nature SUPPLEMENTARY INFORMATION
https://doi.org/10.1038/s41586-019-1346-5

In the format provided by the authors and unedited.

High-dimensional geometry of population responses in visual cortex

Carsen Stringer[1,2,6]*, Marius Pachitariu[1,3,6]*, Nicholas Steinmetz[3,5], Matteo Carandini[4,7] & Kenneth D. Harris[3,7]*

[1]HHMI Janelia Research Campus, Ashburn, VA, USA. [2]UCL Gatsby Computational Neuroscience Unit, University College London, London, UK. [3]UCL Institute of Neurology, University College London, London, UK. [4]UCL Institute of Ophthalmology, University College London, London, UK. [5]Present address: Department of Biological Structure, University of Washington, Seattle, WA, USA. [6]These authors contributed equally: Carsen Stringer, Marius Pachitariu. [7]These authors jointly supervised this work: Matteo Carandini, Kenneth D. Harris. *e-mail: stringerc@janelia.hhmi.org; pachitarium@janelia.hhmi.org; kenneth.harris@ucl.ac.uk

Supplementary Discussion

1. Analysis of cvPCA algorithm

Neuronal responses to sensory stimuli can be divided into signal and noise components. If the response of cell c to repeat r of stimulus s is $f_r(c,s)$, we define the "signal" as the expected response, which will be equal for all repeats: $\phi(c,s) = \mathbb{E}\left[f_r(c,s)\right]$, and the "noise" on repeat r to be the residual after the expected response is subtracted: $\nu_r(c,s) = f_r(c,s) - \phi(c,s)$. This terminology is standard but somewhat unfortunate: by this definition, noise represents any difference between the response on a single trial and the trial-average, even though it may reflect deterministic encoding of behavioral or cognitive variables [1, 2, 3]. We assume that the noise is independently and identically distributed across repeats of a single stimulus; this condition can be approximated in practice by separating the presentation of the stimulus repeats by tens of minutes to avoid temporally correlated activity. By definition, the noise has zero expectation: $\mathbb{E}_{\nu_r}\left[\nu_r(c,s)|c,s\right] = 0$ for all r, c, and s. However, we do not assume that the noise has any particular probability distribution, and we allow its distribution and variance to depend on the stimulus.

We cannot measure the noise-free response $\phi(c,s)$ directly. Although it could in principle be estimated by averaging, this requires a very large numbers of repeats. Instead, we focus directly on our quantity of interest, the signal variance (i.e. the variance of the noise-free response across stimuli), for which an unbiased estimate can be obtained from just two repeats. We first describe how cross-validation allows an unbiased estimate of a single cell's signal variance, before moving on to analyze the cvPCA method in full.

Suppose we would like to compute the variance of cell c's response across a set of N_s stimuli, $s_1 \ldots s_{N_s}$. To simplify the analysis below, we shift means without loss of generality so that the neuron's mean response is zero: $\sum_{i=1}^{N_s} \phi(c,s_i) = 0$. After shifting to zero mean, the sample signal variance is:

$$V_{\text{sig}} = \frac{1}{N_s}\sum_{i=1}^{N_s} \phi(c,s_i)^2.$$

If we estimated the neuron's signal variance using its responses to a single repeat, this would be upwardly biased, as it would contain a contribution from both the signal and the noise:

$$\mathbb{E}_{\nu_1}\left[\frac{1}{N_s}\sum_{i=1}^{N_s} f_1(c,s_i)^2\right] = V_{\text{sig}} + \frac{1}{N_s}\sum_{i=1}^{N_s}\mathbb{E}_{\nu}\left[\nu_1(c,s_i)^2\right],$$

where we have used the fact that the noise has zero expectation to cancel cross-terms. The upward bias introduced by noise variance could be reduced, but not eliminated, by averaging over repeats. However, we can obtain an unbiased estimate by instead computing the covariance across two repeats. Indeed, because noise has mean zero and is uncorrelated between repeats,

$$\mathbb{E}_{\nu_1,\nu_2}\left[f_1(c,s)f_2(c,s)\right] = \mathbb{E}_{\nu_1,\nu_2}\left[\left(\phi(c,s)+\nu_1(c,s)\right)\left(\phi(c,s)+\nu_2(c,s)\right)\right] = \phi(c,s)^2.$$

Thus the covariance of responses between repeats forms an unbiased estimate of the signal variance:

$$\mathbb{E}_{\nu_1,\nu_2}\left[\frac{1}{N_s}\sum_{i=1}^{N_s} f_1(c,s_i)f_2(c,s_i)\right] = V_{\text{sig}}.$$

More generally, given a population of cells $c_1 \ldots c_{N_c}$, and a weight vector $\hat{\mathbf{u}} \in \mathbb{R}^{N_c}$, we can obtain an unbiased estimate of the variance of population activity projected on this dimension. Defining $\mathbf{f}_r(s) \in \mathbb{R}^{N_c}$ to be the vector summarizing these cells' responses to repeat r of stimulus s, and $\boldsymbol{\phi}(s) \in \mathbb{R}^{N_c}$ to be their noise-free responses, we can apply the same reasoning to show that:

$$\mathbb{E}_{\nu_1,\nu_2}\left[\frac{1}{N_s}\sum_{i=1}^{N_s}(\mathbf{f}_1(s_i)\cdot\hat{\mathbf{u}})(\mathbf{f}_2(s_i)\cdot\hat{\mathbf{u}})\right] = \frac{1}{N_s}\sum_{i=1}^{N_s}(\boldsymbol{\phi}(s_i)\cdot\hat{\mathbf{u}})^2$$

Thus, if the vector $\hat{\mathbf{u}}$ is an eigenvector of the stimulus-related variance, we will obtain an unbiased estimate of the corresponding stimulus-variance along this principal component.

1.1 Analysis of cvPCA method

We now consider the full cvPCA method, in which $\hat{\mathbf{u}}$ is not fixed but estimated from the data. First we will show that cvPCA finds a lower bound for the population eigenspectrum (theorem 1). Next, we will show that the expectation of the cvPCA estimate equals the actual value under conditions similar to our recordings (theorem 2).

As before, $\phi(c, s)$ represents the noise-free response of cell c to stimulus s, and $f_r(c, s) = \phi(c, s) + \nu_r(c, s)$ is its noise-corrupted response on repeat r. We consider a fixed set of cells $c_1 \dots c_{N_c}$, and an infinite sequence of stimuli $\{s_i : i \in \mathbb{N}\}$ selected randomly from a probability distribution $\mathbb{P}(s)$. We assume that both signal and noise have finite variance: for all i, $\mathbb{E}_s \left[\phi(c_i, s)^2\right] < \infty$ and $\mathbb{E}_{s,\nu_r} \left[\nu_r(c_i, s)^2\right] < \infty$. For each repeat r, and each possible number of stimuli N_s, define the $N_s \times N_c$ data matrix $\mathbf{F}_r$ to have entries $f_r(c_i, s_k)$. We will consider the limit of applying our analysis to $\mathbf{F}_r$, as $N_s \to \infty$.

Define the signal correlation matrix $\mathbf{G}$ by $G_{i,j} = \mathbb{E}_s [\phi(c_i, s)\phi(c_j, s)]$, and the total correlation matrix $\tilde{\mathbf{G}}$ by $\tilde{G}_{i,j} = \mathbb{E}_{s,\nu} [f(c_i, s)f(c_j, s)]$, which will include contributions from both signal and noise variance. The matrices $\mathbf{G}$ and $\tilde{\mathbf{G}}$ are not random: they are statistical summaries of the (unknown) population distribution. Define the training-set sample correlation matrix $\hat{\mathbf{G}}$ as an average of the sample of N_s stimuli: $\hat{G}_{i,j} = \frac{1}{N_s} \sum_{k=1}^{N_s} f_1(c_i, s_k)f_1(c_j, s_k)$. For any finite value of N_s this is a random matrix, as the dependence on the stimulus choices and noise has not been averaged out. However, in the limit $N_s \to \infty$, it converges to the deterministic total correlation matrix $\tilde{\mathbf{G}}$ with probability 1, by the strong law of large numbers.

Let $\mathbf{u}_n$, $\hat{\mathbf{u}}_n$, and $\tilde{\mathbf{u}}_n$ be the n^{th} eigenvectors of the three correlation matrices $\mathbf{G}$, $\hat{\mathbf{G}}$, and $\tilde{\mathbf{G}}$, arranged in decreasing order of eigenvalues. Because $\lim_{N_s \to \infty} \hat{\mathbf{G}} = \tilde{\mathbf{G}}$ with probability 1, also $\hat{\mathbf{u}}_n \to \tilde{\mathbf{u}}_n$ (assuming $\tilde{\mathbf{G}}$ has no duplicate eigenvalues).

