nature

https://doi.org/10.1038/s41586-020-2130-2

Supplementary information

Fundamental bounds on the fidelity of sensory cortical coding

In the format provided by the authors and unedited

Oleg I. Rumyantsev✉, Jérôme A. Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli✉ & Mark J. Schnitzer✉

Supplementary Mathematical Appendix

Contents

§1. Introduction

The main purpose of this Appendix is to provide the theoretical underpinnings for interpreting and estimating the noise eigenstructure and Fisher information in terms of basic models of area V1 neural network function. In **§2**, we use a simple linear-nonlinear (LN) model of V1 neurons' coding properties to show how the cells' noise covariance matrix depends on both their linear receptive fields and the nonlinearities in their input-output relationships. In **§3**, we derive the optimal fidelity, d'_{opt}, with which two different visual stimuli can be discriminated, based on the coding properties and noise characteristics of a neural ensemble. In **§4,** we use a simple model of cortical coding to show that d'_{opt} saturates as the number of cortical neurons in an ensemble recording, N_o , increases. We also examine how the saturation level depends on the input and output noise levels, and on network amplification factors. In **§5**, we explain why only a small number of the eigenvalues of the cortical noise covariance matrix rise linearly with increases in N_o. Specifically, the number of such eigenvalues relates to the number of linear dimensions

spanned by the space of visual cortical receptive fields, which in turn are confined to a low-dimensional manifold in the space of all visual stimuli that the photoreceptor neural ensemble can encode. In **§6**, we show that one can accurately estimate d'_{opt} using many fewer trials of visual stimulation than the number of recorded neurons. In essence, one can accurately determine the eigenvectors of the cortical noise covariance matrix with the largest eigenvalues, even when the number of trials is sufficently low that any individual neural correlation coefficient cannot be well estimated. The ability to accurately determine these eigenvectors and eigenvalues, which crucially contribute to the limits on information encoding, underlies our capacity to estimate d'_{opt} well despite a limited number of visual stimulation trials. In **§7** we analyze an independently acquired, publicly accessible dataset of large-scale neural Ca^{2+} activity taken in visual cortical area V1 of awake mice. This analysis reveals the same type of information saturation in real but not shuffled datasets that we observe in our own data, and thereby confirms our paper's major conclusions.

Several results in this Appendix build upon analyses performed in Ref. 1 (see **§8** for a list of references). Specifically, results in **§2, §3** and **§4** reproduce those in Ref. 1, though in **§4** we focus on a simpler linear Gaussian model that admits an exact analytical understanding of the relationship between signal, noise and network connectivity. In **§5,** we go beyond previous analyses, which focused on how information saturates with the number of observed neurons, so as to attain a conceptual understanding of how the number of eigenvalues of the cortical noise covariance matrix that grow with the number of neurons is related to the dimensionality of the space of cortical receptive fields. In **§6** we show conceptually why we can accurately estimate these eigenmodes of the noise covariance matrix in large neural ensembles despite using many fewer experimental trials than neurons and despite being unable to accurately estimate individual

correlation coefficients. Thus, this section goes well beyond prior work that considered the accurate estimation of Fisher information in the qualitatively distinct, classical experimental regime in which there were more trials than recorded cells. Our work lies in the regime in which there are more recorded cells than experimental trials, and our large-scale recordings enable accurate determinations of both the noise eigenstructure and information content of large neural ensemble codes.

§2. Signal and noise in a simple model of visual cortical area V1.

In this section we illustrate how the signals and noise correlations of a neural ensemble relate to its coding properties, and we show how the relationship between signal and noise limits the information conveyed. For conceptual clarity, we consider a basic linear-nonlinear (LN) model of single neuron coding properties in area V1, which provides a reasonable approximation of neural responses in V1 to simple visual stimuli[2].

Consider a two-layer circuit with N_s sensory neurons, or photoreceptors, in its first layer and N_c visual cortical neurons in the second layer. We respectively denote the activity patterns of these cells as $\mathbf{s}_a$, for $a = 1, \ldots, N_s$ and $\mathbf{r}_i$, for $i = 1, \ldots, N_c$, and the input-output relationship as:

$$\mathbf{r}_i = F_i\left(\mathbf{w}^i \cdot \mathbf{s} + \boldsymbol{\xi}_i^{\text{in}}\right) + \boldsymbol{\xi}_i^{\text{out}}. \qquad (1)$$

Here $\boldsymbol{\xi}^{\text{in}}$ and $\boldsymbol{\xi}^{\text{out}}$ are vectors, or patterns, of photoreceptor input noise and cortical output noise, respectively, modeled as Gaussian random vectors with zero means and covariance matrices $\boldsymbol{\Sigma}^{\text{in}}$ and $\boldsymbol{\Sigma}^{\text{out}}$. The vector $\mathbf{w}^i$ denotes the synaptic weights onto cortical neuron i, and F_i is a nonlinear (scalar) function.

What is the conditional distribution of cortical activity, $\mathbf{r}^A$, that results from a specific pattern of input activity, $\mathbf{s}^A$, in response to a stimulus? When the level of input noise is small, we can approximate equation (1) as:

$$\mathbf{r}_i^A \approx F_i\left(\mathbf{w}^i \cdot \mathbf{s}^A\right) + F'_i\left(\mathbf{w}^i \cdot \mathbf{s}^A\right)\boldsymbol{\xi}_i^{\text{in}} + \boldsymbol{\xi}_i^{\text{out}}. \quad (2)$$

In this approximation, the mean response to stimulus $\boldsymbol{A}$ is (using brackets to denote averages):

$$\left\langle \mathbf{r}_i^A \right\rangle = F_i\left(\mathbf{w}^i \cdot \mathbf{s}^A\right). \quad (3)$$

Further, as a sum of Gaussian-distributed random variables, the cortical responses to stimulus $\boldsymbol{A}$ in equation (2) are also Gaussian-distributed, with a noise covariance matrix that we denote $\boldsymbol{\Sigma}^{\text{A}}$. This matrix describes the fluctuations about the mean response to $\boldsymbol{A}$, and it is defined by

$$\boldsymbol{\Sigma}_{ij}^{\text{A}} \equiv \left\langle \delta\mathbf{r}_i^A \, \delta\mathbf{r}_j^A \right\rangle \quad (4)$$

where $\delta\mathbf{r}^A = \mathbf{r}^A - \left\langle \mathbf{r}_i^A \right\rangle$. Inserting (2) and (3) into (4) yields the matrix expression

$$\boldsymbol{\Sigma}^{\text{A}} = \mathbf{G}^A \mathbf{W} \, \boldsymbol{\Sigma}^{\text{in}} \mathbf{W}^T \mathbf{G}^{\mathbf{A}} + \boldsymbol{\Sigma}^{\text{out}}, \quad (5)$$

where $\mathbf{G}^A$ is a diagonal N_c by N_c matrix whose diagonal elements reflect the linear gains of each neuron around stimulus $\boldsymbol{A}$: $\mathbf{G}_{ij}^A = \delta_{ij} F'_i\left(\mathbf{w}^i \cdot \mathbf{s}^A\right)$, and F' is the derivative or slope of the neuronal nonlinearity F_i. $\mathbf{W}$ is an N_c by N_s synaptic weight matrix whose $i'th$ row is the vector of synaptic weights $\mathbf{w}^i$ onto cortical neuron i. Thus the cortical noise correlation $\boldsymbol{\Sigma}^{\text{A}}$ reflects both the propagation of input noise correlations $\boldsymbol{\xi}^{\text{in}}$ through the synaptic weight matrix $\mathbf{W}$ and nonlinearities F_i, in addition to the internally generated noise $\boldsymbol{\xi}^{\text{out}}$.

In this model, the only potential stimulus dependence of the cortical noise arises through the nonlinearity; different stimuli $\mathbf{s}^A$ will lead to different neurons i operating at different gain levels on their nonlinear response curve F_i, leading to different gain matrices $\mathbf{G}^A$. If all neurons operate at approximately similar gains, we can neglect the stimulus dependence of the noise correlations in this model. This is exactly true if every neuron is linear and has the same gain, *i.e.* $F_i(x) = gx$ for all i.

To appreciate the significance of equation (5), it is useful to simplify to the case in which the input noise is white, or uncorrelated: $\boldsymbol{\xi}^{\text{in}} = \sigma_{\text{in}}^2 \mathbf{I}$, where $\mathbf{I}$ is the $N_s \times N_s$ identity matrix.

This is the case in which the photoreceptor noise level is uniform. Equation (5) becomes

$$\mathbf{\Sigma}_{ij}^{\mathrm{A}} = \sigma_{\mathrm{in}}^2 \, \mathbf{G}_{ii}^{A} \, \mathbf{G}_{jj}^{A} \, \mathbf{w}^i \cdot \mathbf{w}^j + \mathbf{\Sigma}_{ij}^{\mathrm{out}}. \tag{6}$$

The first term in (6), the photoreceptors' contribution to the noise covariance between two cortical neurons i and j with shared photoreceptor inputs, is simply proportional to the similarity in the two cells' synaptic weights. Neuron pairs that transform their photoreceptor inputs similarly will have positive noise correlations, whereas cell pairs that transform these inputs differently can have negative noise correlations.

§3. Discrimination of two stimuli using the signals of a noisy neural population

Consider two stimuli $\mathbf{s}^A$ and $\mathbf{s}^B$, each of which elicits a conditional distribution of cortical neural ensemble activity, namely $P_A(\mathbf{r}|\mathbf{s}^A)$ and $P_A(\mathbf{r}|\mathbf{s}^B)$. We wish to decode stimulus identity from the population activity using a decision variable that reads out the activity in a linear manner, $v = \hat{\mathbf{w}} \cdot \mathbf{r}$. The two conditional distributions for the ensemble activity lead to conditional distributions for the decision variable, $P_A(v|\mathbf{s}^A)$ and $P_B(v|\mathbf{s}^B)$. The ease with which we can discriminate the two stimuli depends on how well separated these two distributions are.

