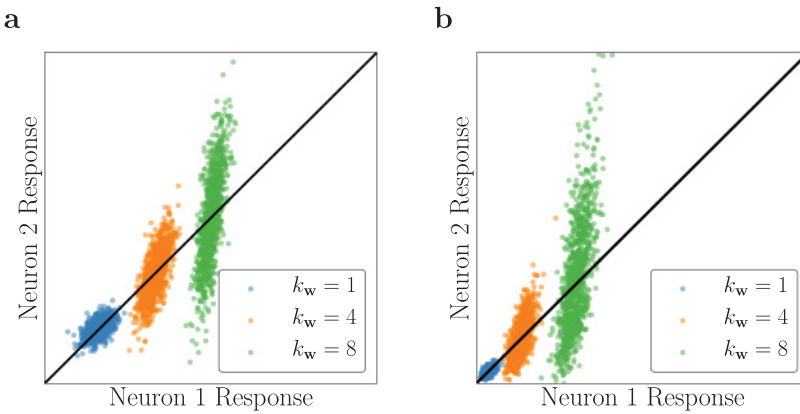

Figure 9: The benefits of increased synaptic weight heterogeneity. (a) The responses of a pair of neurons against the signal space, taken after the linear stage. Colors indicate different choices of $k_{\mathbf{w}}$ (while $k_{\mathbf{v}} = 1$). Each cloud contains 1000 sampled points. (b) Same as panel a, but responses are taken after the quadratic nonlinearity.

as the signal correlation (Hu et al., 2014). Geometrically, the sign rule is a consequence of the intuitive observation that decoding is easier when the noise correlations lie perpendicular to the signal manifold (Averbeck et al., 2006; Zylberberg, Cafaro, Turner, Shea-Brown, & Rieke, 2016; Montijn et al., 2016).

For example, consider the correlated activity for two neurons in the network against their signal space (see the black lines in Figures 9a and 9b) as a function of $k_{\mathbf{w}}$. Note that the signal space is linear. After the linear stage, the larger weight heterogeneity pushes the cloud of neural activity to lie more orthogonal to the signal space. At the same time, the variance becomes observably larger due to the magnification of the common noise (see Figure 9a). Importantly, note that the variability for $k_{\mathbf{w}} = 1$ lies parallel to the signal space, signifying the presence of differential correlations. The correlated variability after the nonlinear stage is similar in that orthogonality to the signal space increases with $k_{\mathbf{w}}$. There is a notable difference: squaring the linear stage ensures nonnegative activities, thereby limiting the response space. Thus, for large enough $k_{\mathbf{w}}$, the rectification manifests strongly enough that the network enters a regime where increased heterogeneity harms decoding. These figures only demonstrate the relationship between a pair of neurons, while the collective correlated variability structure ultimately dictates decoding performance. They do, however, shed light on how the distribution of synaptic weights can radically shape the common noise and thereby the overall structure of the shared variability.

The linear stage of the network constitutes a noisy projection of two signals (one of which is not useful to the network) in a high-dimensional space. Thus, we can assess the entire population by examining the relationship between the projecting vectors $\mathbf{v}$ and $\mathbf{w}$. We might expect that improved decoding occurs when these signals are farther apart in the N-dimensional space (Kanerva, 2009). For a chosen $k_{\mathbf{v}}$, this occurs as $k_{\mathbf{w}}$ is increased when the weights are structured. When the weights are unstructured, the average angle between the stimulus and weight vectors is large as either μ_v or μ_w increases. Increased heterogeneity implies access to a more diverse selection of weights, thus pushing the two signals apart. From this perspective, the nonlinear stage acts as a mapping on the high-dimensional representation. Given that no noise is added after the nonlinear processing stage in the networks, if the nonlinearities were one-to-one, the data processing inequality would ensure that the results from the linear stage would hold. But as we observed earlier, the nonlinear stage benefits from increased heterogeneity only in certain regimes. Thus, the behavior of the nonlinearity is important: the application of the quadratic nonlinearity restricts the high-dimensional space that the neural code can occupy, and thus limits the benefits of diverse synaptic weighting. Validating and characterizing these observations for other nonlinearities (such as an exponential nonlinearity or a squared rectified linear unit) and within the framework of a linear-nonlinear-Poisson cascade model will be interesting to pursue in future studies. For example, we performed a simple experiment numerically assessing the behavior of the linear Fisher information under an exponential nonlinearity. We observed that synaptic weight heterogeneity benefits coding, but information may saturate for a wide range of $k_{\mathbf{w}}$ (see appendix A.3). Thus, the choice of nonlinearity may affect the coding performance in the presence of common noise.

In this work, we considered the coding ability of a network in which a stimulus is corrupted by a single common noise source. However, cortical circuits receive many inputs and must likely contend with multiple common noise inputs. Thus, it is important to examine how our analysis changes as the number of inputs increases. Naively, the neural circuit could structure weights to collapse all common noise sources on a single subspace, but this strategy will fail if the circuit must perform multiple tasks (e.g., the circuit may be required to decode among many of the inputs using the same set of weights). Furthermore, there are brain regions in which the dimensionality is drastically reduced, such as cortex to striatum (10 to 1 reduction) or striatum to basal ganglia (300 to 1 reduction; Bar-Gad, Morris, & Bergman, 2003; Seger, 2008). In these cases, the number of inputs may scale with the size of the neural circuit. In such an underconstrained system, linear decoding will be unable to properly extract estimates of the relevant stimulus. This implies that linear Fisher information, which relies on a linear decoder, may be insufficient to judge the coding fidelity of these populations. Thus, future work could examine how the synaptic weight

distribution affects neural coding with multiple common noise inputs. This includes the case when the number of common noise sources is smaller than the population size or when they are of similar scale, the latter of which may require alternative coding strategies (Davenport, Duarte, Eldar, & Kutyniok, 2012; Garfinkle & Hillar, 2019).