Let λ_n denote the n^{th} eigenvalue of the true signal correlation matrix $\mathbf{G}$, and $\hat{\lambda}_n$ denote the n^{th} cross-validated eigenvalue estimate:

$$\begin{aligned} \hat{\lambda}_n &= (\mathbf{F}_1 \hat{\mathbf{u}}_n) \cdot (\mathbf{F}_2 \hat{\mathbf{u}}_n)/N_s \\ &= \frac{1}{N_s} \sum_{k=1}^{N_s} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \big(\phi(c_i, s_k) + \nu_1(c_i, s_k)\big)\big(\phi(c_j, s_k) + \nu_2(c_j, s_k)\big)\hat{u}_{n,i}\hat{u}_{n,j} \end{aligned} \tag{1}$$

We are now ready to state our first theorem, which establishes the cross-validated eigenvalue estimates as a lower bound on the population signal eigenvalues:

Theorem 1. *With probability 1,*

$$\lim_{N_s \to \infty} \mathbb{E}_{\nu_1,\nu_2} \left[\sum_{n=1}^{N} \hat{\lambda}_n\right] \leq \sum_{n=1}^{N} \lambda_n. \tag{2}$$

Proof. Taking the expectation of equation (1) over the noise on both repeats, we obtain

$$\begin{aligned} \mathbb{E}_{\nu_1,\nu_2} \left[\hat{\lambda}_n\right] &= \frac{1}{N_s} \sum_{k=1}^{N_s} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \mathbb{E}_{\nu_1,\nu_2} \left[\big(\phi(c_i, s_k) + \nu_1(c_i, s_k)\big)\big(\phi(c_j, s_k) + \nu_2(c_j, s_k)\big)\hat{u}_{n,i}\hat{u}_{n,j}\right] \\ &= \frac{1}{N_s} \sum_{k=1}^{N_s} \sum_{i=1}^{N_c} \sum_{j=1}^{N_c} \mathbb{E}_{\nu_1} [\phi(c_i, s_k)\phi(c_j, s_k)\hat{u}_{n,i}\hat{u}_{n,j}] + \mathbb{E}_{\nu_1} [\nu_1(c_i, s_k)\phi(c_j, s_k)\hat{u}_{n,i}\hat{u}_{n,j}], \end{aligned}$$

The term ν_2 is has disappeared from the expectation because it has mean 0 and is independent of $\hat{\mathbf{u}}_n$, which is computed only from repeat 1. However, $\hat{\mathbf{u}}_n$ may be correlated with ν_1, so the expectation of the second term on the right is not necessarily zero. Nevertheless, this term will with probability 1 converge to 0 as $N_s \to \infty$. Indeed, because $\hat{\mathbf{u}}_n$ converges to the non-random quantity $\tilde{\mathbf{u}}_n$,

$$\begin{aligned}\lim_{N_s\to\infty} \mathbb{E}_{\nu_1}\left[\nu_1(c_i,s_k)\phi(c_j,s_k)\hat{u}_{n,i}\hat{u}_{n,j}\right] &= \mathbb{E}_{\nu_1}\left[\nu_1(c_i,s_k)\phi(c_j,s_k)\tilde{u}_{n,i}\tilde{u}_{n,j}\right] \\ &= \mathbb{E}_{\nu_1}\left[\nu_1(c_i,s_k)\right]\phi(c_j,s_k)\tilde{u}_{n,i}\tilde{u}_{n,j} = 0,\end{aligned}$$

since the noise has mean 0 by definition. Thus,

$$\lim_{N_s\to\infty} \mathbb{E}_{\nu_1,\nu_2}\left[\hat{\lambda}_n\right] = \lim_{N_s\to\infty} \frac{1}{N_s}\sum_{k=1}^{N_s}\sum_{i=1}^{N_c}\sum_{j=1}^{N_c}\phi(c_i,s_k)\phi(c_j,s_k)\tilde{u}_{n,i}\tilde{u}_{n,j} = \tilde{\mathbf{u}}_n^\top \mathbf{G}\tilde{\mathbf{u}}_n \tag{3}$$

If the vectors $\tilde{\mathbf{u}}_n$ were eigenvectors of the signal covariance matrix $\mathbf{G}$, then this last term would equal the signal eigenvalue λ_n. Instead, $\tilde{u}_n$ are the eigenfunctions of the total covariance operator $\tilde{\mathbf{G}}$. Nevertheless, as we discuss below in Theorem 2, $\mathbf{u}$ and $\tilde{\mathbf{u}}$ are likely to coincide for the current recordings. Furthermore, even if they did not coincide, we would still obtain a one-sided bound. Indeed, the Courant-Fisher minimax principle (Ref. [4], theorem 4.2.7) implies that for any set of orthonormal vectors $\tilde{\mathbf{u}}_n(c)$, the projection of $\mathbf{G}$ onto them cannot exceed what would be obtained with the true eigenvectors $\mathbf{u}_n$:

$$\sum_{n=1}^{N}\tilde{\mathbf{u}}_n^\top \mathbf{G}\tilde{\mathbf{u}}_n \le \sum_{n=1}^{N}\mathbf{u}_n^\top \mathbf{G}\mathbf{u}_n = \sum_{n=1}^{N}\lambda_n$$

Combining this with equation (3) proves the theorem. □

We now show that the lower bound (2) will be exactly satisfied, i.e cvPCA will provide an unbiased estimate of the population eigenspectrum, when noise arises from a combination of three sources that together form an accurate description of variability in mouse visual cortex: independent variability, multiplicative scaling of population responses, and additive noise along dimensions orthogonal to the directions of signal variance [5, 6, 2]. For the current proof, we will assume that the magnitude of independent variability is equal across neurons; this assumption is not necessary in the limit that $N_c \to \infty$, but proving so will require methods to consider this limit, and is thus delayed until section 3.

Theorem 2. *Consider a noise model*

$$\nu_r(c,s) = \alpha_r(s)\phi(c,s) + \beta_r(c,s) + \gamma_r(c,s)$$

where $\alpha_r(s)$ represents multiplicative noise scaling the entire population's response to stimulus s on repeat r; $\beta_r(c,s)$ is additive noise in dimensions orthogonal to the stimulus; and $\gamma_r(c,s)$ is independent between neurons and stimuli, with α, β and γ statistically independent of each other and of $\phi(c,s)$. Then as $N_s \to \infty$, the eigenvalue estimates $\mathbb{E}_{\nu_1,\nu_2}\left[\hat{\lambda}_n\right]$ converge to the population eigenvalues λ_n together with additional zero eigenvalues corresponding to the additive noise dimensions.

Proof. From equation (3), we see that if an eigenvector $\tilde{\mathbf{u}}_n$ of the total covariance matrix $\tilde{\mathbf{G}}$ was equal to an eigenvector $\mathbf{u}_m$ of the signal covariance matrix $\mathbf{G}$, then $\lim_{N_s\to\infty}\mathbb{E}_{\nu_1,\nu_2}\left[\hat{\lambda}_n\right] = \lambda_m$. It therefore suffices to show that the eigenvectors of $\tilde{\mathbf{G}}$ are the same as those of $\mathbf{G}$, together with additional orthogonal dimensions. To do so, note that the total response

$$f_r(c,s) = \left(1+\alpha_r(s)\right)\phi(c,s) + \beta_r(c,s) + \gamma_r(c,s)$$

Thus,

$$\begin{aligned}\tilde{G}_{i,j} &= \mathbb{E}_{s,\alpha,\beta,\gamma}\left[f_r(c_i,s)f_r(c_j,s)\right] \\ &= \mathbb{E}_{s,\alpha}\left[(1+\alpha_r(s))^2\phi(c_i,s)\phi(c_j,s)\right] + \mathbb{E}_{s,\beta}\left[\beta_r(c,s)\beta_r(c',s)\right] + \mathbb{E}_{s,\gamma}\left[\gamma_r(c,s)\gamma_r(c',s)\right].\end{aligned}$$

In matrix notation:

$$\tilde{\mathbf{G}} = V_\alpha \mathbf{G} + \mathbf{B} + V_\gamma \mathbf{I},$$

where $V_\alpha = \mathbb{E}_{\alpha,s}\left[(1+\alpha_r(s))^2\right]$ is a scale factor resulting from multiplicative modulation, $\mathbf{B}$ represents the correlation of the additive noise, $V_\gamma = \mathbb{E}_{s,\gamma}\left[\gamma_r(c_i,s)^2\right]$ is the independent noise variance, and $\mathbf{I}$ is the identity matrix.

We have assumed the additive noise dimensions to be orthogonal to the signal dimensions, so $\mathbf{B}\mathbf{u}_n = 0$. Thus, $\mathbf{u}_n$ is an eigenvector of $\tilde{\mathbf{G}}$, with eigenvalue $V_\alpha\lambda_n + V_\gamma$. Similarly, if $\mathbf{b}_m$ is an eigenvector of $\mathbf{B}$ with eigenvalue η_m, it is also an eigenvector of $\tilde{\mathbf{G}}$, with eigenvalue $\eta_m + V_\gamma$. The remaining eigenvectors of $\tilde{\mathbf{G}}$ will span the common nullspace of $\mathbf{G}$ and $\mathbf{B}$, and will all have eigenvalue V_γ.

Now, although the eigenvalues of $\mathbf{u}_n$ for the total correlation matrix $\tilde{\mathbf{G}}$ have been inflated by noise to $V_\alpha\lambda_n + V_\gamma$, this does not affect the cvPCA estimate, which is derived from the test set, and thus has expectation $\mathbf{u}_n^\top\mathbf{G}\mathbf{u}_n = \lambda_n$. In addition to the $\mathbf{u}_n$, $\tilde{\mathbf{G}}$ has a new set of eigenvectors $\mathbf{b}_m$ corresponding to the eigenfunctions of the additive noise covariance $\mathbf{B}$, and further eigenvectors spanning the common nullspace of $\mathbf{G}$ and $\mathbf{B}$. These eigenvectors are all orthogonal to the directions of signal covariance, and so are annihilated by $\mathbf{G}$. Thus, the corresponding cvPCA eigenvalue estimates will be the covariance of the training and test noise along these dimensions, which is 0 as training and test set noise is independent. Therefore, under the assumed noise model, cvPCA provides an asymptotically unbiased estimate of population eigenvalues, together with an additional set of zero eigenvalues resulting from additive noise. □

2. Relating geometry to eigenspectrum decay

In this second appendix we consider the relationship between the eigenspectrum of the neural code and its geometry, and discuss the computational consequences of this relationship. Our fundamental conclusions will be that codes whose eigenspectra decay too slowly have pathological properties. We will demonstrate this by proving two theorems, and will illustrate it with three examples. The examples are counterfactual: our experimental results show that eigenspectra decay faster than $n^{-1-2/d}$, and the examples illustrate the problems that would occur if eigenspectra decayed more slowly than this.

We consider here only the geometry of noise-free responses: as shown in Appendix 1, the cvPCA method provides an unbiased estimate of the noise-free eigenspectrum. We consider their geometry in the limit as a large number of neurons are recorded, which allows us to consider neural activity in an infinite-dimensional coding space. The mathematical technicalities and philosophy of this many-neuron limit will be discussed in more detail in Appendix 3.

2.1 Why can't eigenspectra be flat?

Describing our main results will require building up some mathematical machinery concerning fractal dimensions. However, even before doing this it is possible to gain an intuition for the pathology of slowly-decaying eigenspectra, by considering a simple counterfactual example.

Example 1. This first example is the efficient coding hypothesis *in extremis*: all neurons encode the stimulus completely independent of each other. There are no signal correlations, and the eigenspectrum is therefore flat. Although neurons certainly could display independent noise, this example shows why independent encoding of a signal is pathological in large populations.