When the readout vector $\hat{\mathbf{w}}$ samples from many neurons, and the neural populations are weakly correlated, the distributions over v will be approximately Gaussian and thus well characterized by their mean and variance. More generally, a convenient measure of the separation between the two distributions is given by the signal-to-noise ratio (SNR), also known as $(d')^2$. This measure is the squared difference in the means of the two distributions, normalized to the variance:

$$d'(\hat{\mathbf{w}})^2 = \frac{\left[\langle v|\mathbf{s}^A\rangle - \langle v|\mathbf{s}^B\rangle\right]^2}{\frac{1}{2}\langle(\delta v)^2|\mathbf{s}^A\rangle + \frac{1}{2}\langle(\delta v)^2|\mathbf{s}^B\rangle} = \frac{[\hat{\mathbf{w}}\cdot\Delta\boldsymbol{\mu}]^2}{\hat{\mathbf{w}}^T\mathbf{\Sigma}\hat{\mathbf{w}}}, \tag{7}$$

where $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_A - \boldsymbol{\mu}_B$ and $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are the mean neural population patterns for each stimulus, and $\boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{\Sigma}^A + \frac{1}{2}\boldsymbol{\Sigma}^B$ is the average noise covariance matrix of the two conditional distributions of neural activity patterns. This measure of discriminability depends on the statistical structure of the two conditional distributions of neural population activity in response to the two stimuli, and on the linear readout direction $\widehat{\mathbf{w}}$. One can maximize (7) over the choice of readout $\widehat{\mathbf{w}}$ to obtain the optimal readout

$$\mathbf{w}_{\text{opt}} = \boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}, \tag{8}$$

and its associated optimal signal-to-noise ratio

$$\left(d'_{\text{opt}}\right)^2 = \Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}. \tag{9}$$

This optimal SNR depends on the stimulus conditioned neural distributions only through their means and covariances. When the two distributions are exactly Gaussian with the same covariance matrix, this optimal SNR is the Kullback-Leibler divergence between the two distributions. This divergence is a statistical measure of how different the two distributions are from each other and governs the error rate of a hypothesis test attempting to distinguish between them[3]. Furthermore, when the two means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are close to each other, then this Kullback-Leibler divergence becomes proportional to the Fisher information[4] conveyed about the stimulus identity by a single neural activity pattern.

§4. From circuit properties to stimulus discrimination performance

To understand how well one can discriminate between two stimuli based on their evoked patterns of neural ensemble activity, and how this ability depends on the synaptic weights, neuronal nonlinearities and noise properties, here we combine the results from the two prior sections.

We can compute $\left(d'_{\text{opt}}\right)^2$ by inserting into (9) the expressions for the stimulus

conditional means and covariances from equations (3) and (6). To gain insight into how neural correlations limit the information encoded about the external stimuli, we focus here on the case of a linear network, with $F_i(x) = x$ for all i, so that the gain values $\mathbf{G}_{ii}^{A} = 1$ for all neurons and all stimuli. Then $\Delta\boldsymbol{\mu} = \mathbf{W}\Delta\mathbf{s}$, where $\Delta\mathbf{s} = \mathbf{s}^A - \mathbf{s}^B$ and $\boldsymbol{\Sigma} = \sigma_{\text{in}}^2\mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\Sigma}^{\text{out}}$, yielding

$$\left(d'_{\text{opt}}\right)^2 = \Delta\mathbf{s}^T\mathbf{W}^T[\sigma_{\text{in}}^2\mathbf{W}\mathbf{W}^{\mathbf{T}} + \boldsymbol{\Sigma}^{\text{out}}]^{-1}\mathbf{W}\Delta\mathbf{s}. \tag{10}$$

Thus, $\left(d'_{\text{opt}}\right)^2$ is a complex property of the synaptic architecture $\mathbf{W}$ and the cortically generated noise covariance $\boldsymbol{\Sigma}^{\text{out}}$. To attain insight regarding the saturation of information, we further assume the cortically generated noise is white, $\boldsymbol{\Sigma}^{\text{out}} = \sigma_{\text{out}}^2\mathbf{I}$. This corresponds to the simple case in which all cortical correlations originate from common photoreceptor input. In this case, we can analyze (10) through the singular value decomposition (SVD) of the weight matrix:

$$\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \tag{11}$$

Here, $\mathbf{U}$ is an $N_c \times N_p$ matrix whose orthonormal columns are output singular vectors of $\mathbf{W}$, $\mathbf{V}$ is an $N_p \times N_p$ matrix whose orthonormal columns are input singular vectors, and $\mathbf{D}$ is an $N_p \times N_p$ diagonal matrix of singular values, where $\mathbf{D}_{\alpha\alpha} = d_\alpha$. Thus the synaptic weight matrix $\mathbf{W}$ maps special patterns of photoreceptor inputs (the input singular vectors) to special patterns of cortical outputs (the corresponding output singular vectors), with network amplification factors, d_α. The amplification factors are all non-negative, and we arrange them in decreasing order, $d_1 \geq d_2 \ldots \geq d_{N_p}$.

As an important aside, we note a connection between the network amplification factors d_α and the eigenvalues of the cortical noise covariance matrix, which here takes the form

$$\boldsymbol{\Sigma} = \sigma_{\text{in}}^2\mathbf{W}\mathbf{W}^T + \sigma_{\text{out}}^2\mathbf{I} = \sigma_{\text{in}}^2\mathbf{U}\mathbf{D}^2\mathbf{U}^T + \sigma_{\text{out}}^2\mathbf{I}\ . \tag{12}$$

Equation (12) implies the eigenvalues of the noise covariance matrix are simply $\lambda_\alpha = \sigma_{\text{in}}^2 d_\alpha^2 +$

σ^2_{out}. Further, the eigenvectors are the same as the output singular vectors of $\mathbf{W}$. This simple relation between cortical noise eigenstructure and network connectivity $\mathbf{W}$ holds when the response is linear and the internally generated cortical noise is white. More generally, cortical noise eigenstructure will reflect a more complex combination of network connectivity, nonlinear amplification, and cortically generated noise structure.

In this simple setting, by substituting (11) into (10), we obtain an expression for the discriminability, or SNR, as a function of the network amplification factors and the decomposition of the stimulus difference $\Delta\mathbf{s}$ into the special input photoreceptor patterns, or input singular vectors:

$$\left(d'_{\text{opt}}\right)^2 = \frac{1}{\sigma^2_{\text{in}}}\sum_{\alpha=1}^{N_P}[\Delta\mathbf{s}\cdot\mathbf{v}^\alpha]^2\frac{d_\alpha^2}{d_\alpha^2+\frac{\sigma^2_{\text{out}}}{\sigma^2_{\text{in}}}} \leq \frac{1}{\sigma^2_{\text{in}}}\Delta\mathbf{s}\cdot\Delta\mathbf{s}, \tag{13}$$

where $\mathbf{v}^\alpha$ is the α'th column of $\mathbf{V}$. The inequality (13) is saturated when there is no internally generated cortical noise, *i.e.* when $\sigma^2_{\text{out}} = 0$. Moreover, the upper bound on $\left(d'_{\text{opt}}\right)^2$ in equation (13) is simply the SNR in the photoreceptor input layer. Thus, this result reveals that no matter how many cortical neurons there are, the SNR in the cortical layer cannot exceed the SNR in the input layer.

While we have reached this conclusion after having made multiple assumptions, the reach of this result is actually much more general when understood from the perspective of Fisher information. The data processing inequality for Fisher information states that, given a set of signals subject to noise, no amount of computational processing can ever increase the Fisher information governing the ability to estimate a particular statistical parameter from these stochastic signals[5]. Thus, the Fisher information of the cortical neural ensemble cannot exceed that of the photoreceptor ensemble. The intuition underlying this result is that any noise that

enters with the input cannot be removed by subsequent neural processing.