It may seem unreasonable that the neural circuit possesses the ability to weight common noise inputs. However, excitatory neurons receive many excitatory synapses in circuits throughout the brain. Some subset of common inputs across a neural population will undoubtedly be irrelevant for the underlying neural computation, even if these signals are not strictly speaking "noise" and could be useful for other computations. Thus, these populations must contend with common noise sources contributing to their overall shared variability and potentially hampering their ability to encode a stimulus. Our work demonstrates that neural circuits, armed with a good set of synaptic weights, need not suffer adverse impacts due to inadvertently amplifying potential sources of common noise. Instead, broad, heterogeneous weighting ensures that common noise sources will project the signal and noise into a high-dimensional space in such a way that is beneficial for decoding. This observation is in agreement with recent work that explored the relationship between heterogeneous weighting and degrees of synaptic connectivity (Litwin-Kumar, Harris, Axel, Sompolinsky, & Abbott, 2017). Furthermore, synaptic input, irrelevant on one trial, may become the signal on the next: heterogeneous weighting provides a general, robust principle for neural circuits to follow.

We chose the simple network architecture in order to maintain analytic tractability, which allowed us to explore the rich patterns of behavior it exhibited. Our model is limited, however. It is worthwhile to assess how our qualitative conclusions hold with added complexity in the network. For example, interesting avenues to consider include the implementation of recurrence, spiking dynamics, and global fluctuations. In addition, these networks could also be equipped with varying degrees of sparsity and inhibitory connections. Importantly, the balance of excitation and inhibition in networks has been shown to be vital in decorrelating neural activity (Renart et al., 2010). Past work has explored how to approximate both information-theoretic quantities studied here in networks with some subset of these features (Beck, Bejjanki, & Pouget, 2011; Yarrow, Challis, & Seriès, 2012). Thus, analyzing how common noise and synaptic weighting interact in more complex networks is of interest for future work.

We established correlated variability structure in the linear-nonlinear network by taking a linear combination of a common noise source and private noise sources (though our model ignores any noise potentially carried by the stimulus). This was sufficient to establish low-dimensional shared variability observed in neural circuits. As a consequence, our model as devised enforces stimulus-independent correlated variability. Recent work, however, has demonstrated that correlated variability is in fact stimulus

dependent. Such work used both phenomenological (Lin, Okun, Carandini, & Harris, 2015; Franke et al., 2016) and mechanistic (Zylberberg et al., 2016) models in producing fits to the stimulus-dependent correlated variability. These models all share a doubly stochastic noise structure, stemming from both additive and multiplicative sources of noise (Goris, Movshon, & Simoncelli, 2014). It is therefore worthwhile to fully examine how both additive and multiplicative modulation interact with synaptic weighting to influence neural coding. For example, Arandia-Romero et al. (2016) demonstrated that such additive and multiplicative modulation, modulated by overall population activity, can redirect information to specific neuronal assemblies, increasing information for some but decreasing it for others. Synaptic weight heterogeneity, attuned by plasticity, could serve as a mechanism for additive and multiplicative modulation, thereby gating information for specific assemblies.

A Appendix

A.1 Calculation of Fisher and Mutual Information Quantities.

A.1.1 Calculation of Fisher Information, Linear Stage. All variability after the linear stage is gaussian; thus, the Fisher information can be expressed as (Abbott & Dayan, 1999; Kay, 1993)

$$I_F(s) = \mathbf{f}'(s)^T \mathbf{\Sigma}^{-1}(s)\mathbf{f}'(s) + \frac{1}{2}\mathrm{Tr}\left[\mathbf{\Sigma}'(s)\mathbf{\Sigma}^{-1}(s)\mathbf{\Sigma}'(s)\mathbf{\Sigma}^{-1}(s)\right]. \tag{A.1}$$

Our immediate goal is to calculate $\mathbf{f}(s)$, the average response of the linear stage, and $\mathbf{\Sigma}$, the covariance between the responses. The output of the ith neuron after the linear stage is

$$\ell_i = v_i s + w_i \sigma_C \xi_C + \sigma_P \xi_{P,i}, \tag{A.2}$$

so that the average response as a function of s is

$$f_i(s) = \langle \ell_i \rangle = v_i s. \tag{A.3}$$

Thus,

$$\mathbf{f}(s) = \mathbf{v}s \Rightarrow \mathbf{f}'(s) = \mathbf{v}, \tag{A.4}$$

and

$$\langle \ell_i \ell_j \rangle = \langle (v_i s + w_i \sigma_C \xi_C + \sigma_P \xi_{P,i})(v_j s + w_j \sigma_C \xi_C + \sigma_P \xi_{P,j}) \rangle \tag{A.5}$$

$$= v_i v_j s^2 + w_i w_j \sigma_C^2 + \sigma_P^2 \delta_{ij}, \tag{A.6}$$

so that

$$\Sigma_{ij} = \langle \ell_i \ell_j \rangle - \langle \ell_i \rangle \langle \ell_j \rangle \tag{A.7}$$

$$= \sigma_P^2 \delta_{ij} + w_i w_j \sigma_C^2 \tag{A.8}$$

$$\Rightarrow \mathbf{\Sigma} = \sigma_P^2 \mathbf{I} + \sigma_C^2 \mathbf{w}\mathbf{w}^T. \tag{A.9}$$

Notice that the covariance matrix does not depend on s, so the second term in equation A.1 will vanish. We do, however, need the inverse covariance matrix for the first term:

$$\mathbf{\Sigma}^{-1} = \frac{1}{\sigma_P^2} \left(\mathbf{I} - \frac{\sigma_C^2}{\sigma_P^2 + \sigma_C^2 |\mathbf{w}|^2} \mathbf{w}\mathbf{w}^T \right). \tag{A.10}$$

Hence, the Fisher information is

$$I_F(s) = \frac{1}{\sigma_P^2} \mathbf{v}^T \left(\mathbf{I} - \frac{\sigma_C^2}{\sigma_P^2 + \sigma_C^2 |\mathbf{w}|^2} \mathbf{w}\mathbf{w}^T \right) \mathbf{v} \tag{A.11}$$

$$= \frac{1}{\sigma_P^2} \frac{\left(\sigma_P^2/\sigma_C^2\right) |\mathbf{v}|^2 + \left(|\mathbf{v}|^2 |\mathbf{w}|^2 - (\mathbf{v} \cdot \mathbf{w})^2 \right)}{(\sigma_P^2/\sigma_C^2) + |\mathbf{w}|^2}. \tag{A.12}$$