In this example, the stimulus to be encoded is a single number, and neurons encode its binary representation. Let the stimulus s be uniformly distributed between 0 and 1, and let cell c_i encode the i^{th} digit in the binary expansion of s. Because s is uniformly distributed, all neurons are independent, and the eigenspectrum is flat. Real neurons, of course, do not form binary codes: however, by considering this simple counterfactual case, we will gain an insight into the pathological properties any independent code would also have to exhibit.

To understand the pathologies of this code, suppose the stimulus was represented by a population of 1,000 neurons. The first ten cells would represent the stimulus to an accuracy of one part in 1024, which would almost certainly exceed the accuracy to which the stimulus could be measured by sensory systems. The remaining 990 cells would represent the stimulus to a supposed accuracy of more than one part in 10^{300}. Nearly all the variance of the population would therefore be used to encode absurdly fine details of the stimulus. A downstream population could only recover useful information about the stimulus by "finding a needle in a haystack," i.e. by basing its responses on just these 10 neurons, while ignoring all others. If the downstream structure received functional inputs from anything but this tiny fraction of cells, then a minute change in the stimulus would cause it to receive an almost completely different input, meaning that no generalization would be possible.

The code just described is highly susceptible to neural noise, but that is not its fundamental pathology. If the population contained many neurons redundantly encoding each digit in the binary expansion, a downstream structure could recover the signal accurately despite noise by averaging their activity. Still however, 0.1% of the population would encode the stimulus to an accuracy of over one part in 1000, with the remaining 99.9% of neurons encoding irrelevant details.

The same pathology would occur for higher dimensional stimuli. Indeed, consider the example of N completely independent binary neurons encoding a stimulus described by d real variables. The first $10d$ neurons would encode these variables to an accuracy of 1 in 1024, with the remainder of the population again representing features that could not in reality be measured by sensory systems. If $d \ll N$, then nearly all the population variance is devoted to encoding irrelevant stimulus details. Fully uncorrelated stimulus encoding is therefore pathological whenever the number of neurons substantially exceeds the dimension of the stimulus being encoded.

We have presented an example of binary neurons, but it is possible to extend it to an (even more counterfactual) case of continuous firing rates, illustrating the types of pathology that we will prove must happen for any codes whose eigenspectra decay too slowly. Considering again a one-dimensional stimulus, we can obtain a continuous code by smoothly joining the 2^N possible binary response vectors into a smooth curve. Because any two such vectors are separated by a distance of at least 1 in neural space, this curve has length at least 2^N. The average derivative of the neural response vector is the length of the neural manifold divided by the length of the stimulus manifold, which is independent of N. Thus, the derivative scales exponentially with N, and a tiny change in the stimulus would cause a huge change in the population response, reflecting a code which cannot generalize between stimuli.

In the remainder of this section, we will prove that the type of pathologies illustrated by this example are inevitable if eigenspectra do not decay sufficiently rapidly. The key to this argument is that later principal components encode finer details of the stimulus. This is empirically true in our data (Extended Data Fig. 6) and is also a mathematical necessity. The number of slowly-varying orthogonal functions of the stimulus is limited (assuming a finite-dimensional, smooth, and compact stimulus space). Because the principal components are orthogonal, they must therefore asymptotically encode ever finer stimulus details. If the eigenspectrum decays too slowly, population activity will be overwhelmed by representation of these ever finer details, resulting in a code with the same pathologies as the binary example.

2.2 Multiple notions of dimensionality

A key concept in this work is the dimensionality of a neural representation. We will require four different notions of dimensionality, which in general are not equal.

The first notion is *ambient dimension*. By this we mean the dimension of a vector space, i.e. the number of coordinates required to identify a point. For example, we can summarize the responses of N_c cells to a stimulus s by a vector in an N_c-dimensional vector space, which we refer to as the *ambient space*.

The second notion is *planar dimension*. Not every point in the ambient space needs to be a response to a stimulus; indeed, as there are more neurons in V1 than fibers in the optic nerve, the actual response patterns evoked by stimuli must occupy a subset of the ambient space. We define the planar dimension of a neural code to be the dimension of the smallest vector subspace containing all firing patterns that can actually be produced. Thus, if the planar dimension is n, then the response to any stimulus s can be represented as a sum of n basis vectors: $\boldsymbol{\phi}_s = \sum_{i=1}^{n} \phi_i \mathbf{u}_i$.

The third notion is *manifold dimension*. A *manifold* is a generalization of a continuous surface to arbitrary dimensionality. In any local region of a d-dimensional manifold, each point can be uniquely identified by d coordinates. However, unlike a vector space, a manifold may be curved, and the functions relating these coordinates to the points on the manifold can be nonlinear. A manifold can be embedded in a vector space of equal or higher ambient dimension than the manifold dimension. For example, the earth's surface defines a 2-dimensional manifold embedded in a 3-dimensional ambient space.

Our final notion of dimensionality is *fractal dimension*. Fractal dimension measures the roughness of a geometric object. For example, the coastline of a country has manifold dimension 1, but can be "rough" in the sense that the higher the resolution, the more complicated and longer the coastline appears to be. Precise definitions of fractal dimension are based on the idea that the volume of a d-dimensional object with diameter δ should scale as δ^d. The west coast of Britain, for example, has a fractal dimension of approximately 1.25 (Ref. [7]), meaning that the number of times a ruler of length δ can be laid end-to-end around the coastline scales as $\delta^{-1.25}$. A *fractal* is an object whose fractal dimension exceeds its manifold dimension. Smooth manifolds, whose coordinate functions are differentiable, cannot be fractals.

There are multiple subtly-differing ways to formalize the fractal dimension [8, 9]. Here we will use the *upper Minkowski dimension*, which is defined by counting the number of spheres of diameter δ required to cover the object. If we call this number N_δ, the Minkowski dimension is essentially the limit of $\log(N_\delta)/\log(\delta^{-1})$ as $\delta \to 0$ (see section 2.6 for precise definition).

2.3 Summary of results

We are now able to summarize our mathematical results relating the geometry of the neural code to its eigenspectrum. These results demonstrate that unless eigenspectra decay faster than n^{-1}, population codes are pathological, either exhibiting discontinuous responses, or infinite population variance. Furthermore, for stimuli drawn from a set of manifold dimension d, codes with eigenspectra decaying slower than $n^{-1-2/d}$ are also pathological, displaying infinite variance of the code's derivative and fractal geometry of the response manifold. We conclude that our experimental observations of eigenspectrum decay only just faster than $n^{-1-2/d}$ indicate a neural code that is as high-dimensional as possible before hitting the regime where these pathological conditions must occur. More detailed descriptions of each result now follow.

Theorem 3 demonstrates the pathology of neural population codes whose eigenspectra do not decay faster than n^{-1}. We show that if the sum of the population eigenvalues is infinite, then either the neural code is discontinuous or population activity has infinite variance. Because $\sum_{n=1}^{\infty} n^{-\alpha} = \infty$ when $\alpha \leq 1$, a code whose eigenspectrum did not decay faster than n^{-1} must exhibit one of these properties. Both possibilities render a neural code pathological, resulting respectively in complete failure to generalize, or requiring neurons to respond to preferred stimuli with arbitrarily large firing rates.

Theorem 4 illustrates the pathological features of eigenspectra decaying more slowly than $n^{-1-2/d}$. This theorem shows that if the fractal dimension of an object is d, then its covariance eigenspectrum cannot decay slower than $n^{-1-2/d}$. Applying this to the set of neural population responses to a stimulus ensemble of manifold dimension d, we see that if its eigenspectrum decayed slower than $n^{-1-2/d}$, the responses must define a fractal in the ambient space. The "rough" nature of such population codes means they are unsuitable for representing differences between similar stimuli: we prove in Corollary 4.1 that any map from a d-dimensional stimulus manifold to a space of fractal dimension $> d$ cannot be differentiable, so there is no way a population whose eigenspectrum decayed this slowly could smoothly encode a d-dimensional stimulus. The precise form of this non-differentiability is made explicit in the next theorem.

Theorem 5 relates the derivative of the neural population representation with respect to a stimulus of manifold dimension d, to its eigenspectrum. We prove that later eigenfunctions must encode finer stimulus details, and specifically that the derivative magnitude of the n^{th} eigenfunction must grow as order n^{2d}. This implies that unless the eigenspectrum decays faster than $n^{-1-2/d}$, the derivative of the population response vector has infinite expected magnitude. The pathology of non-differentiable population codes occurs because the difference in population response to two close stimuli is proportional to the code's derivative. If the derivative is infinite, then the population codes for extremely similar stimuli can still be very different, resulting in failure to generalize responses to arbitrarily close stimuli. This theorem also shows that the smoothest codes possible for a given eigenspectrum will have principal components whose dependence on the stimulus is given by Laplacian eigenfunctions, as would be found by the Laplacian eigenmap algorithm often used for unsupervised learning [10].

Example 2 illustrates theorems 4 and 5 with a representation of stimulus of manifold dimension 1 in an infinite-dimensional ambient space. Consistent with theorem 4, we show that if the eigenspectrum decays slower than n^{-3} then the set of possible responses is fractal. Consistent with theorem 5, we show that the derivative is infinite if the eigenspectrum decays as n^{-3} or slower. Though analytically tractable, this representation takes the form of coordinates in an abstract Hilbert space, rather than explicitly-modelled neurons.

In example 3 we consider a more explicitly neural representation, built from units with Gaussian radial basis function receptive fields over an abstract stimulus space. We show that if all units have equal tuning radius, the eigenspectrum does not follow a power law but is approximately flat until the eigenfunctions' spatial scale matches this radius, after which the eigenspectrum falls rapidly. We then show how a scale-free mixture of units of varying radii results in a power-law eigenspectrum. By varying the balance of broadly and sharply tuned neurons, we show that the population code is not differentiable if its eigenspectrum decays slower than $n^{-1-2/d}$. We show that in this case the neural representation of the derivative is dominated by increasingly rare neurons of increasingly sharp tuning, resulting in an infinite population gradient with respect to the stimulus, even though every neuron's individual response is an infinitely-differentiable Gaussian function.