However, when the cortical neural ensemble also has its own intrinsic noise, as quantified by a nonzero σ^2_{out}, the gap between d'_{opt} and its theoretical upper bound will be nonzero, but this gap will decline as the number of cortical neurons N_c rises. A natural question then is, how many cortical neurons are needed so that d'_{opt} appreciably saturates and closely approaches the upper bound set by the SNR at the photoreceptor level? The only dependence of d'_{opt} on N_c arises through the network amplification factors d_α. For most typical models (we treat an explicit example below), the largest squared network amplification factors are proportional to the number of cortical neurons: $d^2_\alpha = N_c \bar{d}^2_\alpha$. Then each factor that attenuates the SNR in (10) is given by

$$\frac{N_c \bar{d}^2_\alpha}{N_c \bar{d}^2_\alpha + \frac{\sigma^2_{\text{out}}}{\sigma^2_{\text{in}}}}. \tag{14}$$

A natural measure of the saturation point is when this attenuation factor equals 1/2:

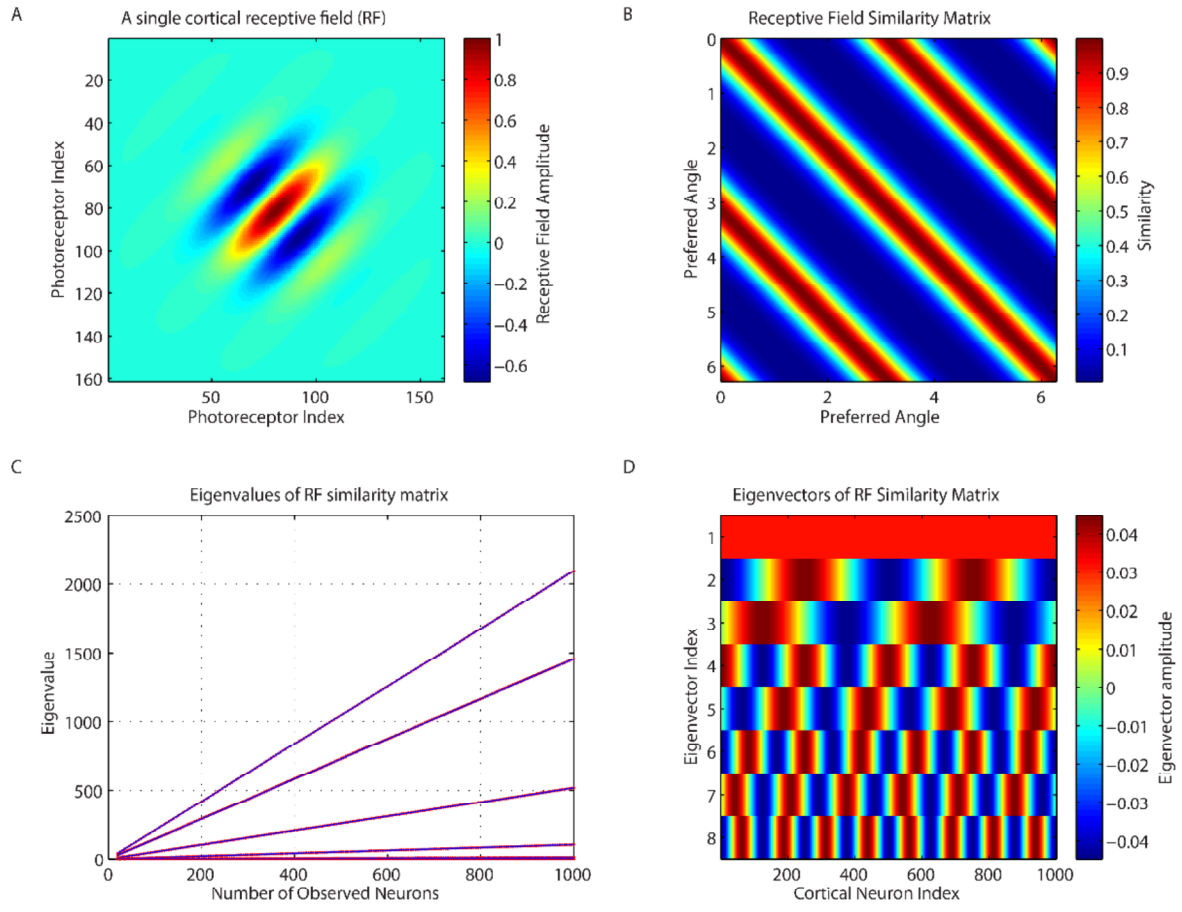
$$N^\alpha_{\frac{1}{2}} = \frac{\sigma^2_{\text{out}}}{\sigma^2_{\text{in}} \bar{d}^2_\alpha}. \tag{15}$$

This result is intuitive: if the noise generated in cortex, σ^2_{out}, is large relative to the (per neuron) amplified input noise, $\sigma^2_{\text{in}} \bar{d}^2_\alpha$, then the impact of the cortical noise can be combatted with additional cortical neurons. In this case the saturation of d'_{opt} occurs at a high value of N_c. On the other hand, if the noise generated in cortex is weak relative to the input noise, then d'_{opt} will saturate quickly at low values of N_c.

§5. Example: modeling the neural responses in V1 as a bank of orientation detectors

The above analyses of noise and SNR levels in terms of synaptic weight matrices might seem abstract. Thus, we illustrate the key ideas here by modeling the set of ensemble neural responses

in area V1 as a bank of Gabor filters[6]. Our main goal in this section is to show, in a simple conceptual model, how the network amplification factors d_α of the prior section and thus the empirically determined eigenvalues of the noise covariance matrix depend on the set of neural receptive fields. Specifically, we want to know how many of these eigenvalues grow appreciably as the number of *recorded* cortical neurons, N_0, increases. We keep N_0 distinct from the total number of cells, N_c, in the cortical circuit, as we wish to consider the empirical capacity to determine the largest eigenvalues.



Appendix Fig. 1 | Eigenvalues of the noise covariance matrix rise with the number of observed cortical neurons.

(A) An example of an individual V1 neural receptive field, defined using a Gabor function in equation (16), with $\sigma = 2$, $\frac{2\pi}{\lambda} = 3$, and $\theta^i = \frac{\pi}{4}$. The photoreceptors occupy a spatial range of $L = 16$, with spacing $dx = 0.1$, yielding a 160×160 array of photoreceptors.

(B) The normalized receptive field similarity matrix $\mathbf{W}\mathbf{W}^T$ for $N_0 = 1000$ cortical neurons with Gabor receptive fields defined as in **(A)**, but with preferred angles regularly spaced between 0 and 2π. In a simple model for the generation of noise correlations in equation (12), the receptive field similarity matrix determines the eigenvalues and eigenvectors of the noise covariance matrix. Both have the same number of eigenvalues growing with N_0 and the same eigenvectors.

(C) The largest eigenvalues of the $N_0 \times N_0$ receptive field similarity matrix $\mathbf{W}\mathbf{W}^T$ plotted as a function of the number of recorded neurons, N_0.

(D) The largest eight eigenvectors of the similarity matrix in panel (B), ordered according to the corresponding eigenvalues, from largest to smallest.

For our example we consider how well two visual stimuli, centered at the same point in visual space but with distinct orientations, can be discriminated. We thus restrict our attention to the set of $V1$ cells whose receptive fields are centered at the stimulus location, which we take to be the origin. The set of synaptic weights, $\mathbf{w}^i$, onto cortical neuron i convey inputs from a rectangular grid of photoreceptors arranged in physical space with retinotopic coordinates $\mathbf{x} = (x_1, x_2)$. We take the range of each coordinate to be $-\frac{L}{2} \leq x_i \leq \frac{L}{2}$, and the photoreceptor

spacing to be dx. Thus, there are a total of $N_p = \left(\frac{L}{dx}\right)^2$ photoreceptors covering the stimuli, and the photoreceptor density is $\rho = \frac{N_p}{L^2} = \frac{1}{dx^2}$. We model the strength of the synaptic input to cortical neuron i from a photoreceptor located in the array at position $\mathbf{x}$ as a $2D$ Gabor filter (**Appendix Fig. 1**):

$$\mathbf{w}^i(\mathbf{x}) = \mathcal{G}\left(\mathbf{x}; \sigma, \lambda, \theta^i\right) = \frac{1}{\rho} e^{-\frac{\mathbf{x}^T\mathbf{x}}{2\sigma^2}} \cos\left(\frac{2\pi\left(x_1\cos\theta^i + x_2\sin\theta^i\right)}{\lambda}\right). \tag{16}$$

Here, σ is the length scale of the Gaussian envelope, and λ and θ^i are the the length scale and orientation of the carrier wave, respectively. For simplicity, we take the length scales σ and λ to be the same for all cortical neurons in the population. We normalize the Gabor filters by the photoreceptor density so that a response to a visual stimulus similar in size to that of the receptive field remains $O(1)$ as the photoreceptor density ρ becomes large.

Now given any two cortical neurons i and j with preferred orientations θ^i and θ^j, the similarity of their receptive fields contributes to the cortical noise covariance in (12). In the limit of high photoreceptor density, this similarity can be well-approximated via an integral:

$$\mathbf{w}^i \cdot \mathbf{w}^j = \frac{1}{\rho} \int_{-\frac{L}{2}}^{\frac{L}{2}} \int_{-\frac{L}{2}}^{\frac{L}{2}} dx_1\, dx_2\, \mathcal{G}\left(\mathbf{x}; \sigma, \lambda, \theta^i\right) \mathcal{G}\left(\mathbf{x}; \sigma, \lambda, \theta^j\right) \tag{17}$$

$$= \frac{\pi\sigma^2}{2\rho} e^{-\kappa} \left[e^{-\kappa\cos\left(\Delta^{ij}\right)} + e^{+\kappa\cos\left(\Delta^{ij}\right)}\right]. \tag{18}$$

Here $\Delta^{ij} = \theta^i - \theta^j$ is the angular difference in the two preferred orientations, and the constant κ is given by $\kappa = 2\left(\frac{\sigma\pi}{\lambda}\right)^2$. The similarity between two receptive fields is thus a sum of two von-Mises functions[7], in the angular difference of preferred orientations, with sharpness parameter κ.

Now suppose we track the dynamics of N_o cortical neurons, out of a large set of N_c cells. How will the observed network amplification factors d_α, defined in (11), and

consequently the observed cortical noise eigenvalues $\lambda_\alpha = \sigma_{\text{in}}^2 d_\alpha^2 + \sigma_{\text{out}}^2$, behave as a function of N_o? The noise covariance matrix for the set of recorded cells is $N_o \times N_o$ in size and will depend on the preferred angles θ^i for $i = 1, \ldots, N_o$. We assume that these angles are uniformly distributed from 0 to 2π. Although in general one would observe random samples from this distribution, for simplicity we neglect the stochasticity of the sampling process and assume the observed set of preferred angles is regularly spaced from 0 to 2π, *i.e.* $\theta^i = \frac{i}{N_o} 2\pi$. In the limit of large N_o, this regularity approximation yields a good description of the noise covariance eigenstructure, since fluctuations in the sampling of preferred orientations can be neglected in the large N_o limit.