A.1.2 Calculation of Mutual Information, Linear Stage. The mutual information is given by

$$I[s, \boldsymbol{\ell}] = \int d\boldsymbol{\ell} ds P[s] P[\boldsymbol{\ell}|s] \log \frac{P[\boldsymbol{\ell}|s]}{P[\boldsymbol{\ell}]} \tag{A.13}$$

$$= H[\boldsymbol{\ell}] + \int ds P[s] \int d\boldsymbol{\ell} P[\boldsymbol{\ell}|s] \log P[\boldsymbol{\ell}|s]. \tag{A.14}$$

Note that $P[\boldsymbol{\ell}]$ and $P[\boldsymbol{\ell}|s]$ are both multivariate gaussians. The (differential) entropy of a multivariate gaussian random variable X with mean $\boldsymbol{\mu}$ and covariance $\mathbf{\Sigma}$ is given by

$$H[X] = \frac{1}{2} \log \left(\det \mathbf{\Sigma} \right) + \frac{N}{2} (1 + \log(2\pi)). \tag{A.15}$$

Therefore, by the gaussianity of the involved distributions,

$$P[\boldsymbol{\ell}|s] = \frac{1}{\sigma_P^{N-1} \sqrt{(2\pi)^N (\sigma_P^2 + \sigma_C^2 |\mathbf{w}|^2)}}$$

$$\times \exp \left[-\frac{1}{2\sigma_P^2} (\boldsymbol{\ell} - \mathbf{v}s)^T \left(\mathbf{I} - \frac{\sigma_C^2 \mathbf{w}\mathbf{w}^T}{\sigma_P^2 + \sigma_C^2 |\mathbf{w}|^2} \right) (\boldsymbol{\ell} - \mathbf{v}s) \right] \tag{A.16}$$

$$P[\boldsymbol{\ell}] = \frac{1}{\sqrt{(2\pi)^N \sigma_P^{2N-4}\kappa}} \exp\left[-\frac{1}{2}\boldsymbol{\ell}^T\left(\sigma_P^2\mathbf{I} + \sigma_S^2\mathbf{v}\mathbf{v}^T + \sigma_C^2\mathbf{w}\mathbf{w}^T\right)^{-1}\boldsymbol{\ell}\right], \tag{A.17}$$

where

$$\kappa = (\sigma_P^2 + \sigma_C^2|\mathbf{w}|^2)(\sigma_P^2 + \sigma_S^2|\mathbf{v}|^2) - \sigma_C^2\sigma_S^2(\mathbf{v}\cdot\mathbf{w})^2. \tag{A.18}$$

Thus,

$$H[\boldsymbol{\ell}] = \frac{1}{2}\log\left(\sigma_P^{2N-4}\kappa\right) + \frac{N}{2}(1 + \log(2\pi)) \tag{A.19}$$

and

$$\int d\boldsymbol{\ell} P[\boldsymbol{\ell}|s]\log P[\boldsymbol{\ell}|s] = -\frac{1}{2}\log(\sigma_P^{2N-2}(\sigma_P^2 + \sigma_C^2|\mathbf{w}|^2)) - \frac{N}{2}(1 + \log(2\pi)), \tag{A.20}$$

which is notably independent of s. Thus, the integral over s will marginalize away. We are left with

$$I[s, \boldsymbol{\ell}] = \frac{1}{2}\log\left(\frac{\kappa}{\sigma_P^2(\sigma_P^2 + \sigma_C^2|\mathbf{w}|^2)}\right) \tag{A.21}$$

$$= \frac{1}{2}\log\left(1 + \sigma_S^2 I_F(s)\right). \tag{A.22}$$

A.1.3 Calculation of Linear Fisher Information, Quadratic Nonlinearity. We repeat the calculation of the first section, but after the nonlinear stage. In this case, we consider a quadratic nonlinearity. Instead of the Fisher information, we calculate the linear Fisher information (since it is analytically tractable). The output of the network is

$$r_i = (v_i s + w_i\sigma_C\xi_C + \sigma_P\xi_{P,i})^2 \tag{A.23}$$

$$= v_i^2 s^2 + w_i^2\sigma_C^2\xi_C^2 + \sigma_P^2\xi_{P,i}^2 + 2sv_i w_i\sigma_C\xi_C + 2sv_i\sigma_P\xi_{P,i} + 2w_i\sigma_C\sigma_P\xi_C\xi_{P,i}. \tag{A.24}$$

Thus, the average is then

$$f_i(s) = \langle r_i\rangle = v_i^2 s^2 + w_i^2\sigma_C^2 + \sigma_P^2, \tag{A.25}$$

which implies

$$\langle r_i \rangle \langle r_j \rangle = (v_i^2 s^2 + w_i^2 \sigma_C^2 + \sigma_P^2)(v_j^2 s^2 + w_j^2 \sigma_C^2 + \sigma_P^2) \tag{A.26}$$

$$= \sigma_P^4 + s^2 \sigma_P^2 (v_i^2 + v_j^2) + \sigma_P^2 \sigma_C^2 (w_i^2 + w_j^2) + s^2 \sigma_C^2 (v_i^2 w_j^2 + v_j^2 w_i^2) + s^4 v_i^2 v_j^2 + \sigma_C^4 w_i^2 w_j^2. \tag{A.27}$$

Next, the covariate can be written as

$$\langle r_i r_j \rangle = \sigma_P^4 + s^2 \sigma_P^2 (v_i^2 + v_j^2) + \sigma_P^2 \sigma_C^2 (w_i^2 + w_j^2) + s^2 \sigma_C^2 (v_i^2 w_j^2 + v_j^2 w_i^2) + s^4 v_i^2 v_j^2 + 3\sigma_C^4 w_i^2 w_j^2 + 4s^2 \sigma_C^2 v_i v_j w_i w_j. \tag{A.28}$$

The off-diagonal terms of the covariance matrix are then

$$\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle = 2\sigma_C^4 w_i^2 w_j^2 + 4s^2 \sigma_C^2 v_i v_j w_i w_j. \tag{A.29}$$