2.4 Relation to kernel learning theory

To consider the computational consequences of different cortical coding geometries, one should consider how these geometries affect the ability of downstream structures to form associations between sensory stimuli and appropriate behaviors. This reveals an intriguing parallel between our findings and the theory of kernel machines, a class of machine learning algorithms that provide a useful analogy to describe features of cortical coding [11, 12]. We will discuss this parallel using a model where the cortical code is fixed, and downstream structures learn stimulus-behavior associations through Hebbian mechanisms such as the delta rule [13], leading to a behavioral output determined by nonlinear function of a linear projection of cortical population activity (linear-nonlinear function). Neither our results nor their interpretation rely on this assumption, but it provides a simple way to illustrate the computational consequences of different representational geometries.

The way animals generalize responses to sensory stimuli will depend on the overlap of the corresponding population codes. Suppose an animal had learned to produce a behavioral response to a sensory stimulus s. If the animal later experienced another stimulus s' that evoked a similar pattern of cortical population activity, this stimulus would likely produce the same behavioral response, even if the animal had never experienced s' before. However, if s and s' activate different cortical populations, such generalization would be unlikely to occur.

Sensory inputs that generate strong responses in a large number of cortical cells will be salient, meaning that the animal is likely to rapidly learn responses to them. Indeed, if downstream structures learn by Hebbian plasticity, then pairing a training signal with a stimulus that strongly drives many cortical cells will strengthen many synapses, which will strongly drive downstream neurons when the stimulus next appears. However pairing the training signal with a stimulus that weakly drives only a few cortical cells would only strengthen a few synapses, and so lead to only a weak downstream response.

These intuitive notions can be formalized in geometric terms. The geometrical relationship between the population vectors $\boldsymbol{\phi}_s$ and $\boldsymbol{\phi}_{s'}$ evoked by two stimuli can be summarized by their inner product $\langle \boldsymbol{\phi}_s, \boldsymbol{\phi}_{s'} \rangle$, which is termed the *kernel function* $K(s, s')$. The total amount of population activity evoked by stimulus s can be summarized by the length of the population vector, $\|\boldsymbol{\phi}_s\|$, which is given by $\sqrt{K(s,s)}$. The similarity of the responses to s and s' can be captured by the correlation coefficient, which is equal to $K(s,s')/\sqrt{K(s,s)K(s',s')}$. Thus, not only is the geometry of the cortical representation critical to downstream Hebbian learning, but the important features of this geometry are captured by the kernel function.

The kernel function summarizes the similarity of cortical responses to two stimuli, rather than the physical similarity of the stimuli themselves (i.e. the patterns of light falling on the retina). For example, sensory processing might render the population responses to physically similar but behaviorally different stimuli (such as an edible and a poisonous mushroom) as orthogonal, but cause two physically different stimuli of similar behavioral relevance (such as different instantiations of a similar visual texture) to have a similar population representation.

To understand how the kernel function determines downstream learning, it is useful to adopt the "function space" description of kernel learning systems [14, 15]. The kernel function is defined not just for the particular sample of stimuli shown in an experiment, but over the entire set of possible stimuli that the animal might potentially encounter. This function represents an infinite-dimensional generalization of a matrix, and we can define its eigenspectrum, although because the set of potential stimuli is infinite, there may now be an infinite number of eigenvalues. Indeed (given a technical condition known as compactness) the kernel function has a set of eigenfunctions $v_n(s)$ and real non-negative eigenvalues λ_n such that $\mathbb{E}_{s'}\left[K(s,s')v_n(s')\right] = \lambda_n v_n(s)$ (Ref. [4], Theorem 4.2.4). This definition, based on the expectation over a continuous random variable, generalizes the finite-dimensional notion of eigenvectors based on matrix multiplication. Similarly to how finite-dimensional PCA produces orthonormal eigenvectors, these eigenfunctions are uncorrelated and have unit variance: $\mathbb{E}_s\left[v_n(s)v_m(s)\right] = \delta_{n,m}$.

We can formalize a response that the system might need to learn as a function $f(s)$, again defined over the infinite set of all stimuli that might potentially be encountered. If downstream learning occurs via a delta rule, the dynamics by which this function will be learned from examples is determined by the kernel eigenspectrum. The system will first learn the projection of f onto the eigenfunctions of large eigenvalues, corresponding to the dimensions of maximum variance of the neural representation. The projection of f onto eigenfunctions with smaller eigenvalues will be added to the system's output only after more extensive training, at a speed proportional to the eigenvalues [16, 17, 18].

The kernel eigenspectrum therefore determines the flexibility of the set of functions the system is able to learn. If the eigenspectrum decays rapidly, associations will be biased towards functions expressible as combinations of those few eigenfunctions with large eigenvalues. Because fine-scale stimulus features are represented in eigenfunctions with small eigenvalues (shown in theorem 5 below), rapid eigenspectrum decay means that a crude representation will be learned quickly, but fine details cannot be distinguished without extensive training. If the eigenspectrum decays slowly, the system is able to learn much more detailed associations, but this comes at a price: the system is less able generalize from one stimulus to the next, as the representation of these stimuli is more likely to be orthogonal. Thus, the rate of eigenspectrum decay can be seen as a smoothing parameter: slower eigenspectrum decay means less smoothing across stimuli, which enables representation of fine stimulus features but impairs generalization.

This viewpoint provides a new way of understanding the n^{-1} and $n^{-1-2/d}$ bounds. If the eigenspectrum decayed slower than n^{-1}, then the representation of fine details would overwhelm the representation of broad stimulus features: because $\sum_{n=1}^{\infty} n^{-1}$ does not converge, the leading eigenfunctions will make up an infinitesimal fraction of the total neural variance, the functions the system learned would be dominated by fine details, and convergence to the correct function f would not occur after any finite training time. If the eigenspectrum decayed slower than $n^{-1-2/d}$, the functions learned by the system would fail to be differentiable, and responses to one stimulus need not generalize to other nearby stimuli.

2.5 Implications of n^{-1} eigenspectra

We now move to a formal characterization of the pathologies that occur if eigenspectra decay too slowly. To a reader interested in learning more of the relevant background material, we recommend Ref. [4], which covers almost all the required material on probability in infinite-dimensional spaces, and Ref. [9] which covers all the required material on fractal geometry. For applications of similar mathematics in machine learning, we recommend refs. [19, 20, 21], and for a more general background in functional analysis, refs. [22, 23]. The key material on Riemannian and spectral geometry can be found in Ref. [24].

Our first theorem concerns the pathologies of neural representations whose eigenspectra decay slower than $1/n$. We show that if the population eigenspectrum decays slower than $1/n$, then either the kernel function is discontinuous, or the variance of the population is infinite. The former case would mean that the neural code cannot generalize responses to new stimuli. The latter would mean that population responses are dominated by rare neurons responding at arbitrarily high rates to certain stimuli. From a mathematical perspective, this theorem follows immediately from basic results; this subsection will therefore be devoted to introducing the terminology required to describe why.

To characterize the geometry of the population code in the limit of a large number of cells, we must consider activity to reside in an infinite-dimensional vector space. For finite neural populations, we can define the distance between responses to two stimuli using Euclidean distance: if the vectors $\boldsymbol{\phi}_s$ and $\boldsymbol{\phi}_{s'}$ contain the responses of cells $\{c_1, c_2, ..., c_{N_c}\}$ to stimuli s and s', we define the distance between them as $\|\boldsymbol{\phi}_s - \boldsymbol{\phi}_{s'}\|^2 = \frac{1}{N_c}\sum_{i=1}^{N_c}(\phi(c_i, s) - \phi(c_i, s'))^2$. The normalization factor $\frac{1}{N_c}$ allows us to take a limit as the number of cells increases, as described in more detail in section 3. We define the distance in this limit as the expected squared difference in the firing rate of a randomly chosen neuron: $\|\boldsymbol{\phi}_s - \boldsymbol{\phi}_{s'}\|^2 = \mathbb{E}_c\left[(\phi(c, s) - \phi(c, s'))^2\right]$. This distance makes the space of responses into a *Hilbert space*, a type of infinite-dimensional space with many properties similar to finite-dimensional Euclidean spaces. Hilbert spaces, like finite-dimensional spaces, have an inner product, $\langle \boldsymbol{\phi}_s, \boldsymbol{\phi}_{s'} \rangle = \mathbb{E}_c\left[\phi(c, s)\phi(c, s')\right]$. Like finite-dimensional vector spaces, they can be given an orthonormal basis, although now there may be an infinite number of basis vectors. A vector in a Hilbert space can be expressed as an infinite linear combination of basis vectors: $\boldsymbol{\phi} = \sum_{i=1}^{\infty} \phi_i \mathbf{u}_i$, and Pythagoras' theorem holds: $\|\boldsymbol{\phi}\|^2 = \sum_{i=1}^{\infty} \phi_i^2$. Unlike the finite-dimensional case, however not every set of coordinates defines a vector: a vector only belongs to the Hilbert space if it has finite length, $\sum_{i=1}^{\infty} \phi_i^2 < \infty$.