In this situation, the $N_o \times N_o$ matrix of receptive field similarity $\mathbf{WW^T}$, which contributes to the cortical noise covariance, (12), becomes a circulant matrix in which each row is a shifted version of the sum of von Mises functions in (18). An example of such a matrix of receptive field similarity is shown in **Appendix Fig. 1B**. The eigenvectors of such a circulant matrix, due to circular symmetry, are simply Fourier modes, or sinusoidal functions of each neuron's preferred angle. Therefore its eigenvalues are simply the Fourier amplitudes of the central row of $\mathbf{WW^T}$, multiplied by the size of the matrix N_o. The Fourier decomposition of a von-Mises function, shifted by θ_0, is given by

$$e^{\kappa \cos(\theta - \theta_0)} = I_0(\kappa) + 2 \sum_{\alpha=1}^{\infty} I_\alpha(\kappa) \cos\big(\alpha(\theta - \theta_o)\big), \tag{19}$$

where $I_\alpha(\kappa)$ is a Bessel function of the first kind. Using this relation, we find that the eigenvalues of $\mathbf{WW^T}$, or equivalently the squared network amplification factors d_α^2, are

$$d_\alpha^2 = N_0 \frac{\pi\sigma^2}{2\rho} e^{-\kappa} \big[I_\alpha(\kappa) + I_\alpha(\kappa) \cos\big(\alpha(\theta - \pi)\big) \big], \quad \text{for} \quad \alpha = 1, \ldots, N_0. \tag{20}$$

An example of this spectrum, as a function of the number of recorded neurons is shown in

Appendix Fig. 1C. The associated cortical noise eigenvectors, which are the same as those of the receptive field similarity matrix, due to equation (12), are shown arranged according to the values of the corresponding eigenvalues in **Appendix Fig. 1D**. We see that, just as in the real data (**Fig. 4d,e**; **Extended Data Fig. 10c**), only a small number of noise eigenmodes have an eigenvalue that grows linearly with the number of observed neurons N_0.

What is the origin of this simplicity, or low dimensionality? Algebraically, the Bessel functions $I_\alpha(\kappa)$ for any fixed κ are rapidly (exponentially) decreasing functions of α. Thus only a small number of modes α can be detected through linear amplification via N_0. Geometrically, the reason for this is the low-dimensional space of receptive fields present in the entire cortical population. We can view each cortical neuron's receptive field as a point in N_p-dimensional photoreceptor space. In our model, as in real neural circuits, there is a diversity of receptive fields, but the entire set of receptive fields lie on a low-dimensional, nonlinear curved manifold. In our model, this manifold is simply a circle parameterized by preferred angle. A principal components analysis of this set of receptive fields would yield a small number of nontrivial eigenvalues (identical to d_α^2), whose number corresponds to the approximate dimensionality of a linear subspace containing the curved nonlinear manifold. Each of these nontrivial eigenvalues will grow with the number of sampled points N_0 on this manifold. However, the other eigenvalues will be too small to detect even at relatively large values of N_0, because the additional observed cortical receptive fields are still approximately confined to the low-dimensional manifold.

In summary, the above analysis explains why a small number of noise covariance eigenvalues grow linearly with the number of observed cortical neurons N_0, and yields an exact formula, in (20), for the network amplification factors d_α. In turn, because noise and signal are

both amplified in the same small number of directions by the nontrivial network amplification factors d_α, we obtain the saturation of d'_{opt} with N_0, as described in (13).

§6. Determinations of noise eigenstructure and Fisher information given limited trials.

The calculation of the optimal readout vector $\mathbf{w}_{\mathrm{opt}}$ in (8), and consequently the calculation of d'_{opt} (or equivalently the Fisher information) in (9), requires knowledge of the difference in the two stimulus conditional mean responses, $\Delta\boldsymbol{\mu}$, and the stimulus-averaged noise covariance matrix, $\boldsymbol{\Sigma}$. In a real experiment we do not have direct access to the exact values of these quantities and instead must estimate them from the neural ensemble activity data. In this section, we show analytically how it is possible to attain accurate estimates of the eigenvectors with the largest eigenvalues of $\boldsymbol{\Sigma}$, given only a limited set of sensory stimulation trials. This is an important point, because in general one needs orders of magnitude more trials to accurately estimate the individual matrix elements of $\boldsymbol{\Sigma}$.

Before proceeding with the theory, we first note that the extended data figures show through analyses of both simulated and experimental data that we were neither trial-limited nor neuron-limited in our ability to accurately estimate cortical noise eigenstructure and Fisher information. Specifically, **Extended Data Figs. 7g** and **8a,b** show that modest reductions in either the number of available trials, or the number of available neurons, did not substantially change the estimate of d'. This result indicates we have enough neurons (about 1500–2000) and enough trials (about 250–300 per stimulus) to accurately estimate d' and the Fisher information; in essence, recording more neurons or more trials would not substantially change the estimates of these quantities.

Extended Data Fig. 8c–h further support this conclusion through analyses of simulated datasets that matched the actual data regarding the number of neurons, the number of trials, and

the statistical properties of the encoded visual signals and the noise eigenstructure. With these simulated datasets, we demonstrated that two qualitatively distinct methods for estimating d' (PLS and L_2-regularized regression) both recovered the known ground truth values of d' with reasonable accuracy in a variety of scenarios. Further, **Extended Data Fig. 7j** shows that our estimates of d' were stable when computed using two disjoint halves of all the trials. Finally, **Extended Data Fig. 10a,b** show there were sufficient experimental trials to estimate the noise eigenvalues and the alignment between the noise eigenvectors and the signal direction, within a PLS subspace identified as most relevant for stimulus discrimination. Notably, we were able to accurately estimate both the Fisher information and the noise eigenstructure in an experimental regime in which the moderate number of experimental trials did *not* allow accurate estimation of the individual noise correlation coeffiicents for pairs of neurons (*i.e.*, the individual matrix elements of $\mathbf{\Sigma}$), as shown in **Fig. 2d** and **Extended Data Fig. 10f**.

These results thus raise an important conceptual question: what theoretical principles govern our ability to accurately estimate both the Fisher information and noise eigenstructure in the modern experimental regime in which there are many more recoreded neurons than experimental trials, given that we cannot even estimate the individual noise correlation coefficients? Here we provide an analytic theory that addresses this question by building on known results in high dimensional statistics. The advantage of this analytic theory, beyond our empirical demonstrations of accuracy in **Extended Data Figs. 7g**, **7j**, **8a–h** and **10a,b**, lies in its ability to guide the design of future experiments, in the increasingly relevant experimental regime of large-scale neural recordings in which it is becoming commonplace to record many more cells than experimental trials. Key results from this theory are presented in **Extended Data Fig. 10f–k**, and here we present the mathematical derivations.

To develop this theory, consider a scenario in which we obtain cortical neural activity patterns $\mathbf{r}^{A,\mu}$ and $\mathbf{r}^{B,\mu}$ for $\mu = 1, \ldots, \frac{P}{2}$ trials, for each of two stimuli A and B. An estimate of the *true* noise covariance matrix $\mathbf{\Sigma}$ is given by the *empirical* covariance matrix,

$$\widehat{\mathbf{\Sigma}} = \frac{1}{\frac{P}{2}} \sum_{\mu=1}^{\frac{P}{2}} \delta\mathbf{r}^{A,\mu} (\delta\mathbf{r}^{A,\mu})^T + \frac{1}{\frac{P}{2}} \sum_{\mu=1}^{\frac{P}{2}} \delta\mathbf{r}^{B,\mu} (\delta\mathbf{r}^{b,\mu})^T \ . \tag{21}$$

In our experiments, the typical number of observed neurons is $N_o = 1000$, whereas a typical number of trials per stimulus is 300, yielding $P = 600$. Thus, for any individual pairwise correlation $\mathbf{\Sigma}_{ij}$, the error in its estimate $\widehat{\mathbf{\Sigma}}_{ij}$ is expected to be $O\left(\frac{1}{\sqrt{P}}\right) \approx 0.04$. However, the measured correlations themselves are also quite small; the root-mean-squared (RMS) scale of all off-diagonal neural correlations is 0.058 ± 0.005 (mean $\pm$ s.d. across 5 mice). Thus, our error in the estimate of any of the pairwise correlation coefficients is about the same size as the actual values we seek to determine. These considerations lead to a potential concern that we might be in a trial-limited regime in which we might be underestimating d'_{opt}, and misestimating the eigenstructure of the noise covariance, due to the limited number of trials. Thankfully, such concerns are unfounded, as we show here by outlining a theory for the estimation of the noise eigenstructure and Fisher information.

A key reason we can accurately estimate the Fisher information and noise eigenstructure is that the low-dimensional space of V1 receptive fields and the large number of recorded neurons imply $\mathbf{\Sigma}$ has only a *small* number of eigenvectors with nontrivially *large* eigenvalues (**Appendix Fig. 1C**). It is the structure of these eigenvectors, and specifically their angles with respect to the mean stimulus separation, $\Delta\boldsymbol{\mu}$, that determine the contribution to d'_{opt}, *e.g.* in equation (13). (See also **Fig. 4g,h** and **Extended Data Fig. 10c–e** for how d' is decomposed into noise eigenmodes of the neocortical recordings.) Thus, the ability to estimate the principal

eigenvectors with large eigenvalues of the true covariance matrix $\mathbf{\Sigma}$, not the matrix elements themselves, is essential for correctly estimating d'_{opt}. However, if the number of trials is sufficiently limited that we cannot accurately estimate any individual matrix element of the covariance matrix, then we must take special care to show we can still estimate the principal eigenvectors well.