Finally, the variance of r_i (the diagonal terms of the covariance matrix) is given by

$$\mathrm{Var}(r_i) = \langle r_i^2 \rangle - \langle r_i \rangle^2 \tag{A.30}$$

$$= 3\sigma_P^4 + 6s^2 \sigma_P^2 v_i^2 + 6\sigma_P^2 \sigma_C^2 w_i^2 + 6s^2 \sigma_C^2 v_i^2 w_i^2 + s^4 v_i^4 + 3\sigma_C^4 w_i^4 - \left(v_i^2 s^2 + w_i^2 \sigma_C^2 + \sigma_P^2\right)^2 \tag{A.31}$$

$$= 2\sigma_C^4 w_i^4 + 4s^2 \sigma_C^2 v_i^2 w_i^2 + 2\sigma_P^4 + 4s^2 \sigma_P^2 v_i^2 + 4\sigma_P^2 \sigma_C^2 w_i^2. \tag{A.32}$$

Thus, the total covariance, which takes the variance into consideration, is

$$\Sigma_{ij} = \delta_{ij} \left(2\sigma_P^4 + 4\sigma_P^2 (s^2 v_i^2 + \sigma_C^2 w_i^2)\right) + 4s^2 \sigma_C^2 v_i v_j w_i w_j + 2\sigma_C^4 w_i^2 w_j^2. \tag{A.33}$$

In vector notation, this can be expressed as

$$\mathbf{\Sigma} = 2\sigma_P^4 \mathbf{I} + 4\sigma_P^2 s^2 \mathrm{diag}(\mathbf{V}) + 4\sigma_P^2 \sigma_C^2 \mathrm{diag}(\mathbf{W}) + 4s^2 \sigma_C^2 \mathbf{X}\mathbf{X}^T + 2\sigma_C^4 \mathbf{W}\mathbf{W}^T, \tag{A.34}$$

where

$$\mathbf{V} = \mathbf{v} \odot \mathbf{v}, \tag{A.35}$$

$$\mathbf{W} = \mathbf{w} \odot \mathbf{w}, \tag{A.36}$$

$$\mathbf{X} = \mathbf{v} \odot \mathbf{w}, \tag{A.37}$$

where $\odot$ indicates the Hadamard product (element-wise product). We now proceed to the linear Fisher information:

$$I_{LFI}(s) = \mathbf{f}'(s)^T \boldsymbol{\Sigma}(s)^{-1} \mathbf{f}'(s). \tag{A.38}$$

We start by calculating the inverse covariance matrix, which we will achieve with repeated applications of the Sherman-Morrison formula (Sherman & Morrison, 1950). We can write

$$\boldsymbol{\Sigma}^{-1} = (\mathbf{M} + 2\sigma_C^4 \mathbf{W}\mathbf{W}^T)^{-1} \tag{A.39}$$

$$= \mathbf{M}^{-1} - \frac{\mathbf{M}^{-1}(2\sigma_C^4 \mathbf{W}\mathbf{W}^T)\mathbf{M}^{-1}}{1 + 2\sigma_C^4 \mathbf{W}^T \mathbf{M}^{-1} \mathbf{W}} \tag{A.40}$$

$$= \mathbf{M}^{-1} - \frac{2\sigma_C^4}{1 + 2\sigma_C^4 \mathbf{W}^T \mathbf{M}^{-1} \mathbf{W}} \mathbf{M}^{-1} \mathbf{W}\mathbf{W}^T \mathbf{M}^{-1}, \tag{A.41}$$

where

$$\begin{aligned} \mathbf{M}^{-1} \equiv & \left(2\sigma_P^4 + 4\sigma_P^2 s^2 v_i^2 + 4\sigma_P^2 \sigma_C^2 w_i^2\right)^{-1} \delta_{ij} \\ & - \frac{s^2 \sigma_C^2}{\sigma_P^4 + 2s^2 \sigma_C^2 \sigma_P^2 \sum_i \dfrac{v_i^2 w_i^2}{\sigma_P^2 + 2s^2 v_i^2 + 2\sigma_C^2 w_i^2}} \\ & \times \frac{v_i v_j w_i w_j}{\left(\sigma_P^2 + 2s^2 v_i^2 + 2\sigma_C^2 w_i^2\right)\left(\sigma_P^2 + 2s^2 v_j^2 + 2\sigma_C^2 w_j^2\right)}. \end{aligned} \tag{A.42}$$

Note that

$$\mathbf{f}'(s) = 2s\mathbf{V}, \tag{A.43}$$

so the Fisher information is

$$I_{LFI}(s) = 4s^2 \left(\mathbf{V}^T \mathbf{M}^{-1} \mathbf{V} - \frac{2\sigma_C^4}{1 + 2\sigma_C^4 \mathbf{W}^T \mathbf{M}^{-1} \mathbf{W}} \mathbf{V}^T \mathbf{M}^{-1} \mathbf{W}\mathbf{W}^T \mathbf{M}^{-1} \mathbf{V} \right) \tag{A.44}$$

$$= 4s^2 \left(\mathbf{V}^T \mathbf{M}^{-1} \mathbf{V} - \frac{2\sigma_C^4}{1 + 2\sigma_C^4 \mathbf{W}^T \mathbf{M}^{-1} \mathbf{W}} \left(\mathbf{V}^T \mathbf{M}^{-1} \mathbf{W}\right)^2 \right). \tag{A.45}$$

To facilitate the matrix multiplications, we will define the following notation:

$$\{v, w\}_{m,n} = \sum_i \frac{v_i^m w_i^n}{\sigma_P^2 + 2s^2 v_i^2 + 2\sigma_C^2 w_i^2}. \tag{A.46}$$

Thus,

$$\mathbf{V}^T\mathbf{M}^{-1}\mathbf{V} = \frac{1}{2\sigma_P^2}\sum_i \frac{v_i^4}{\sigma_P^2 + 2s^2v_i^2 + 2\sigma_C^2 w_i^2} - \frac{s^2\sigma_C^2}{\sigma_P^4 + 2s^2\sigma_C^2\sigma_P^2\{v,w\}_{2,2}}\left(\sum_i \frac{v_i^3 w_i}{\sigma_P^2 + 2s^2v_i^2 + 2\sigma_C^2 w_i^2}\right)^2 \tag{A.47}$$

$$= \frac{1}{2\sigma_P^2}\{v,w\}_{4,0} - \frac{s^2\sigma_C^2}{\sigma_P^4 + 2s^2\sigma_C^2\sigma_P^2\{v,w\}_{2,2}}\{v,w\}_{3,1}^2 . \tag{A.48}$$