Recall that the *trace* of a finite-dimensional matrix is the sum of its diagonal elements. A standard theorem of linear algebra holds that the trace of a matrix equals the sum of its eigenvalues (e.g. Ref. [25], pp. 336). Traces can also be defined in infinite-dimensional Hilbert spaces. The infinite-dimensional equivalent of a matrix is an *operator*: a function that linearly maps infinite-dimensional vectors to infinite-dimensional vectors. For our current application, the most important operators are *integral operators*. For example, given a probability space $\mathcal{S}$ of possible stimuli, the space of all functions $a(s)$ with finite squared expectation ($\mathbb{E}_s\left[a(s)^2\right] < \infty$) defines an infinite-dimensional Hilbert space termed $L^2(\mathcal{S})$. The kernel function $K(s,s')$ defines an operator $\mathscr{K}$ on $L^2(\mathcal{S})$, as the linear mapping that takes a function $a(s)$ to the function $\mathscr{K}a(s) = \mathbb{E}_{s'}\left[K(s',s)a(s')\right]$. (It is called an integral operator because the expectation can be expressed as an integral over s'.) The trace of the operator $\mathscr{K}$ is defined as the sum of its diagonal elements with respect to any orthonormal basis e_i: $\text{Tr}\,(\mathscr{K}) = \sum_i \mathbb{E}_s\left[e_i(s)\mathscr{K}e_i(s)\right]$. As in finite dimensions, the trace of an operator is equal to the sum of its eigenvalues: $\text{Tr}\,(\mathscr{K}) = \sum_i \lambda_i$. However, unlike in finite dimensions, the trace need not be finite; there are an infinite number of eigenvalues, which might not have a finite sum. An operator whose trace is finite is called a *trace class* operator (Ref. [4], section 4.5). For integral operators with a continuous kernel, the trace can be expressed as an expectation along the diagonal (Ref. [4] theorem 4.6.7; Ref. [26], lemma to theorem XI.31):

$$\text{Tr}\,(\mathscr{K}) = \mathbb{E}_s\left[K(s,s)\right].$$

Because $K(s,s) = \mathbb{E}_c\left[\phi(c,s)^2\right]$, this establishes:

Theorem 3. *Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ be the eigenspectrum for a population code $\phi(c,s)$ whose kernel function $K(s,s')$ is continuous. Then*

$$\sum_{i=1}^{\infty} \lambda_i = \mathbb{E}_{s,c}\left[\phi(c,s)^2\right].$$

In particular, the eigenvalue sum is finite if and only if the population variance is finite. □

The relevance of this to eigenspectrum decay comes from the fact that $\sum_{n=1}^{\infty} n^{-1}$ is infinite. Thus, a continuous kernel and finite population variance means that the eigenspectrum must decay faster than this. Formally:

Corollary 3.1. *If $\mathbb{E}_{s,c}\left[\phi(c,s)^2\right]$ is finite and K is continuous then $\lambda_n = o(n^{-1})$.*

Proof. Recall that $\lambda_n = o(n^{-1})$ means that for any $\epsilon > 0$ there exists N such that $\lambda_n \leq \epsilon n^{-1}$ for all $n \geq N$. Theorem 3 shows $\sum_{i=1}^{\infty} \lambda_i$ is finite, thus for any ϵ there exists N such that $\sum_{i=n/2}^{\infty} \lambda_i \leq \epsilon/2$ for all $n \geq N$, where $n/2$ is understood to be rounded down for odd n. Because λ_n is non-negative and non-increasing, $n\lambda_n/2 \leq \sum_{i=n/2}^{n} \lambda_i \leq \sum_{i=n/2}^{\infty} \lambda_i \leq \epsilon/2$. Thus, $\lambda_n \leq \epsilon n^{-1}$ for all $n \geq N$. □

Thus, a code whose eigenspectrum decayed more slowly than n^{-1} would be pathological in one of two ways. The first way, a discontinuous kernel function, would mean no generalization: the response of the population to neighboring stimuli can be completely different, no matter how close these stimuli are. The completely uncorrelated code described in example 1 was an example of a code with discontinuous kernel function; we will make this statement mathematically precise in section 3.1. The second way, infinite population variance would mean neurons respond to preferred stimuli at arbitarily high rates. We will provide an example of this other form of pathological code as example 5.

2.6 Relating fractal dimension to covariance eigenvalues

The previous section showed that population codes with eigenspectra decaying slower than n^{-1} are highly pathological, exhibiting either kernel discontinuity or infinite variance. However, the potential pathologies of neural codes are not restricted to these extreme problems. A neural population code can exhibit a more subtle form of pathology by failing to be *differentiable*. Functions can be continuous but not differentiable: for example, the function $y = \sqrt[3]{x}$ is continuous but not differentiable at $x = 0$, as the curve has an infinite slope at this point. Furthermore, there are functions that are continuous everywhere but differentiable nowhere [27]. While such non-differentiable functions are surprising, they can be useful models of some phenomena occurring in nature. Indeed, because fractals cannot have differentiable coordinate functions, objects such as coastlines can be modeled by functions that are continuous but not differentiable.

In infinite dimensions, the concept of differentiability has an extra subtlety. Consider a vector $\boldsymbol{\phi}(x) = \sum_{i=1}^{\infty} \phi_i(x)\mathbf{u}_i$, in a Hilbert space with orthonormal basis $\{\mathbf{u}_i\}$, depending on a scalar parameter x. As in finite dimensions, we can formally express the derivative of the vector as $\frac{d\boldsymbol{\phi}}{dx} = \sum_{i=1}^{\infty} \frac{d\phi_i}{dx}\mathbf{u}_i$. In a finite-dimensional space, $\boldsymbol{\phi}(x)$ will be differentiable if all of its coordinates $\phi_i(x)$ are differentiable. But in infinite dimensions this need not be the case: $\boldsymbol{\phi}(x)$ can fail to be differentiable even if every coordinate function is differentiable, if the infinite sum $\sum_{i=1}^{\infty} \left|\frac{d\phi_i}{dx}\right|^2$ does not converge.

We now prove two theorems characterizing the relationship between eigenspectrum decay and differentiability of the population code. In the first we prove that the eigenspectrum of neuronal responses constrained to a set of fractal dimension d must decay at least as fast as a power law of exponent $\alpha = 1 + 2/d$. This means that if the eigenspectrum of a code of manifold dimension d decays slower than this, the response set must be fractal.

There are several ways of formalizing the notion of fractal dimension (See e.g. Refs. [8, 9]), all based on the idea that the volume of a d-dimensional object with diameter δ should scale as δ^d. Here, we will use the *upper Minkowski dimension*. Given a subset $\mathcal{F}$ of a metric space, we define the the *covering number* $N_\delta(\mathcal{F})$ to be the smallest number of spheres of diameter δ required to cover $\mathcal{F}$. Intuitively, $N_\delta(\mathcal{F})$ should scale as δ^{-d} for a space of dimension d. The upper Minkowski dimension of $\mathcal{F}$ is written as $\overline{\dim}_M\mathcal{F}$, and defined to be $\inf\{d' : \limsup_{\delta\to 0} N_\delta(\mathcal{F})\delta^{d'} = 0\}$. This means that for any $d' > \overline{\dim}_M\mathcal{F}$ and for any $C > 0$, there exists an $\epsilon > 0$ such that for all $\delta < \epsilon$, $N_\delta(\mathcal{F})\delta^{d'} < C$; furthermore $\overline{\dim}_M\mathcal{F}$ is the smallest number with this property. We can now prove our theorem.

Theorem 4. *Let $\boldsymbol{\phi}(s)$ be a function of a random element s, supported on a set $\mathcal{F}$ of upper Minkowski dimension d inside a separable Hilbert space $\mathbb{H}$, with $\mathbb{E}_s\left[\|\boldsymbol{\phi}(s)\|^2\right] < \infty$. Write the eigenvalues of Cov$(\boldsymbol{\phi}(s))$ as $\lambda_1 \geq \lambda_2 \geq \ldots$. Then for all $d' > d$, $\lambda_n = O(n^{-1-2/d'})$ as $n \to \infty$.*

Proof. We assume without loss of generality that $\mathbb{E}_s[\boldsymbol{\phi}(s)] = 0$. Fix $d' > \overline{\dim}_M\mathcal{F}$ and $C > 0$. By the defintion of upper Minkowski dimension, there exists n_0 such that for any $n \geq n_0$, we can cover $\mathcal{F}$ with n balls of diameter at most $Cn^{-1/d'}$. For each s, let b_s be an integer between 1 and n identifying a ball in this cover that contains $\boldsymbol{\phi}(s)$. Define a mean vector for each ball: $\boldsymbol{\mu}_s = \mathbb{E}_{s'}[\boldsymbol{\phi}(s')|b_{s'} = b_s]$. We can decompose the variance of $\boldsymbol{\phi}(s)$ using an ANOVA decomposition:

$$\mathbb{E}_s\left[\|\boldsymbol{\phi}(s)\|^2\right] = \mathbb{E}_s\left[\|\boldsymbol{\phi}(s) - \boldsymbol{\mu}_s\|^2\right] + \mathbb{E}_s\left[\|\boldsymbol{\mu}_s\|^2\right]. \tag{4}$$

Let $\mathscr{P}$ be the projection operator onto the $\leq n$-dimensional subspace spanned by mean vectors $\{\boldsymbol{\mu}_s\}$. Then the total variance of $\boldsymbol{\phi}(s)$ in this subspace is at least as big as the variance between means:

$$\mathbb{E}_s\left[\|\mathscr{P}\boldsymbol{\phi}(s)\|^2\right] = \mathbb{E}_s\left[\|\mathscr{P}(\boldsymbol{\phi}(s) - \boldsymbol{\mu}_s)\|^2\right] + \mathbb{E}_s\left[\|\boldsymbol{\mu}_s\|^2\right] \geq \mathbb{E}_s\left[\|\boldsymbol{\mu}_s\|^2\right].$$

The variance in any n-dimensional subspace cannot exceed the sum of the first n covariance eigenvalues: $\mathbb{E}_s\left[\|\mathscr{P}\boldsymbol{\phi}(s)\|^2\right] \leq \sum_{i=1}^{n} \lambda_i$ (e.g. Ref. [4], theorem 7.2.8). Thus,

$$\sum_{i=1}^{n} \lambda_i \geq \mathbb{E}_s\left[\|\boldsymbol{\mu}_s\|^2\right]. \tag{5}$$

Write the sum of all eigenvalues above n as $\Lambda_n = \sum_{i>n} \lambda_i$. Because the eigenvalues must sum to the total variance, $\sum_{i=1}^{\infty} \lambda_i = \mathbb{E}_s\left[\|\boldsymbol{\phi}(s)\|^2\right]$, combining (4) and (5) gives $\Lambda_n \leq \mathbb{E}_s\left[\|\boldsymbol{\phi}(s) - \boldsymbol{\mu}_s\|^2\right]$. Furthermore, because the balls have diameter at most $Cn^{-1/d'}$, this implies that for all $n \geq n_0$,

$$\Lambda_n \leq C^2 n^{-2/d'}$$

Informally we can see why λ_n should follow an $n^{-1-2/d'}$ power law by differentiating with respect to n. Formally, consider any even $n > 2n_0$. Then $\Lambda_{n/2} = \sum_{i=n/2+1}^{\infty} \lambda_i \geq \sum_{i=n/2+1}^{n} \lambda_i \geq \frac{n}{2}\lambda_n$, as the λ_i are positive and non-increasing. Thus $\lambda_n \leq \frac{2}{n}\Lambda_{n/2} \leq \frac{2}{n}C^2(\frac{n}{2})^{-2/d'} = 2^{1+2/d'}C^2 n^{-1-2/d'}$. Because C was arbitrary, $\lambda_n = O(n^{-1-2/d'})$. □

We now prove a corollary to this theorem, which explains its relevance to stimulus encoding. Suppose a stimulus is drawn from a manifold of dimension d, and mapped to the neural space by a code with bounded derivative (which means it satisfies a property of Lipschitz continuity). The corollary shows the response set must have fractal dimension $\leq d$, and thus an eigenspectrum decaying faster than $n^{-1-2/d'}$ for any $d' > d$. Thus, even though theorem 4 only speaks to the geometry of the response set, it has implications for the neural coding of sensory stimuli. These implications will be discussed further in the next section.