For example, with $N_o = 1000$ neurons, there are $\frac{N_o(N_o-1)}{2} \approx 500{,}000$ unique pairwise correlation coefficients across the observed cell population, and it is unclear *a priori* if small errors in the estimates of all these numbers will proliferate to create large errors in the estimate of the noise eigenvectors, or whether they will average out to yield relatively accurate eigenvector estimates. Fortunately, the latter scenario is what prevails.

To illustrate, consider a simple model of the cortical noise covariance matrix, as in (12):

$$\mathbf{\Sigma} = \sigma_{\text{in}}^2 \mathbf{W}\mathbf{W}^T + \sigma_{\text{out}}^2 \mathbf{I} = \sigma_{\text{out}}^2 \left[\sum_{\alpha=1}^{K} \frac{d_\alpha^2 \sigma_{\text{in}}^2}{\sigma_{\text{out}}^2} \mathbf{u}^\alpha \mathbf{u}^{\alpha T} + \mathbf{I} \right]. \tag{22}$$

Here we have modeled the noise covariance matrix by keeping only the top K non-trivial eigenmodes $\mathbf{u}^\alpha$, corresponding to the largest network amplification factors d_α. Now consider an empirical covariance matrix estimate, $\widehat{\mathbf{\Sigma}}$, obtained from P trials as in equation (21). This empirical covariance matrix will have its own eigenvectors $\widehat{\mathbf{u}}^\alpha$. The key question is how correlated are the empirical eigenvectors $\widehat{\mathbf{u}}^\alpha$ of $\widehat{\mathbf{\Sigma}}$ and the true eigenvectors $\mathbf{u}^\alpha$ of $\mathbf{\Sigma}$? This model is known as the 'spiked covariance' model[8] , in which the spikes refer to the K outlier eigenvalues in the true covariance matrix (22), and it has been well studied in the literature on random matrix theory[8]. In particular, the accuracy of the eigenvector estimate $\widehat{\mathbf{u}}^\alpha$, as measured by its absolute correlation with the true eigenvector, $C_\alpha = |\widehat{\mathbf{u}}^\alpha \cdot \mathbf{u}^\alpha|$, depends on the ratio $\beta = \frac{P}{N_o}$ of trials to observed neurons as well as the eigenvalue SNR $\Delta_\alpha = \frac{d_\alpha^2 \sigma_{\text{in}}^2}{\sigma_{\text{out}}^2}$. This latter

eigenvalue SNR measures the excess amplitude of the outlier eigenvalues in (22), after normalizing by the overall level of the output noise σ^2_{out}. Intuitively, if either β or Δ_α increase, so should the accuracy of the eigenvector estimate C_α. An asymptotic analytic formula exists for $C(\beta, \Delta)$ when both N and P are large and is given by

$$C(\beta,\Delta) = \begin{cases} 0, & \text{for } \beta \le \beta_c(\Delta) = \frac{1}{\Delta^2} \\ \sqrt{\frac{1-\frac{1}{\beta\Delta^2}}{1+\frac{1}{\beta\Delta}}}, & \text{for } \beta \ge \beta_c(\Delta) = \frac{1}{\Delta^2} \end{cases} . \tag{23}$$

This formula can be applied for each α, as long as the eigenvalues are non-degenerate, so we have suppressed the α index. Thus, the accuracy of eigenvector estimation undergoes a phase transition; below a critical number of trials, one cannot estimate the eigenvector, but above this critical number, one can, with an accuracy that steadily improves with increased trial counts.

To connect back to our neural imaging experiment, what is a reasonable order of magnitude estimate for the scale of the important eigenvalue SNR, Δ? Consider the RMS amplitude of an individual off-diagonal correlation coefficient, which we have determined to be about $c_0 = 0.06$ in the experimental data. The RMS strength of an off-diagonal correlation coefficient in the model is $\sum_{\alpha=1}^{K} \Delta_\alpha \mathbf{u}_i^\alpha \mathbf{u}_j^\alpha \approx \frac{\sqrt{K}\Delta}{N_0}$, since each eigenvector is normalized so that each component is $O\left(\frac{1}{\sqrt{N_o}}\right)$. Equating the model RMS correlation $\frac{\sqrt{K}\Delta}{N_0}$ with the measured RMS correlation c_0 yields the eigenvalue scale $\Delta = \frac{N_o c_0}{\sqrt{K}}$. Thus, in analogy to the data, in this simple model as one records more neurons, if one observes the same RMS correlation strength c_0 and the same number of nontrivial modes *K,* then the eigenvalue Δ associated with each noise eigenmode will grow with the number of recorded neurons N_o. Note that another possibility is that the number of nontrivial modes *K* may itself grow with the number of recorded neurons N_0,

yielding a sublinear growth of the noise eigenvalue Δ with N_0. However, this possibility would be ruled out if the space of receptive fields characterizing the ensemble of V1 neurons has a fixed, low-dimensional structure, independent of the total number of recorded neurons, as discussed after equation (20) above in **§5**.

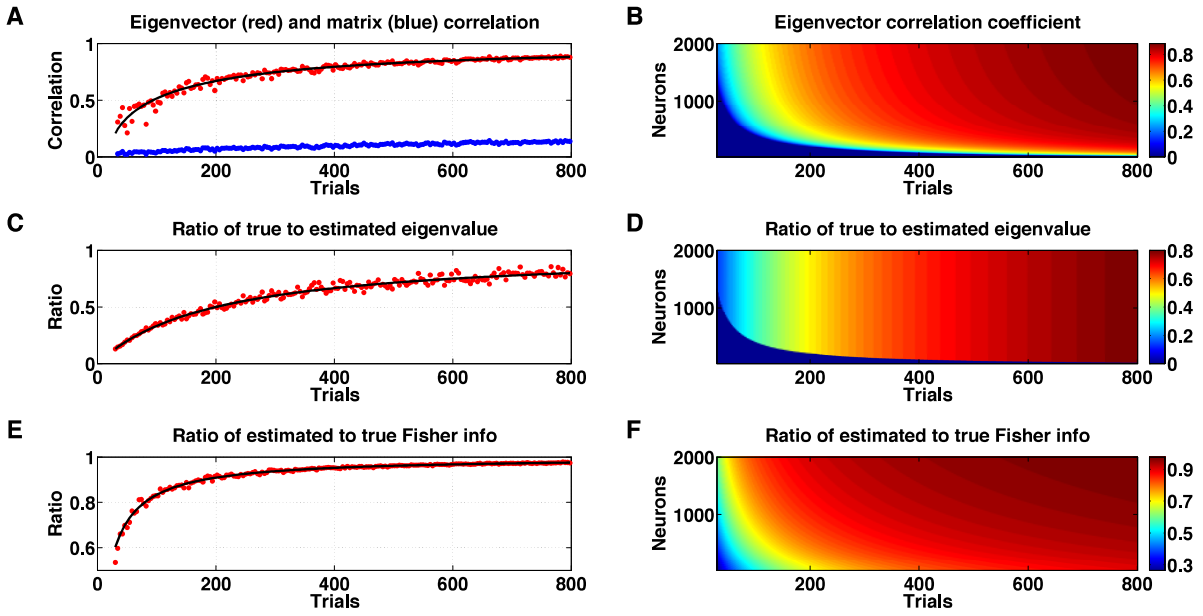
Now, to estimate the scale Δ of the noise eigenvalues, we take the measured value $c_0 \approx 0.06$ for the RMS correlation strength, and the estimate of about $K = 6$ nontrivial modes, and find $\Delta \approx 0.02\, N_0$. For $N_o = 1000$ neurons, the scale of the eigenvalue SNR is thus 20 and therefore quite high. We can compare this to direct estimates of the rate of growth of eigenvalues with N_o, as obtained from **Fig. 4d,e**. The different eigenvalues grow at different rates and range from about 5 to 50 at $N_o = 1000$ neurons, corresponding to growth rates ranging from $\Delta \approx 0.005 N_0$ to $\Delta \approx 0.05 N_0$. This is consistent with our crude estimate of $\Delta \approx 0.02 N_0$ which was obtained by simplistically assuming that the RMS correlation strength $c_0 \approx 0.06$ was divided *evenly* across $K = 6$ modes. Given that the rate of growth of Δ with N_0 is a key parameter determining the accuracy of eigenvector estimation, we introduce a new constant c to denote this growth rate. Thus c obeys the relation $\Delta = \mathrm{c} N_0$. Larger values of c will make eigenvector estimation more accurate, so to probe the difficulty of eigenvector estimation a reasonable numerical value for c is $\mathrm{c} = 0.005$, corresponding to the lower end of the range of eigenvalues in **Fig. 4d,e** and **Extended Data Fig. 10c**.

Inserting $\Delta = \mathrm{c} N_0$ and $\beta = \frac{P}{N_o}$ into equation (23), we obtain the correlation $\mathscr{C}$ between the estimated and true eigenvector as a function of the number of trials P, neurons N_0, and the eigenvalue growth rate c:

$$\mathcal{C}^2 = \begin{cases} 0 & \text{for } c^2 N_0 P \leq 1 \\ \frac{cP - \frac{1}{cN_0}}{cP + 1} & \text{for } c^2 N_0 P \geq 1 \end{cases} . \tag{24}$$

This formula exhibits a phase transition: we start to acquire information about the true eigenvector as soon as $N_0 P \geq 1/c^2$. Thus what really matters for eigenvector estimation is the *product* of two very different experimental resources: the number of neurons N_0 and the number of trials P. This implies the practical result that we can tradeoff these two resources. To obtain the same accuracy in eigenvector estimation, one actually requires *fewer* trials if one records *more* neurons.