Furthermore,

$$\mathbf{W}^T\mathbf{M}^{-1}\mathbf{W} = \frac{1}{2\sigma_P^2}\{v,w\}_{0,4} - \frac{s^2\sigma_C^2}{\sigma_P^4 + 2s^2\sigma_C^2\sigma_P^2\{v,w\}_{2,2}}\{v,w\}_{1,3}^2 \tag{A.49}$$

and, finally,

$$\mathbf{V}^T\mathbf{M}^{-1}\mathbf{W} = \frac{1}{2\sigma_P^2}\{v,w\}_{2,2} - \frac{s^2\sigma_C^2}{\sigma_P^4 + 2s^2\sigma_C^2\sigma_P^2\{v,w\}_{2,2}}\{v,w\}_{1,3}\{v,w\}_{3,1} . \tag{A.50}$$

Inserting this expression into equation A.45 and simplifying, we can write the Fisher information as

$$I_{LFI}(s) = 4s^2\left(\frac{1}{\sigma_P^2}\{v,w\}_{4,0} - \frac{2s^2\sigma_C^2}{\sigma_P^2+2s^2\sigma_P^2\sigma_C^2\{v,w\}_{2,2}}\{v,w\}_{3,1}^2 + \frac{\sigma_P^2\sigma_C^4\{v,w\}_{2,2}+2s^2\sigma_C^6(\{v,w\}_{2,2}-2\{v,w\}_{1,3}\{v,w\}_{3,1})}{\sigma_P^4+\sigma_P^2(\sigma_C^4\{v,w\}_{0,4}+2s^2\sigma_C^2\{v,w\}_{2,2})+2s^2\sigma_C^6(\{v,w\}_{0,4}\{v,w\}_{2,2}-2\{v,w\}_{1,3}^2)}\right). \tag{A.51}$$

A.2 Information Saturation and Differential Correlations. In section 3.2.1, we observed that the Fisher information saturates in particular instances of the nonlinear network. Specifically, for the nonlinear network, Fisher information saturates for $k_w = 1$ and $k_w = 2$, but not for $k_w > 3$. Additionally, Fisher information saturates for $k_w \sim O(N)$. To understand why we observe saturation in some cases and not others, it is helpful to examine the eigenspectrum of the covariance matrix $\mathbf{\Sigma}$ describing the neural responses. Here, we rely on an analysis in the supplement of Moreno-Bote et al. (2014).

The linear Fisher information can be written in terms of the eigenspectrum of $\mathbf{\Sigma}$ as

$$I_{LFI} = \mathbf{f}'^T \boldsymbol{\Sigma}^{-1} \mathbf{f}' \tag{A.52}$$

$$= \mathbf{f}'^T \mathbf{f}' \sum_k \frac{\cos^2 \theta_k}{\sigma_k^2}, \tag{A.53}$$

where σ_k^2 is the kth eigenvalue and θ_k is the angle between the kth eigenvector and $\mathbf{f}'$. We consider the cases in which I_{LFI} saturates with the population size N. First, note that the squared norm of the tuning curve derivative $\mathbf{f}'^T \mathbf{f}'$ will scale as $O(N)$, since there are N terms in the sum. This implies that the summation must shrink at least as fast as $O(1/N)$ for information to saturate. This implies that any eigenvalues scaling as $O(1)$ must have their corresponding cosineangles shrink faster than $O(1/N)$. If there are $O(N)$ such eigenvalues, they must shrink faster than $O(1/N^2)$.

In the case of $k_w = 1$, one eigenvalue grows as $O(N)$ while the others remain constant (see Figure 10a, left). Meanwhile, the cosine-angles of the constant eigenvalues are effectively zero. This case is the easiest to understand: the principal eigenvector aligns with $\mathbf{f}'$ while all other directions are effectively orthogonal to $\mathbf{f}'$. For $k_w \geq 1$, however, two eigenvalues grow as $O(N)$ while the others grow as $O(1)$ (see Figure 10a, middle and right). In this case, the behavior of the cosine-angles corresponding to the constant growth eigenvalues varies depending on k_w.

As in Moreno-Bote et al. (2014) we split up equation A.53 into two groups: those with eigenvalues that scale as $O(N)$, denoted by the set S_N, and those that scale as $O(1)$, denoted by the set S_1:

$$I_{LFI} = \mathbf{f}'^T \mathbf{f} \sum_{m \in S_N} \frac{\cos^2 \theta_m}{\sigma_m^2} + \mathbf{f}'^T \mathbf{f}' \sum_{n \in S_1} \frac{\cos^2 \theta_n}{\sigma_n^2}. \tag{A.54}$$

The left sum contains one term when $k_w = 1$ and two terms when $k_w > 1$. Information saturation is dictated by the right sum, which we call R_{k_w}:

$$R_{k_w} = \sum_{n \in S_1} \frac{\cos^2 \theta_n}{\sigma_n^2}. \tag{A.55}$$

The addends of R_{k_w} correspond to the $O(1)$ eigenvalues, whose eigenvectors must have cosine-angles that vanish more quickly than $O(1/N)$ since there are $O(N)$ such eigenvalues. As expected, for $k_w = 1$, R_1 quickly vanishes (see Figure 10a: gray line). We observe similar behavior for $k_w = 2$: the summation R_2 eventually vanishes as well (see Figure 10b: red line). However, for $k_w > 2$, this no longer occurs: the cosine-angles scale to zero slowly enough that R_3 approaches a constant value (thereby preventing information saturation). Thus, going to larger k_w ensures that the majority of the eigenvectors of $\boldsymbol{\Sigma}$ do not become orthogonal to $\mathbf{f}'$ quickly enough for information saturation to occur.