Corollary 4.1. *If $\mathcal{F} \subset \mathbb{H}$ is a the image of a set of upper Minkowski dimension* d *under a Lipschitz map, or if $\mathcal{F}$ is a* d*-dimensional compact differentiable submanifold of $\mathbb{H}$, then for all* $\mathrm{d}' > d$*, $\lambda_n \sim O(n^{-1-2/d'})$.*

Proof. In both cases we must show that $\mathcal{F}$ has upper Minkowski dimension at most d. It is a standard result that if f is a Lipschitz map, $\overline{\dim}_M(f(\mathcal{F})) \leq \overline{\dim}_M(\mathcal{F})$, and that if f is bi-Lipschitz, $\overline{\dim}_M(f(\mathcal{F})) = \overline{\dim}_M(\mathcal{F})$ (Ref. [9], Prop. 2.5).

By a d-dimensional differentiable submanifold of $\mathbb{H}$ we mean a set $\mathcal{F} \subset \mathbb{H}$, where every point is contained in an open set (of the subspace topology on $\mathcal{F}$) that bijects to an open set of $\mathbb{R}^d$ via a chart with a continuous Frechet derivative that is nonzero everywhere. To show such a set has upper Minkowski dimension d, observe that each chart must be bi-Lipschitz, and thus each point is contained in an open set of upper Minkowski dimension d. By compactness, $\mathcal{F}$ is therefore covered by finite number of sets of upper Minkowski dimension d, and because $\overline{\dim}_M(\mathcal{F}_1 \cup \mathcal{F}_2) = \max(\overline{\dim}_M\mathcal{F}_1, \overline{\dim}_M\mathcal{F}_2)$ (Ref. [9], p.35), $\overline{\dim}_M\mathcal{F} = d$. □

2.7 Insufficient decay means infinite derivative

To gain additional insight into the $n^{-1-2/d}$ bound, we now explicitly show how if eigenvalues did not decay this fast, the expected squared derivative of the neural code would have to be infinite. To do so, we consider the space of stimuli $\mathcal{S}$ to be a Riemannian manifold. A Riemannian manifold has more structure than a differentiable manifold: a differentiable manifold allows one to compute a derivative, but a Riemannian manifold also allows one to measure its magnitude. We will make use of three theorems from Riemannian geometry. The first of these, Green's theorem (Ref. [24], theorem III.10), provides an estimate of the magnitude of the derivative of a smooth function on a Riemannian manifold. If $f(s)$ is a smooth and and compactly supported function on a Riemannian manifold, Green's theorem implies

$$\int |\nabla f(s)|^2 d\mu(s) = \int f(s) \triangle f(s) d\mu(s).$$

Here, $|\nabla f(s)|^2$ represents the squared magnitude of the derivative of f with regards to the Riemannian metric, and $\int d\mu(s)$ represents integration on the manifold using the canonical measure defined by the Riemannian metric. $\triangle$ represents the Laplace-Beltrami operator, a generalization of the Laplacian operator $-\sum_i \frac{\partial^2}{\partial x_i^2}$ familiar from physics in Euclidean spaces, to the case of general Riemannian manifolds.

The second standard result we use concerns the eigenvalues of the Laplace-Beltrami operator. The operator has a set of eigenfunctions $w_n, n \in \mathbb{N}$, with corresponding eigenvalues η_n, such that

$$\triangle w_n(s) = \eta_n w_n(s)$$

The eigenfunctions w_n form a basis for $L^2(\mathcal{S})$, the Hilbert space of square-integrable functions on the manifold, and the eigenvalues $0 \leq \eta_1 \leq \eta_2 \ldots$ are non-negative, increasing, and unbounded (Ref. [24], theorem III.18).

The third result we require is Weyl's asymptotic law (Ref. [24], theorem III.36), which relates the growth of the Laplace-Beltrami eigenvalues to the dimension of the manifold. This law states that for a mainfold of dimension d,

$$\eta_n \sim n^{2/d}$$

In other words, there exists a constant $C_{\mathcal{S}}$ such that $\lim_{n\to\infty} \eta_n n^{-2/d} = C_{\mathcal{S}}$. (The value of $C_{\mathcal{S}}$ can be computed using the total volume of the manifold, but its precise value does not concern us here.) This is a remarkable result, as it means the asymptotic distribution of larger eigenvalues does not depend on the shape of the manifold, only its volume and dimensionality.

With these preliminaries, we are now ready to prove our next theorem. We consider stimuli to be drawn at random from a closed, compact d-dimensional Riemannian manifold $\mathcal{S}$ according to its canonical Riemannian measure. We assume that the responses have finite squared expectation: $\mathbb{E}_s\left[\|\boldsymbol{\phi}(s)\|^2\right] < \infty$. This activity defines a kernel function $K(s,s') = \langle \boldsymbol{\phi}(s), \boldsymbol{\phi}(s)\rangle$, with positive bounded eigenvalues λ_n, arranged in decreasing order: $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. Then we have

Theorem 5. *If the expected derivative magnitude of $\boldsymbol{\phi}$ is finite, i.e. $\mathbb{E}_s\left[\|\nabla_s \boldsymbol{\phi}(s)\|^2\right] < \infty$, then $\lambda_n = o(n^{-1-2/d})$.*

Proof. Our proof rests on relating the kernel eigenfunctions $v_i(s)$ to the Laplacian eigenfunctions $w_i(s)$. Since they are both orthonormal bases for $L^2(\mathcal{S})$, we may define a set of orthonormal matrix elements $\{a_{i,j}\}$ such that $v_i = \sum_j a_{i,j} w_j$ and $w_j = \sum_i a_{i,j} v_i$, with $\sum_i a_{i,j} a_{i,k} = \delta_{j,k}$, and $\sum_j a_{i,j} a_{k,j} = \delta_{i,k}$.

Because the kernel eigenfunctions $v_i(s)$ are an othonormal basis for $L^2(\mathcal{S})$, we can write $\boldsymbol{\phi}(s) = \sum_i \mathbf{u}_i v_i(s)$, where the Hilbert space vector $\mathbf{u}_i = \mathbb{E}_s\left[\boldsymbol{\phi}(s) v_i(s)\right]$. These vectors have magnitude

$$\|\mathbf{u}_i\|^2 = \mathbb{E}_{s,s'}\left[\langle \boldsymbol{\phi}(s), \boldsymbol{\phi}(s')\rangle v_i(s) v_i(s')\right] = \mathbb{E}_{s,s'}\left[K(s,s') v_i(s) v_i(s')\right] = \lambda_i,$$

since the v_i are kernel eigenfunctions. Thus the expected derivative magnitude

$$\begin{aligned}
\mathbb{E}_s\left[|\nabla_s \boldsymbol{\phi}(s)|^2\right] &= \mathbb{E}_s\left[\sum_i \|\mathbf{u}_i\|^2 |\nabla v_i(s)|^2\right] \\
&= \mathbb{E}_s\left[\sum_i \lambda_i v_i(s) \triangle v_i(s)\right] \\
&= \mathbb{E}_s\left[\sum_i \lambda_i \left(\sum_j a_{i,j} w_j(s)\right) \triangle \left(\sum_k a_{i,k} w_k(s)\right)\right] \\
&= \mathbb{E}_s\left[\sum_{i,j,k} \lambda_i a_{i,j} a_{i,k} w_j(s) \triangle w_k(s)\right] \\
&= \mathbb{E}_s\left[\sum_{i,j,k} \lambda_i \eta_k a_{i,j} a_{i,k} w_j(s) w_k(s)\right] \\
&= \sum_{i,j,k} \lambda_i \eta_k a_{i,j} a_{i,k} \mathbb{E}_s\left[w_j(s) w_k(s)\right] \\
&= \sum_{i,j} \lambda_i \eta_j a_{i,j}^2
\end{aligned}$$

Recall that the kernel eigenvalues λ_i are a non-increasing sequence, while the Laplacian eigenvalues η_j increase asymptotically as $n^{2/d}$. Informally, we can guess that the orthonormal matrix $a_{i,j}$ minimizing this sum will be the identity, so the sum only converges if $\lambda_n \eta_n = o(n^{-1})$ and thus $\lambda_n = o(n^{-1-2/d})$.