As an example, in our experiments the scale of the product of experimental resources is given by $N_0 P = (1000)(600) = 600{,}000$. Thus we can expect to begin to acquire information about eigenmodes whose eigenvalue grows at a rate of $c \geq \frac{1}{\sqrt{N_0 P}} \approx 0.001$. Moreover, with 2000 neurons we have $N_0 P = 1.2 \cdot 10^6$, and so we can detect modes with growth rate $c \geq 0.0009$. These growth rates are well below what we actually observe in **Fig. 4e** and **Extended Data Fig. 10c**, thereby providing a theoretical explanation of the empirically observed stability of our eigenvector estimates with respect to a reduction in the number of trials, as demonstrated in **Extended Data Fig. 10b**. Moreover, **Appendix Fig. 2A** demonstrates the validity of equation (24) and reveals that it is indeed possible to correctly estimate the principal eigenvector of the cortical noise covariance matrix, in the parameter regime in which our data lives, while still being unable to accurately estimate any individual correlation coefficient. An illustration of the tradeoff between neurons and trials is given in **Appendix Fig. 2B**.



Appendix Fig. 2 | Both the principal eigenvector of the cortical noise covariance matrix and the Fisher information can be accurately determined given a limited number of sensory stimulation trials.

(A) Here we consider a true covariance matrix, $\mathbf{\Sigma}$, for $N_0 = 1000$ neurons, with a signal eigenvalue of SNR $\Delta = (0.005) \cdot N_0 = 5$. This scaling of the eigenvalue SNR with N_0 qualitatively matches the data of **Fig. 4e** for the weaker, and therefore harder to estimate, noise modes. For each set of P trials, we empirically estimate the covariance matrix, $\hat{\mathbf{\Sigma}}$, and its principal eigenvector, $\hat{\mathbf{u}}^1$, based on the P samples. We then plot the correlation coefficient between the estimated and true off-diagonal matrix elements of $\hat{\mathbf{\Sigma}}$ (*blue dots*), the correlation coefficient between the estimated eigenvector $\hat{\mathbf{u}}^1$ and true eigenvector $\mathbf{u^1}$ (*red dots*), and the theoretical prediction of equation (24) for this latter correlation (*black line*). With a number of trials that is only ~20% the number of neurons, one can estimate the principal eigenvector well, even though this is not possible for the individual matrix elements. More precisely,

for this same set of parameters, to estimate the individual matrix elements of $\mathbf{\Sigma}$ well, with a correlation coefficient >0.9, one needs >10,000 trials (simulated data not shown here).

(B) More generally, we plot the correlation coefficient between the estimated eigenvector $\hat{\mathbf{u}}^1$ and true eigenvector $\mathbf{u^1}$ predicted by equation (24), for the same parameters as in panel (A), now as a joint function of the number of recorded neurons N_0 and the number of trials P. Iso-contours of constant correlation have a hyperbolic shape, indicating a tradeoff between neurons and trials, with datasets of *more* neurons actually requiring *fewer* trials for accurate estimation of the noise eigenvectors.

(C) For the same parameters as in panel (A) we compute the ratio $\Re_\lambda$ between the true eigenvalue of $\mathbf{\Sigma}$ associated with eigenvector $\mathbf{u^1}$ and the estimated eigenvalue of $\hat{\mathbf{\Sigma}}$ associated with eigenvector $\hat{\mathbf{u}}^1$. The black curve is the theoretically predicted value for $\Re_\lambda$ obtained in equation (25), whereas the red dots are obtained from simulations.

(D) We plot the ratio $\Re_\lambda$ predicted by equation (25) for the same parameters as in (A), now as a joint function of both the number of observed neurons N_0 and the number of trials P.

(E) We plot the fraction of the total Fisher information in a neural population that we can estimate, as a function of the number of trials P to which we have access. The black curve indicates the theoretical prediction for this fraction, $\left(\frac{d'}{d'_{\text{opt}}}\right)^2$, obtained by combining equations (28, 29). The red dots indicate the ratios obtained via numerical simulations of single component PLS regression applied to the rank 1 signal and noise model with parameters $\sigma_{\text{in}}^2 = 0.002$ and $\Delta s^2 = 0.04$, both chosen

to match properties of the Fisher information curves measured in **Extended Data Fig. 8a,b** (see text before equation (29) for details).

(F) More generally, for the same parameters as in panel (E), the theoretically predicted values of $\left(\frac{d'}{d'_{\mathrm{opt}}}\right)^2$ obtained from equations (28, 29) are plotted as a joint function of the numbers of observed neurons N_0 and trials P. As in (B), there is a hyperbolic tradeoff between neurons and trials, with datasets of *more* neurons actually requiring *fewer* trials for accurate estimation of the Fisher information.

In summary, the discrepancy between our ability to estimate eigenvectors well, but not individual matrix elements, reflects that the latter are small, whereas the former have a very large associated signal eigenvalue, $\Delta = cN_0$, due to the simultaneous observation of many neurons N_0. This large eigenvalue SNR, which depends critically on being able to simultaneously monitor many neurons (**Fig. 1**), leads to our ability to estimate the noise eigenstructure accurately with fewer trials than neurons. Indeed, the greater the number of neurons N_0, the larger the eigenvalue SNR $\Delta = cN_0$ and the fewer the critical number of trials $P_c = \frac{1}{c^2 N_0}$ required to start to accurately estimate the noise eigenmode.

A similar analysis can be derived for the accuracy of estimates of the corresponding eigenvalues. In the spiked covariance model[8] of equation (22), if a true large eigenvalue of $\mathbf{\Sigma}$ in equation (22) is $\lambda^\alpha = 1 + \Delta^\alpha$ (here, without loss of generality, we work in units where the overall scale σ^2_{out} in equation (22) is set to 1), and the corresponding estimated eigenvalue of $\widehat{\mathbf{\Sigma}}$ in equation (21) is denoted by $\hat{\lambda}^\alpha$, then the ratio $\Re_\lambda$ of the true eigenvalue to its estimated value can be read off from Ref. 8 as

$$\Re_\lambda = \begin{cases} 0, & \text{for } c^2 N_0 P \le 1 \\ \frac{cP}{cP+1} & \text{for } c^2 N_0 P \ge 1 \end{cases} . \qquad (25)$$

Here, as above, we have assumed $\Delta^\alpha = cN_0$, and the answer does not depend on which large eigenvalue λ^α is chosen, so we have suppressed the eigenvalue index. We note again, as in equation (24), the product $c^2 N_0 P$ must exceed 1 to obtain any information about the eigenvalue, which is reasonable because this same product must exceed 1 to obtain any information about the eigenvector. This again implies a tradeoff between neurons and trials in eigenvalue estimation. However, after this inequality is met, the accuracy in eigenvalue estimation depends only on the number of trials P and eigenvalue growth rate c, rising sigmoidally with the product cP. The validity of equation (25) is demonstrated in **Appendix Fig. 2C**, and the joint dependence on the numbers of neurons and trials is illustrated in **Appendix Fig. 2D**.

We now move beyond the estimation theory of noise eigenvectors, and their associated eigenvalues, to an analysis of how accurately we can estimate d'_{opt}. We already demonstrated the ability to estimate d'_{opt} numerically both in the experimental data (**Extended Data Figure 8a,b**) and in simulations with parameters matched to those of the real data (**Extended Data Figure 8c–h)** using two different algorithms (PLS and L_2-regularized regression) that combat overfitting due to limited trial numbers. To gain further insight into why we can accurately estimate d'_{opt} with far fewer trials than neurons, we develop a theory for the estimation of d'_{opt} by PLS analysis in the simple case of a rank 1 signal and noise data model, identical to the model used to generate **Extended Data Fig. 8c**. We derive a simple, exact formula as a function of the numbers of recorded neurons N_0 and trials P for the estimate of d'_{opt} for this rank-1 signal and noise model.

In this simple rank-1 data model, the signal direction is given by $\Delta\boldsymbol{\mu} = \Delta s \sqrt{N_0}\,\mathbf{u}$ and the noise covariance is given by $\boldsymbol{\Sigma} = \Delta\,\mathbf{u}\mathbf{u}^T + \mathbf{I}$. Thus, both the signal and noise are spread out along the same direction $\mathbf{u}$ in firing rate space, where $\mathbf{u}$ is any unit norm vector. This data model corresponds to a special case of our V1 model described near equation (10) in **§4**. In this special case, we have just one photoreceptor input, with outgoing connectivity vector $\mathbf{w} = \sqrt{N_0}\,\mathbf{u}$ to a population of N_0 cortical neurons. The weight vector $\mathbf{w}$ is normalized so that each individual synaptic strength is O(1), independent of the number of observed neurons N_0. Moreover, in the V1 model we have chosen $\sigma^2_{\text{out}} = 1$, and so equation (22) implies $\Delta = \sigma^2_{\text{in}} N_0$. In this simple rank-1 model, the optimal direction in cortical firing rate space for reading out the photoreceptor input is simply the nontrivial noise eigenvector direction $\mathbf{u}$, which coincides with the signal direction. Thus by inserting $\mathbf{u}$ into equation (7) or equivalently using equation (9), we obtain for this model

$$\left(d'_{\text{opt}}\right)^2 = \frac{N_0\,\Delta s^2}{1+N_0\,\sigma^2_{\text{in}}}. \tag{26}$$

As expected, when N_0 becomes large, d'_{opt} saturates to the Fisher information in the input, namely $\left(d'_{\text{opt}}\right)^2 = \frac{\Delta s^2}{\sigma^2_{\text{in}}}$.