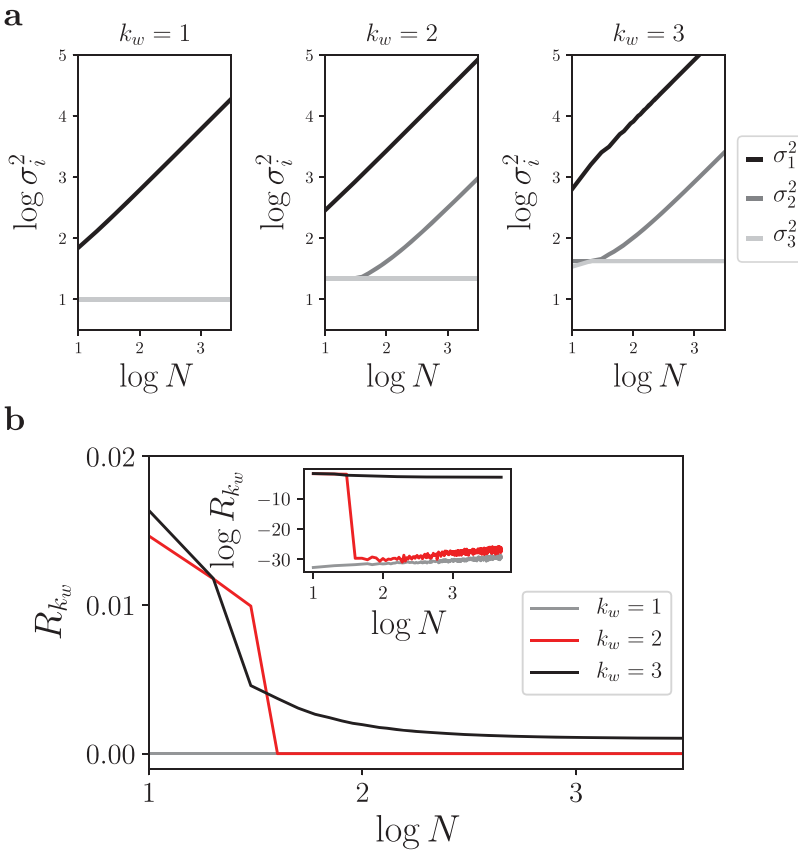

Figure 10: Characterizing the scaling of the eigenvalues and the shrinking of the cosine-angles for the nonlinear stage covariance. (a) Behavior of the largest three eigenvalues σ_1^2, σ_2^2, and σ_3^2 for the cases of $k_w = 1, 2, 3$. The aspect ratio is chosen so that unit steps on each axis appear of equal length. (b) The behavior of cosine-angle sum R_i corresponding to the constant-growth eigenvalues, for each of $k_w = 1, 2, 3$. The inset depicts the same curves, but on a log-log scale.

In the case of $k_w \sim O(N)$, however, the behavior of the covariance matrix is different. Recall that the covariance matrix takes on the form

$$\mathbf{\Sigma} = 2\sigma_P^4\mathbf{I} + 4\sigma_P^2 s^2 \mathrm{diag}(\mathbf{V}) + 4\sigma_P^2\sigma_C^2\mathrm{diag}(\mathbf{W}) + 4s^2\sigma_C^2\mathbf{X}\mathbf{X}^T + 2\sigma_C^4\mathbf{W}\mathbf{W}^T. \tag{A.56}$$

The dominant contribution to the covariance matrix is $2\sigma_C^4\mathbf{W}\mathbf{W}^T$. Thus, the scaling of the trace of $\mathbf{\Sigma}$ is

$$\mathrm{Tr}[\boldsymbol{\Sigma}] \sim \mathrm{Tr}[\mathbf{W}\mathbf{W}^T] = \mathrm{Tr}[(\mathbf{w} \odot \mathbf{w})(\mathbf{w} \odot \mathbf{w})^T]. \tag{A.57}$$

$$= (\mathbf{w} \odot \mathbf{w})^T (\mathbf{w} \odot \mathbf{w}) \tag{A.58}$$

$$\sim \sum_{i=1}^{N} \left(i^2\right)^2 \sim O(N^5). \tag{A.59}$$

Since the trace of the covariance matrix is equal to the sum of the eigenvalues, some subset of the eigenvalues can scale as $O(N^5)$ as well. In fact, all eigenvalues scale at least as $O(N)$, with the largest eigenvalue scaling as $O(N^5)$. In this scenario, the Fisher information must saturate because the cosine-angle can at most scale to a constant. In plainer terms, the variances of the covariance matrix scale so quickly that the differential correlation direction is irrelevant. We interpret this behavior as the neurons simply exhibiting too much variance for any meaningful decoding to occur. Note, however, that the saturation can be avoided if the behavior of $\mathbf{f}'$, which we assumed scales as $O(N)$, instead scales more quickly. This can occur, for example, when $k_v \sim O(N)$. However, it is unreasonable to expect that the synaptic weights of a neural circuit scale with the population size, making this scenario biologically implausible.

A.3 Linear Fisher Information under an Exponential Nonlinearity. The application of an exponential nonlinearity to the output of the linear stage $g_i(\ell_i) = \exp(\ell_i)$ implies that the output of the network $\mathbf{r} = \mathbf{g}(\boldsymbol{\ell})$ follows a multivariate log-normal distribution (since the linear stage is gaussian). The linear stage is described by the distribution

$$\boldsymbol{\ell} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}^L), \tag{A.60}$$

$$\boldsymbol{\mu} = \mathbf{v}s, \tag{A.61}$$

$$\boldsymbol{\Sigma}^L = \sigma_P^2 \mathbf{I} + \sigma_C^2 \mathbf{w}\mathbf{w}^T. \tag{A.62}$$

The multivariate log-normal distribution has first- and second-order statistics given by

$$\mathrm{E}[\mathbf{r}]_i = \exp\left[\mu_i + \frac{1}{2}\Sigma_{ii}^L\right], \tag{A.63}$$

$$\mathrm{Var}\left[\mathbf{r}_{ij}\right] = \exp\left[\mu_i + \mu_j + \frac{1}{2}\left(\Sigma_{ii}^L + \Sigma_{jj}^L\right)\right]\left(\exp(\Sigma_{ij}^L) - 1\right). \tag{A.64}$$

Thus, the mean activity and its derivative with respect to s are given by

$$f_i(s) = \exp\left[\frac{1}{2}\sigma_P^2 + v_i s + \frac{1}{2}\sigma_C^2 w_i^2\right], \tag{A.65}$$