Formally, we define $A_j = \sum_i \lambda_i a_{i,j}^2$, so $\mathbb{E}_{s,c}\left[|\nabla_s \phi(c,s)|^2\right] = \sum_j A_j \eta_j$. Next, consider any integer $n \geq 1$, and define $B_i = \sum_{j \leq n} a_{i,j}^2$. By orthonormality of $a_{i,j}$ we therefore have $B_i \leq 1$ for all i, and also $\sum_{i=1}^{\infty} B_i \leq n$. By definition,

$$\sum_{j\leq n} A_j = \sum_i \lambda_i B_i$$

Thus, $\sum_{j\leq n} A_j$ cannot be more than the largest possible value of $\sum_i \lambda_i b_i$ over all b_i satisfying the same constraints as B_i: $b_i \leq 1$ for all i and $\sum_i b_i \leq n$. Because the λ_i are non-increasing, this maximum is achieved when $b_i = 1$ for $i \leq n$ and zero thereafter, and thus

$$\sum_{j\leq n} A_j \leq \sum_{i\leq n} \lambda_i. \tag{6}$$

Write $\Lambda_n = \sum_{i>n} \lambda_i$. By orthonormality of $a_{i,j}$, $\sum_{j=1}^{\infty} A_j = \sum_{i=1}^{\infty} \lambda_i$. Therefore (6) means that $\sum_{j>n} A_n \geq \Lambda_n$, and so

$$\eta_n \Lambda_n \leq \mathbb{E}_{c,s}\left[|\nabla_s \phi(c,s)|^2\right] - \sum_{j\leq n} \eta_j A_j$$

Because $\sum_{j=1}^{\infty} \eta_j A_j$ converges to $\mathbb{E}_{c,s}\left[|\nabla_s \phi(c,s)|^2\right]$, if this limit is finite then $\lim_{n\to\infty} \eta_n \Lambda_n = 0$. Because $\eta_n \sim n^{2/d}$, this means that $\Lambda_n = o(n^{-2/d})$. The same argument used at the end of theorem 4 now tells us that $\lambda_n = o(n^{-1-2/d})$. □

Note that the eigenvalue decay $\lambda_n = o(n^{-1-2/d})$ that this theorem tells us is required for finite derivative, is faster than the decay theorem 4 tells us is required for Minkowski dimension d (for all $d' > d$, $\lambda_n = O(n^{-1-2/d'})$). The following example illustrates both theorems.

Example 2. We now consider a simple example, in which a 1-dimensional circle is continuously mapped into Hilbert space, with a eigenspectrum $n^{-\alpha}$ by construction. This will show the bounds of theorem 4 cannot be improved, in the sense that for any $d \geq 1$, there is a probability distribution supported on an $\mathcal{F} \subset \mathbb{H}$ with $\overline{\dim}_M(\mathcal{F}) = d$ and covariance eigenvalues that decay $\sim n^{-1-2/d}$. This example is illustrated in Fig. 4d-f.

Parametrize the circle by an angle s, uniformly distributed between 0 and 2π. We define a vector $\boldsymbol{\phi}(s)$ in an infinite-dimensional Hilbert space, such that the covariance eigenvalues follow a power law $n^{-\alpha}$. Specifically, for $n \geq 1$:

$$\boldsymbol{\phi}(s)_{2n-1} = \frac{\cos(ns)}{n^{\alpha/2}}$$
$$\boldsymbol{\phi}(s)_{2n} = \frac{\sin(ns)}{n^{\alpha/2}}$$

The dimensions of $\boldsymbol{\phi}$ are uncorrelated, and computing their variances shows that $\lambda_{2n-1} = \lambda_{2n} = n^{-\alpha}/2$. The sine and cosine functions in each coordinate of $\boldsymbol{\phi}$ are the Laplacian eigenfunctions: the eigenfunction $\cos(ns)$ satisfies $\triangle \cos(ns) = n^2 \cos(ns)$.

Observe that $\|\boldsymbol{\phi}(s)\|^2 = \zeta(\alpha)$, where $\zeta(\alpha) = \sum_{n\geq 1} n^{-\alpha}$ is the Riemann zeta function. Thus, consistent with theorem 3, population activity only has finite expectation if $\alpha > 1$. The derivative of the population code, $\left\|\frac{d\boldsymbol{\phi}(s)}{ds}\right\|^2 = \sum_{n\geq 1} n^{-(\alpha-2)}$. Thus, consistent with theorem 5, the derivative only has finite expectation if $\alpha > 3$.

We now compute $\overline{\dim}_M(\mathcal{F})$. First we compute the distance between two points in $\mathcal{F}$. By circular symmetry we may without loss of generality set one of them to 0. We obtain

$$\begin{aligned} D(s)^2 = \|\boldsymbol{\phi}(s) - \boldsymbol{\phi}(0)\|^2 &= \sum_{n\geq 1} \frac{(\cos(ns)-1)^2 + \sin^2(ns)}{n^{\alpha}} \\ &= \sum_{n\geq 1} \frac{2 - 2\cos(ns)}{n^{\alpha}} \\ &= 2\zeta(\alpha) - \mathrm{Li}_{\alpha}(e^{is}) - \mathrm{Li}_{\alpha}(e^{-is}) \end{aligned}$$

where $\mathrm{Li}_{\alpha}(z) = \sum_{n\geq 1} z^n n^{-\alpha}$ is the polylogarithm function.

To compute $\overline{\dim}_M(\mathcal{F})$ we consider how $D(s)^2$ depends on s as $s \to 0$. For non-integer α, the polylogarithm has a series expansion [28]:

$$\mathrm{Li}_\alpha(e^x) = \Gamma(1-\alpha)(-x)^{\alpha-1} + \sum_{k\geq 0} \frac{\zeta(\alpha-k)}{k!} x^k.$$

Substituting this in, we obtain

$$\begin{aligned} D(s)^2 &= -\Gamma(1-\alpha)s^{\alpha-1}(i^{\alpha-1} + (-i)^{\alpha-1}) + \zeta(\alpha-2)s^2 + O(s^4) \\ &= -2\Gamma(1-\alpha)s^{\alpha-1}\cos((\alpha-1)\pi/2) + \zeta(\alpha-2)s^2 + O(s^4) \end{aligned}$$

For any value of s we can cover $\mathcal{F}$ with $2\pi/s$ balls no smaller than $D(s)$. Thus, we have upper Minkowski dimension of d if $D(s) \sim s^{1/d}$ as $s \to 0$. For $\alpha > 3$, the dominant term as $s \to 0$ is $\zeta(\alpha-2)s^2$, so $D(s) \sim s$, and we have $\overline{\dim}_M(\mathcal{F}) = 1$. However, for $1 < \alpha < 3$, the first term of order $s^{\alpha-1}$ dominates, so $\overline{\dim}_M(\mathcal{F}) = \frac{2}{\alpha-1}$, which is greater than 1 indicating a fractal structure. At the critical value of $\alpha = 3$, we can use a Laurent expansion of the the ζ and Γ functions to obtain $\Delta(s) \sim s^2 \log \frac{1}{s}$. Thus for $\alpha = 3$, the dimension is still 1, even though $\boldsymbol{\phi}(s)$ is not differentiable.

Recall that theorem 4 predicts that for dimension d, we must have $\alpha \geq 1 + 2/d$. This is consistent with the above calculations: $d = 1$ when $\alpha \geq 3$, and $d = 2/(\alpha-1)$ when $1 < \alpha < 3$. Thus the bound is saturated for $1 < \alpha \leq 3$, but loose for $\alpha > 3$.

Example 3. In the previous example we did not explicitly model the activity of individual neurons: instead, we simply constructed a Hilbert space corresponding to their principal components, to demonstrate how differentiability and fractal dimension relate to power-law eigenspectrum decay. We now consider an example explicitly constructed from the activity of a modeled neural population with Gaussian receptive fields; this "radial basis kernel" representation is also often used in machine learning [19, 21]. We show that if all cells' receptive fields have the same size, the eigenspectrum is not a power law; however, by mixing together neurons with different radii in a scale-free manner, we obtain power-law decay, with differentiability for $\alpha > 1 + 2/d$.

Let the stimulus $\mathbf{s}$ be a point in a d-dimensional vector space, and let cell c respond according to a spherical Gaussian function, with center $\mathbf{x}_c$ and variance W. Thus the cell's response is

$$\phi(c, \mathbf{s}) = \frac{A_W}{(2\pi W)^{d/2}} e^{-|\mathbf{x}_c - \mathbf{s}|^2/2W} = A_W N(\mathbf{s}; \mathbf{x}_s, W),$$

where $N(\mathbf{x}; \mu, \sigma^2)$ represents a spherical multivariate Gaussian density of mean μ and variance σ^2, and A_W is an amplitude scaling factor that can potentially depend on the response width W. Note that this is not a model of image processing specifically: the Gaussians are not linear filters applied to an 2d image, but radial basis function units working on an abstract representation. Thus, for natural image stimuli, the stimulus space could have dimension $d \gg 2$, for example representing the results of preprocessing by a convolutional network, and the radial basis function units would then produce nonlinear receptive fields.

We will pick A_W to ensure that a cell's mean p^{th}-power activity is independent of scale: $\mathbb{E}_s\left[\phi(c, s)^p\right] = 1$ for all cells c, where p is a parameter of the model. If $p = 1$, this corresponds to the standard Gaussian normalization; while if $p = 2$, the cells' mean square firing is independent of scale. A straightforward calculation then shows that

$$A_W = V^{\frac{1}{p}} (2\pi W)^{\frac{d}{2}(1-\frac{1}{p})} p^{\frac{d}{2p}}, \tag{7}$$

where we have assumed that the stimuli are evenly distributed over a volume V. Observe that if $p = 1$, a cell's peak firing rate is larger the more sharply it is tuned (in order to obtain constant mean rate); if $p > 1$ then peak rate grows more slowly with tuning sharpness, and if $p = \infty$ then peak activity is independent of width.

Now let us compute the kernel function for two stimuli $\mathbf{s}_1$ and $\mathbf{s}_2$. We assume that the cell centers x_c are evenly distributed over the volume V, and for now consider all cells to have a fixed width W. Then

$$\begin{aligned} K_W(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{E}_c\left[\phi(c, \mathbf{s}_1), \phi(c, \mathbf{s}_2)\right] &= A_W^2 V^{-1} \int N(\mathbf{s}_1; \mathbf{x}_c, W) N(\mathbf{s}_2; \mathbf{x}_c, W) d^d\mathbf{x}_c \\ &= A_W^2 V^{-1} N(\mathbf{s}_1; \mathbf{s}_2, 2W), \end{aligned}$$

where we have used the fact that the convolution of two Gaussian densities is a Gaussian of summed variance.