Now, after setting up this model, we can analytically address the issue of understimation of d'_{opt} due to the finite number of trials P. A main issue is that we do not have direct access to the optimal read out direction $\mathbf{u}$; instead we must estimate it through $P/2$ samples of cortical activity patterns for each of the two stimuli. With fewer trials than cells we cannot perform a direct regression that would inform us how to use the neural activity patterns for optimal estimation of the stimulus. We must additionally regularize the regression. PLS performs this regularization by first computing a subspace that is most highly informative about the

relationship between neural activity and stimulus identity. For the simple rank-1 data model, good performance can be obtained by choosing a one-dimensional PLS subspace. This one dimension is spanned by the direction in neural activity space that maximally correlates with the stimulus identity. This direction is none other than the difference between the empirical mean responses between the two stimuli A and B,

$$\widehat{\Delta\boldsymbol{\mu}} = \frac{1}{\frac{P}{2}}\sum_{\mu=1}^{\frac{P}{2}} \mathbf{r}^{A,\mu} \; - \frac{1}{\frac{P}{2}}\sum_{\mu=1}^{\frac{P}{2}} \mathbf{r}^{B,\mu} \, .$$

The resulting PLS estimate of Fisher information is obtained by inserting $\widehat{\mathbf{w}} = \widehat{\Delta\boldsymbol{\mu}}$, the true mean difference $\Delta\boldsymbol{\mu} = \Delta s \sqrt{N_0}\, \mathbf{u}$, and the noise covariance $\boldsymbol{\Sigma} = \sigma_{\text{in}}^2 N_0\, \mathbf{u}\mathbf{u}^T + \mathbf{I}$ into equation (7), yielding

$$(d'_{PLS})^2 = \frac{N_0\, \Delta s^2}{\frac{1}{\mathcal{C}_{PLS}^2} + N_0\, \sigma_{\text{in}}^2} \quad , \tag{27}$$

where

$$\mathcal{C}_{PLS}^2 = \frac{\widehat{\Delta\boldsymbol{\mu}}^T \mathbf{u}}{\widehat{\Delta\boldsymbol{\mu}}^T \widehat{\Delta\boldsymbol{\mu}}} = \frac{\Delta s^2 P + 4(\sigma_{\text{in}}^2 + 1/N_0)}{\Delta s^2 P + 4(\sigma_{\text{in}}^2 + 1)} \tag{28}$$

is the squared cosine of the angle between the readout direction $\widehat{\Delta\boldsymbol{\mu}}$ estimated by PLS and the optimal readout direction $\mathbf{u}$. The final expression for $\mathcal{C}_{PLS}^2$ in equation (28) is computed by averaging over the Gaussian distribution of $\widehat{\Delta\boldsymbol{\mu}}$. Examining equation (27) we see that d'_{PLS} approaches d'_{opt} in equation (26) from below as $\mathcal{C}_{PLS}^2$ approaches 1. Further, equation (28) reveals that $\mathcal{C}_{PLS}^2$ itself monotonically increases with P and N_0, approaching 1 as both become large. Together, equations (26-28) constitute a full theory for the estimation of d' in the rank-1 data model under a finite number of P trials and N_0 neurons.

We now estimate numerical values for the parameters Δs^2 and σ_{in}^2 so as to match important measured quantities in the real neural data. In particular, we match the growth rate of the Fisher

information as a function of N_0 in equation (26) to that of the experimentally measured Fisher information. This matching implies we should set $\sigma_{\text{in}}^2 = 0.002$ to match the growth rate of Fisher information as a function of N_0 in the real data, as quantified by the finding $\varepsilon \approx 0.002$ (**Fig. 3i**). We can also match the asymptotic Fisher information at large values of N_0, which is given by the ratio $\frac{\Delta s^2}{\sigma_{\text{in}}^2}$. We set this value to 20, which is within the range of $\left(d'_{\text{opt}}\right)^2$ values in the real data (**Fig. 3h**), and which yields $\Delta s^2 = 0.04$. The same logic and the same values of $\sigma_{\text{in}}^2 = 0.002$ and $\Delta s^2 = 0.04$ were used for the simulations of **Extended Data Fig. 8c**. A critical question then is, for the data-matched model parameters of $\sigma_{\text{in}}^2 = 0.002$ and Δs^2=0.04, how quickly does d'_{PLS} approach d'_{opt} from below as a function of the numbers of trials P and neurons N_0? The ratio $\Re$ of $(d'_{PLS})^2$ in equation (27) to $(d'_{opt})^2$ in equation (26) is given by

$$\Re = \frac{\left(d'_{PLS}\right)^2}{\left(d'_{opt}\right)^2} = \frac{1+N_0\,\sigma_{\text{in}}^2}{\frac{1}{c_{PLS}^2}+N_0\,\sigma_{\text{in}}^2} \quad . \tag{29}$$

Both the theoretically predicted behavior of $\Re$ and numerical simulations of the estimate of $\Re$ are shown as a function of P for $N_0 = 2000$ neurons in **Appendix Fig. 2E.** These curves demonstrate that our theory in equations (24-28) matches the numerical simulations, and that the numbers of trials in our experiment suffice to yield d'_{PLS} estimates close d'_{opt}. This is despite our recording far more neurons than trials, and despite being unable to estimate individual pairwise noise correlations (as demonstrated in **Appendix Fig. 2A**).

More generally, to explore the tradeoff between the numbers of neurons N_0 and trials P, in **Appendix Fig. 2F** we plot the ratio $\Re$ as a joint function of P and N_0. This figure reveals a hyperbolic tradeoff between neurons and trials for estimating the Fisher information, similar to that for estimating the noise eigenvectors (**Appendix Fig. 2B**). Thus, for estimating the Fisher

information, as with estimates of the noise eigenstructure, one actually requires fewer experimental trials given recordings of more neurons.

While equations (24-28) constitute an analytic theory for estimating Fisher information in large populations of neurons with few trials in a simple rank-1 model, the qualitative features exhibited in **Appendix Fig. 2E,F** are expected to hold for higher-rank models with more noise modes, as long as the eigenvalues of the K nontrivial noise modes grow with N_0, and the squared distance between mean activity patterns grows with N_0. Indeed, these two conditions hold in the higher-rank V1 model described near equation (10) in **§4**. The basic intution for why we can estimate Fisher information well under these conditions is that with such large eigenvalues, we can estimate the PLS subspace with few trials, and then employ regression within this subspace to find a readout close to the optimal one, achieving a value of $\mathscr{C}_{PLS}^2$ in equation (26) close to 1 and thereby achieving an underestimate ratio $\Re$ in equation (27) that is also close to 1.

In summary our analytic theory for the estimation of both noise eigenstructure and Fisher information in large neural populations with few trials provides important conceptual insights. Most importantly, in scenarios where both signal and noise are confined to a low-dimensional space, one can tradeoff two very different experimental resources (neurons and trials) to accurately estimate both eigenmodes and information. The elucidation of this tradeoff appears likely to play an important role in both the interpretion and design of future large-scale recording experiments, especially as novel recording technologies dramatically increase the number of cells that can be simultaneously recorded in a reasonable time frame. Thus, the knowledge that recordings of more neurons may actually require fewer experimental trials for studies of neural ensemble coding is of substantial relevance to experiments in modern neuroscience.

§7. Information limiting correlations in an independent set of neural recordings.

A recent study, available in preprint form, examined visual cortical coding using two-photon Ca^{2+} imaging in area V1 of awake mice as they viewed square gratings projected across the entirety of a video monitor[9]. These data are publicly accessible[10], which allowed us to apply the same analytic methods that we developed and used to study our own data. As with the datasets we acquired, the analytic results reveal strong information-limiting correlations in the real but not in trial-shuffled neural activity datasets, thereby confirming our paper's major conclusions.

We focused on 5 publicly accessible data files[10], labeled TX38, TX39, TX40, TX41, and TX42. Each of these comprised large-scale neural recordings acquired in area V1 as mice viewed ~4000 trials in which square gratings were presented at orientations chosen independently from a uniform distribution between 43–47 degrees[9]. We used L_2-regularized regression to decode the stimulus orientation values based on the evoked neural activity, with 5-fold cross-validation. The resulting d' values determined on held-out data as a function of the angular separation, $\Delta\theta$, between two classes of gratings are shown in **Appendix Fig. 3A**.