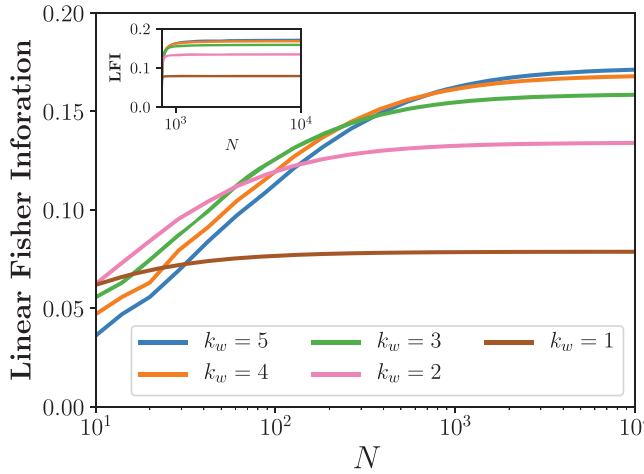

Figure 11: The behavior of linear Fisher information for an exponential nonlinearity as a function of population size. Colors denote different choices of k_w. Inset shows the same plot but on a regular scale.

$$f_i'(s) = v_i \exp\left[\frac{1}{2}\sigma_P^2 + v_i s + \frac{1}{2}\sigma_C^2 w_i^2\right]. \tag{A.66}$$

These equations provide us the tools to calculate the linear Fisher information. The inversion of the covariance matrix (see equation A.64) is not tractable, but we can proceed numerically.

We calculated the linear Fisher information numerically under the same conditions as in Figure 4a, but with $k_w = 1, \ldots, 5$ and for a wider range of population sizes. In Figure 11, we plot the linear Fisher information as a function of N for these choices of k_w. We observe that for large enough N, synaptic weight heterogeneity results in improved coding performance. However, we also observe what appears to be saturation of the Fisher information. Since we cannot write the Fisher information as a function of N, we cannot validate this observation analytically. This does, however, suggest that the choice of nonlinearity can have a dramatic impact on the behavior of the linear Fisher information.

Acknowledgments

We thank Ruben Coen-Cagli for useful discussions. P.S.S. was supported by the Department of Defense through the National Defense Science and Engineering Graduate Fellowship Program. J.A.L. was supported through the Lawrence Berkeley National Laboratory-internal LDRD "Deep Learning for Science" led by Prabhat. M.R.D. was supported in part by the U.S.

Army Research Laboratory and the U.S. Army Research Office under Contract No. W911NF-13-1-0390.

References

Abbott, L. F., & Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural Computation*, *11*(1), 91–101.

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*(2), 284–299.

Arandia-Romero, I., Tanabe, S., Drugowitsch, J., Kohn, A., & Moreno-Bote, R. (2016). Multiplicative and additive modulation of neuronal tuning with population activity affects encoded information. *Neuron*, *89*(6), 1305–1316.

Arieli, A., Sterkin, A., Grinvald, A., & Aertsen, A. (1996). Dynamics of ongoing activity: Explanation of the large variability in evoked cortical responses. *Science*, *273*(5283), 1868–1871.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, *7*(5), 358.

Averbeck, B. B., & Lee, D. (2006). Effects of noise correlations on information encoding and decoding. *Journal of Neurophysiology*, *95*(6), 3633–3644.

Bar-Gad, I., Morris, G., & Bergman, H. (2003). Information processing, dimensionality reduction and reinforcement learning in the basal ganglia. *Progress in Neurobiology*, *71*(6), 439–473.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*, 217–234.

Beck, J., Bejjanki, V. R., & Pouget, A. (2011). Insights from a simple expression for linear Fisher information in a recurrently connected population of spiking neurons. *Neural Computation*, *23*(6), 1484–1502.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., & Pouget, A. (2012). Not noisy, just wrong: The role of suboptimal inference in behavioral variability. *Neuron*, *74*(1), 30–39.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research*, *37*(23), 3327–3338.

Brinkman, B. A., Weber, A. I., Rieke, F., & Shea-Brown, E. (2016). How do efficient coding strategies depend on origins of noise in neural circuits? *PLOS Computational Biology*, *12*(10), e1005150.

Brunel, N., & Nadal, J.-P. (1998). Mutual information, Fisher information, and population coding. *Neural Computation*, *10*(7), 1731–1757.

Cafaro, J., & Rieke, F. (2010). Noise correlations improve response fidelity and stimulus encoding. *Nature*, *468*(7326), 964.

Cohen, M. R., & Kohn, A. (2011). Measuring and interpreting neuronal correlations. *Nature Neuroscience*, *14*(7), 811.

Cohen, M. R., & Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature Neuroscience*, *12*(12), 1594.

Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. Hoboken, NJ: Wiley.

Davenport, M. A., Duarte, M. F., Eldar, Y. C., & Kutyniok, G. (2012). Introduction to compressed sensing. In Y. Eldar& G. Kutyniok (Eds.), *Compressed sensing: Theory and applications* (pp. 1–64). Cambridge: Cambridge University Press.

Deweese, M. R., & Zador, A. M. (2004). Shared and private variability in the auditory cortex. *Journal of Neurophysiology*, *92*(3), 1840–1855.

Ecker, A. S., Berens, P., Tolias, A. S., & Bethge, M. (2011). The effect of noise correlations in populations of diversely tuned neurons. *Journal of Neuroscience*, *31*(40), 14272–14283.

Emerson, R. C., Korenberg, M. J., & Citron, M. C. (1992). Identification of complex-cell intensive nonlinearities in a cascade model of cat visual cortex. *Biological Cybernetics*, *66*(4), 291–300.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292.

Franke, F., Fiscella, M., Sevelev, M., Roska, B., Hierlemann, A., & da Silveira, R. A. (2016). Structures of neural correlation and how they favor coding. *Neuron*, *89*(2), 409–422.

Gao, S., Ver Steeg, G., & Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. In *Proceedings of the Eighteenth International Conference onArtificial Intelligence and Statistics* (pp. 277–286).

Garfinkle, C. J., & <, C. J., (2019). On the uniqueness and stability of dictionaries for sparse representation of noisy signals. *IEEE Transactions on Signal Processing*, *67*(23), 5884–5892.

Goris, R. L., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, *17*(6), 858.

Hu, Y., Zylberberg, J., & Shea-Brown, E. (2014). The sign rule and beyond: Boundary effects, flexibility, and noise correlations in neural population codes. *PLOS Computational Biology*, *10*(2), e1003469.