We first show that our analytic approach based on the d' measure is consistent with the analysis of the probability of decoder error, P_e, in Ref. 9. The decoding methods of Ref. 9 achieved $P_e = 0.25$ for a hypothesis test between two different gratings presented at an angular separation of 0.64 degrees, as shown in Fig. 3C in version 1 of the preprint[9]. Here we define the angular *separation* between 2 gratings presented at $\pm\theta$ degrees as $\Delta\theta = 2\theta$, which is twice the angular *difference* defined on the horizontal axis in Fig. 3C of Ref 9. To compare the authors' analysis to ours, we note that the error probability for a hypothesis test between two Gaussian distributions (here corresponding to the statistical distributions of the projections of ensemble neural activity patterns onto the regression vector) separated by a given d' value, is

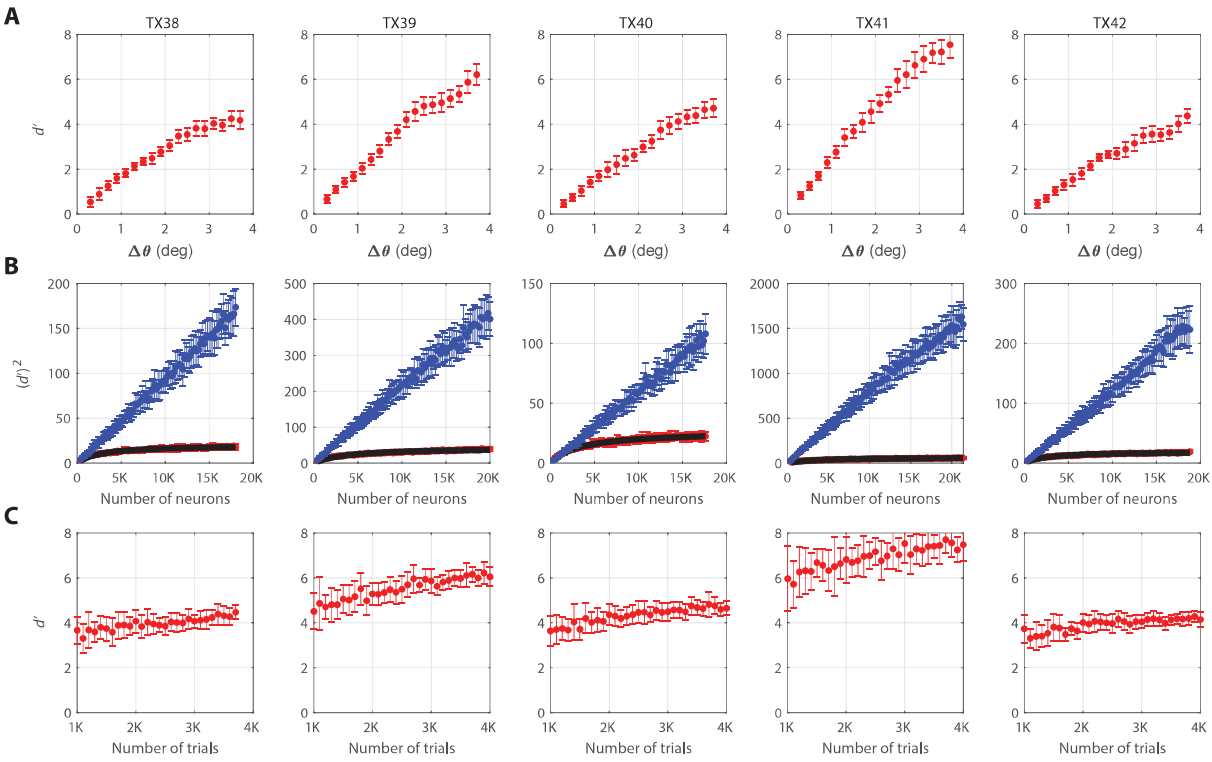
$$P_e = \int_{\frac{d'}{2}}^{\infty} \frac{dx}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad .$$

Thus, a probability of error of $P_e = 0.25$ corresponds to a d' value of approximately 1.35. In **Appendix Fig. 3A** we show that for all 5 datasets, TX38–TX42, d' is about 1.35 when the angular separation $\Delta\theta$ is in the range of 0.6–0.8 degrees. Thus, our estimate of d' is broadly consistent with the determinations of error probability in Ref. 9.

Now, to test for the presence of information-limiting correlations, it is vital to compare how d' values depend on the number of neurons in both real and trial-shuffled versions of the data. Since Ref. 9 does not include this analysis, we performed it for $\Delta\theta = 3.7$ degrees and attained results that plainly demonstrate the presence of strong information-limiting correlations (**Appendix Fig. 3B**). In all 5 data files, TX38–TX42, $(d')^2$ rose linearly with the number of cells in trial-shuffled versions of the data, but $(d')^2$ values saturated in the real data. When the analyses used the activity traces of all the recorded cells, the $(d')^2$ values for the real and shuffled datasets differed by an order of magnitude. Based on the theoretical results of **§6** above, we fit a parametrized curve of the form $(d')^2 = \frac{cN}{1+\varepsilon N}$ to the empirical plots of $(d')^2$ as a function of N in the real data, which yielded excellent fits for all 5 data files (**Appendix Fig. 3B**). We also found similar saturation behavior and correspondences between our theory and the empirical data for smaller angular separations (data not shown).

Further, we analyzed how the empirical determinations of d' varied with the number of experimental trials (**Appendix Fig. 3C**). In all 5 data files, as the number of trials rose from 1000–4000, we found only a modest rise in d' that was not much larger than the s.d. in the estimates of d' across the different cross-validation folds, and thus far less than the differences in the d' values observed in the real and shuffled datasets. Plainly, information-limiting

correlations are a strong and robust feature of both our and the publicly available datasets[10].



Appendix Fig. 3 | Information limiting correlations in an independent dataset.

(A) Plots of d' as a function of the orientation difference, $\Delta\theta$, between pairs of gratings. In each of 5 datasets, TX38–TX42, we analyzed the relationship between the grating orientation value, which was uniformly distributed between –2 and +2 degrees, and the neural activity of ~20,000 neurons. We found the regression vector using L_2-regularized regression and 5-fold cross validation with ~4000 experimental trials. We computed d' values on held-out trials that allowed us to compare two classes of gratings, one with orientations in the range $\theta \pm 0.15$ degrees and the other with orientations in the range $-\theta \pm 0.15$ degrees, yielding a mean angular separation of $\Delta\theta = 2\theta$. Error bars in this and all subsequent panels denote the s.d. of d' estimates computed across the 5-fold cross validation and 20 different randomly

chosen subsets of cells and/or experimental trials.

(B) For $\Delta\theta = 3.7$ degrees and ~4000 trials, we computed d' as a function of the number of cells included in the analysis, both in the real (red points) and trial-shuffled (blue points) datasets. In the shuffled datasets, $(d')^2$ rises linearly with the number of neurons, but in the original datasets $(d')^2$ saturates, revealing information-limiting correlations. The d' values in the real data match well with a parametric fit (black curve) of the form derived in **§6**, $(d')^2 = \frac{cN}{1+\varepsilon N}$, where N is the number of cells.

(C) For the maximal angular separation of $\Delta\theta = 3.7$ and ~20,000 cells, we computed d' as a function of the number of trials included in the analysis, which revealed only a modest increase in d'values over the range of 1000–4000 trials.

We note that ε values in the datasets TX38–TX42 (**Appendix Fig. 3B**) are plainly smaller than those in our datasets (**Fig. 3i**; **Extended Data Fig. 9c**). We surmise this is due, at least in part, to the use of visual grating stimuli in Ref. 9 that are larger than those we used, since stimuli spanning a greater portion of the visual field will elicit neural signals that reach area V1 via a greater diversity of anatomical connections. This in turn will reduce ε, which characterizes the degree of overlap in the active inputs to V1. The resulting low values of ε likely play a major role in decreasing the value of the minimum discriminable angular difference in visual grating orientation, $2\theta_{min}$, inferred from the cortical activity data of Ref. 9, in comparison to that inferred from our data. The stimuli of Ref. 9 also appear to be of greater visual salience than our stimuli, which would further decrease θ_{min} by increasing the parameter c in the equation $(d')^2 = \frac{cN}{1+\varepsilon N}$. However, an exact comparison of the stimuli is not possible because Ref. 9 does not fully specify the luminance and other stimulus parameters. The distribution of grating

orientations across the set of all stimulus presentations will also influence d' values owing to sensory adapation. Notwithstanding differences in the values of c, ε and hence d', both our datasets and those of Ref. 9 reveal a saturation of d' values at large values of N, confirming that information-limiting noise correlations place a fundamental bound on the accuracy of cortical coding.

Finally, it is worth noting that whereas Ref. 9 claims that mice cannot visually distinguish gratings whose orientations differ by <25°, visual behavioral studies[11] show that mice can in fact discriminate gratings with orientation differences as small as 4.6°, using grating stimuli similar to ours. This behavioral threshold for orientation discrimination is numerically similar to the 4.8° value we estimated from our recordings of visual cortical ensemble activity. However, given that multiple stimulus parameters influence d' values, rigorous comparisons between the accuracies of sensory cortical coding and psychophysical discriminations will require concurrent evaluations in individual animals, using identical stimuli.

§8. References for Mathematical Appendix

1 Kanitscheider, I., Coen-Cagli, R. & Pouget, A. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E6973-E6982, doi:10.1073/pnas.1508738112 (2015).

2 Wandell, B. A. *Foundations of vision*. (Sinauer Associates, 1995).

3 Cover, T. M. & Thomas, J. A. *Elements of information theory*. (John Wiley & Sons, 2012).

4 Fisher, R. A. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* **22**, 700-725 (1925).

5 Amari, S.-i. & Nagaoka, H. *Methods of information geometry*. Vol. 191 (American Mathematical Soc., 2007).

6 Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607 (1996).

7 Fisher, R. Dispersion on a sphere. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **217**, 295-305 (1953).

8 Johnstone, I. M. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics* **29**, 295-327, doi:DOI 10.1214/aos/1009210544 (2001).

9 Stringer, C., Michaelos, M. & Pachitariu, M. High precision coding mouse visual cortex. *BioRxiv*, doi:10.1101/679324 (2019).

10 Pachitariu, M., Michaelos, M. & Stringer, C. Recordings of neurons from V1 in response to oriented stimuli. doi:10.25378/janelia.8279387.v2 (2019).

11 Glickfeld, L. L., Histed, M. H. & Maunsell, J. H. Mouse primary visual cortex is used to detect both orientation and contrast changes. *J Neurosci* **33**, 19416-19422, doi:10.1523/JNEUROSCI.3560-13.2013 (2013).