Iyer, R., Menon, V., Buice, M., Koch, C., & Mihalas, S. (2013). The influence of synaptic weight distribution on neuronal population dynamics. *PLOS Computational Biology*, *9*(10), e1003248.

Kafashan, M., Jaffe, A., Chettih, S. N., Nogueira, R., Arandia-Romero, I., Harvey, C. D., Drugowitsch, J. (2020). *Scaling of information in large neural populations reveals signatures of information-limiting correlations.* bioRxiv:2020.01.10.90217.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representation with high-dimensional random vectors. *Cognitive Computation*, *1*(2), 139–159.

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

Karklin, Y., & Simoncelli, E. P. (2011). Efficient coding of natural images with a population of noisy linear-nonlinear neurons. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems*, *24* (pp. 999–1007). Red Hook, NY: Curran.

Kay, S. M. (1993). *Fundamentals of statistical signal processing*. Upper Saddle River, NJ: Prentice Hall.

Kohn, A., Coen-Cagli, R., Kanitscheider, I., & Pouget, A. (2016). Correlations and neuronal population information. *Annual Review of Neuroscience*, *39*, 237–256.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, *69*(6), 066138.

Kulkarni, J. E., & Paninski, L. (2007). Common-input models for multiple neural spike-train data. *Network: Computation in Neural Systems*, *18*(4), 375–407.

Lin, I.-C., Okun, M., Carandini, M., & Harris, K. D. (2015). The nature of shared cortical variability. *Neuron*, *87*(3), 644–656.

Litwin-Kumar, A., Harris, K. D., Axel, R., Sompolinsky, H., & Abbott, L. (2017). Optimal degrees of synaptic connectivity. *Neuron*, *93*(5), 1153–1164.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*(11), 1432.

Montijn, J. S., Liu, R. G., Aschner, A., Kohn, A., Latham, P. E., & Pouget, A. (2019). *Strong information-limiting correlations in early visual areas*. bioRxiv:842724.

Montijn, J. S., Meijer, G. T., Lansink, C. S., & Pennartz, C. M. (2016). Population-level neural codes are robust to single-neuron variability from a multidimensional coding perspective. *Cell Reports*, *16*(9), 2486–2498.

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., & Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, *17*(10), 1410.

Nogueira, R., Peltier, N. E., Anzai, A., DeAngelis, G. C., Martínez-Trujillo, J., & Moreno-Bote, R. (2020). The effects of population tuning and trial-by-trial variability on information encoding and behavior. *Journal of Neuroscience*, *40*(5), 1066–1083.

Pagan, M., Simoncelli, E. P., & Rust, N. C. (2016). Neural quadratic discriminant analysis: Nonlinear decoding with V1-like computation. *Neural Computation*, *28*(11), 2291–2319.

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, *15*(4), 243–262.

Pillow, J. W., Paninski, L., Uzzell, V. J., Simoncelli, E. P., & Chichilnisky, E. (2005). Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *Journal of Neuroscience*, *25*(47), 11003–11013.

Renart, A., De La Rocha, J., Bartho, P., Hollender, L., Parga, N., Reyes, A., & Harris, K. D. (2010). The asynchronous state in cortical circuits. *Science*, *327*(5965), 587–590.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. S. (1999). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.

Sakai, K., & Tanaka, S. (2000). Spatial pooling in the second-order spatial structure of cortical complex cells. *Vision Research*, *40*(7), 855–871.

Sargent, P. B., Saviane, C., Nielsen, T. A., DiGregorio, D. A., & Silver, R. A. (2005). Rapid vesicular release, quantal variability, and spillover contribute to the precision and reliability of transmission at a glomerular synapse. *Journal of Neuroscience*, *25*(36), 8173–8187.

Seger, C. A. (2008). How do the basal ganglia contribute to categorization? Their roles in generalization, response selection, and learning via feedback. *Neuroscience and Biobehavioral Reviews*, *32*(2), 265–278.

Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: Implications for connectivity, computation, and information coding. *Journal of Neuroscience*, *18*(10), 3870–3896.

Shamir, M., & Sompolinsky, H. (2006). Implications of neuronal diversity on population coding. *Neural Computation*, *18*(8), 1951–1986.

Sherman, J., & Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Annals of Mathematical Statistics*, *21*(1), 124–127.

Sompolinsky, H., Yoon, H., Kang, K., & Shamir, M. (2001). Population coding in neuronal systems with correlated noise. *Physical Review E*, *64*(5), 051904.

Song, S., Sjöström, P. J., Reigl, M., Nelson, S., & Chklovskii, D. B. (2005). Highly nonrandom features of synaptic connectivity in local cortical circuits. *PLOS Biology*, *3*(3), e68.

Vidne, M., Ahmadian, Y., Shlens, J., Pillow, J. W., Kulkarni, J., Litke, A. M., Paninski, L. (2012). Modeling the impact of common noise inputs on the network activity of retinal ganglion cells. *Journal of Computational Neuroscience*, *33*(1), 97–121.

Wei, X.-X., & Stocker, A. A. (2016). Mutual information, Fisher information, and efficient coding. *Neural Computation*, *28*(2), 305–326.

Wilke, S. D., & Eurich, C. W. (2002). Representational accuracy of stochastic neural populations. *Neural Computation*, *14*(1), 155–189.

Wu, S., Nakahara, H., & Amari, S.-I. (2001). Population coding with correlation and an unfaithful model. *Neural Computation*, *13*(4), 775–797.

Yarrow, S., Challis, E., & Seriès, P. (2012). Fisher and Shannon information in finite neural populations. *Neural Computation*, *24*(7), 1740–1780.

Yoon, H., & Sompolinsky, H. (1999). The effect of correlations on the Fisher information of population codes. In M. J. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems, 11* (pp. 167–173). Cambridge, MA: MIT Press.

Zohary, E., Shadlen, M. N., & Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, *370*(6485), 140.

Zylberberg, J., Cafaro, J., Turner, M. H., Shea-Brown, E., & Rieke, F. (2016). Direction-selective circuits shape noise to ensure a precise population code. *Neuron*, *89*(2), 369–383.

Zylberberg, J., Pouget, A., Latham, P. E., & Shea-Brown, E. (2017). Robust information propagation through noisy neural circuits. *PLOS Computational Biology*, *13*(4), e1005497.

Received September 25, 2019; accepted February 24, 2020.