Now, the eigenspectrum of a translation-invariant kernel can be found simply by Fourier transformation [19, 21]. Recalling that the Fourier transform of $N(\mathbf{x}; 0, \sigma^2)$ at spatial frequency $\mathbf{k}$ is $e^{-\sigma^2|\mathbf{k}|^2/2}$, we have

$$\hat{K}_W(\mathbf{k}) = A_W^2 V^{-1} e^{-W|\mathbf{k}|^2}.$$

Plugging the distribution of spatial frequencies $\mathbf{k}$ into this function yields the eigenspectrum of the kernel function. Note that the calculation we just performed appears to give a continuous eigenspectrum. However, this is because of the "physicists'" approximation we made earlier, assuming that the volume occupied by potential stimuli was large but not explicitly modeling a probability distribution for the stimuli $\mathbf{x}_c$; had we done so, we could have obtained a discrete spectrum, but at the cost of much more work. Continuing in this informal vein, we note that the number of possible spatial frequencies of magnitude at most $|\mathbf{k}|$ will scale as $g|\mathbf{k}|^d$, where g is a geometric scale factor related to the shape and size of the distribution of $\mathbf{x}_c$. Thus, the n^{th} eigenfunction has spatial frequency magnitude $|\mathbf{k}| = (n/g)^{1/d}$. We therefore find that the eigenspectrum is

$$\lambda_n \propto \exp\left(-W(n/g)^{2/d}\right).$$

This is not a power law. For $n \ll gW^{d/2}$ it is approximately constant (as can be seen by Taylor expansion of the exponential), but for $n > gW^{d/2}$ it falls faster than any power. Thus, the eigenspectrum of a population code constructed from Gaussian neurons of a single width can be approximated by two regimes. At long distance scales, the eigenspectrum is flat, indicating a completely different population response for any two stimuli separated substantially more than the tuning width. At short distance scales however the eigenspectrum decays rapidly, indicating that the population code is barely sensitive to distances substantially shorter than the tuning width.

We may generate a power-law eigenspectrum by now considering a mixture of radial units with different variances. We model the distribution of response widths as following a power law density:

$$f(W) = \begin{cases} BW^{\gamma-1}, & \text{if } W_0 \leq W \leq W_1 \\ 0, & \text{otherwise,} \end{cases}$$

where the normalization factor $B = \gamma / \left(W_1^\gamma - W_0^\gamma\right)$. The parameter γ measures the relative distribution of sharply and broadly tuned neurons in the population: values of $\gamma > 1$ indicate a bias towards neurons of large radii, and $\gamma < 0$ indicate a bias towards small radii. We will see that the allowed range of radii $W_0 \ldots W_1$ is not important provided W_0 is small enough and W_1 large enough.

We can compute the kernel corresponding to this scaled distribution as

$$\begin{aligned} \hat{K}(\mathbf{k}) = \mathbb{E}_W\left[\hat{K}_W(\mathbf{k})\right] &= \int_{W_0}^{W_1} A_W^2 V^{-1} e^{-W|\mathbf{k}|^2} BW^{\gamma-1} dW \\ &= C \int_{W_0}^{W_1} e^{-W|\mathbf{k}|^2} W^{d(1-1/p)+\gamma-1} dW, \end{aligned}$$

where the constant C captures all factors with no dependence on W or $|\mathbf{k}|$. Making a substitution $z = |\mathbf{k}|^2\, W$, we have

$$\hat{K}(\mathbf{k}) = C'\, |\mathbf{k}|^{-2d(1-1/p)-2\gamma} \int_{z_0}^{z_1} e^{-z} z^{d(1-1/p)+\gamma-1} dz. \tag{8}$$

Now, if $d(1-1/p)+\gamma > 0$, the above integral will not depend on the exact values of its limits $z_0 = |\mathbf{k}|^2\, W_0$ and $z_1 = |\mathbf{k}|^2\, W_1$: provided they are small and large enough, the integral will converge to a gamma function, $\Gamma(d(1-1/p)+\gamma)$, which shows no dependence on $|\mathbf{k}|$. We thus have $\hat{K}(\mathbf{k}) \propto |\mathbf{k}|^{-2d(1-1/p)-2\gamma}$, and

$$\lambda_n \propto n^{-\alpha},$$

where

$$\alpha = 2 - \frac{2}{p} + \frac{2\gamma}{d}. \tag{9}$$

Thus, if the basis function radii are distributed according to a scale-free power law, we obtain a power law eigenspectrum also.

To check that $\alpha > 1 + 2/d$ corresponds to finite expected derivative, we may differentiate ϕ explicitly. We find that

$$\mathbb{E}_s \left[\left| \frac{\partial \phi(c, \mathbf{s})}{\partial \mathbf{s}}^2 \right| \right] = \mathbb{E}_W \left[\frac{A_W^2}{(2\pi W)^d W^2} \mathbb{E}_{\mathbf{s}} \left[|\mathbf{x}_c - \mathbf{s}|^2 e^{-|\mathbf{x}_c - \mathbf{s}|^2/W} \right] \right]$$

Because $\mathbb{E}_{\mathbf{s}} \left[|\mathbf{x}_c - \mathbf{s}|^2 e^{-|\mathbf{x}_c - \mathbf{s}|^2/W} \right] \propto W^{1+d/2}$, and using (7),

$$\begin{aligned} \mathbb{E}_s \left[\left| \frac{\partial \phi(c, \mathbf{s})}{\partial \mathbf{s}}^2 \right| \right] &\propto \mathbb{E}_W \left[W^{d(1/2-1/p)-1} \right] \\ &\propto \int_{W_0}^{W_1} W^{\gamma + d(1/2-1/p)-2} \, dW, \end{aligned}$$

This integral will diverge as $W_0 \to 0$ unless $\gamma > 1 - d(1/2 - 1/p)$, which from (9) we see is equivalent to $\alpha > 1 + 2/d$. Thus, confirming theorem 5, the expected derivative will be infinite if eigenvalues do not decay faster than $n^{-1-2/d}$, and this infinite derivative will reflect unbounded contributions from the neurons with the smallest values of W, i.e. neurons of sharpest tuning.

3. The $N \to \infty$ limit

To make formal statements about the asymptotic decay of the eigenspectrum requires defining a limit in which both the number of neurons recorded, and the number of stimuli presented grow large. While most statistical analyses involve a limit of a large number of observations, the fact that we consider two such limits means we must consider not just randomly distributed numbers, but randomly distributed vectors in infinite-dimensional spaces. This approach, known as functional data analysis, has not yet been commonly applied in neuroscience, but has a well-developed theoretical basis thanks to applications in other domains [29, 4].

The logic underlying functional data analysis is the standard logic of statistical inference: using the properties of a limited sample to make inferences about a larger, unobserved population. Specifically, by analyzing how the particular neurons we recorded responded to the particular stimuli we presented, we infer facts about how a larger population of neurons would respond to a larger ensemble of stimuli. To do so, we employ a standard assumption: that our sample of neurons and stimuli are randomly sampled from the larger population and the larger ensemble.

A simple example of inference by random sampling would be estimating the distribution of lengths our mice's tails. If we were to say that these lengths followed a Gaussian distribution (or indeed any continuous distribution), that could not be correct: we studied a finite number of mice, so their tail lengths follow a discrete distribution. However, we can use the sample we measured to make an inference about the population they are drawn from. This larger population is not just the set of all mice in the lab colony, or even all mice currently alive, both of which are finite. It is a hypothetical infinite population of all mice that could have been tested – only such an infinite population can follow a continuous probability distribution such as a Gaussian.

The assumption of such an infinite population can have counter-intuitive consequences, some of which are relevant for our theorems. For example, distributions of wealth and income can be modeled by a Pareto distribution[30]: $\mathbb{P}(x) \propto x^{-1-\gamma}$. When $\gamma \leq 2$, this distribution has infinite variance, $\mathbb{E}\left[x^2\right] = \infty$. Any finite sample drawn from this distribution will have finite variance, but the variance of the sample will continue increasing without bound as the sample size grows. Conversely, a statistician who found that sample variance converges to a finite limit as sample size grows could infer that the population has finite variance.

In our case, we consider the visual stimuli presented to be drawn at random from a probability distribution. We would like to make an inference about neural coding of not just these stimuli, nor even of the entire ImageNet database, but rather of a hypothetical infinite population of similar natural images we could in principle have presented. If we find that properties of the observed neural code converge to a limit as the number of images analyzed increases, we infer that it can be modelled by a probability distribution predicting this limit.

We similarly consider the recorded neurons to be drawn randomly from a population distribution. Again, this population is not just the full set of V1 neurons in the particular mouse studied, but the infinite set of neurons which that mouse's brain might have contained, consistent with our experimental sample. Again, if we find that properties of the observed neural code converge to a limit as the number of neurons analyzed increases, we infer that this infinite population can be modelled by a probability distribution predicting this limit.

The setting for our theorems thus involves three elements. First, a probability distribution for possible stimuli $\mathbb{P}(s)$, over a hypothetically infinite stimulus set $\mathcal{S}$; second, a probability distribution for possible recorded cells $\mathbb{P}(c)$, over a hypothetically infinite cell set $\mathcal{C}$; and third, a function $\phi(c, s)$ that predicts the expected firing rate of cell c to stimulus s. (As before, $\phi(c, s)$ describes the neuron's deterministic mean response to the stimulus, excluding neuronal noise). It is of course impossible to obtain a complete description of these three elements, which would summarize the response of any conceivable V1 cell to any conceivable natural image stimulus. However, it is possible to infer some of their properties using the logic of random sampling described above.

3.1 Fully independent coding revisited

The random sampling framework allows us to mathematically formalize the scenario of example 1, showing that if a set of neurons independently code a single continuous variable, the kernel function becomes discontinuous in a limit that the number of neurons tends to infinity. Constructing this limit requires some mathematical subtleties. If we considered the neurons as sampled with replacement from a finite set of N possibilities, the eigenspectrum would not be flat: once the sample becomes larger than N it would contain multiple cells whose activity would be perfectly correlated. The eigenspectrum would thus be flat for the first N eigenvalues, and exactly zero thereafter. Similarly, if we considered the neurons to be drawn from a discrete and (countably) infinite distribution, a large enough population would contain neurons with identical tuning properties, and thus all neurons could not be independent. This means that the binary coding scheme described in example 1 cannot be taken to a limit of infinite neurons.

Example 4. We therefore consider neurons as drawn from a continuous probability distribution. Again with the stimulus s uniformly distributed between 0 and 1, let each cell c respond to each possible stimulus s with an independent standard Gaussian variate. Cell c's tuning function $\phi(c, s)$ is then with probability 1 discontinuous, and not even measurable (a set theoretic requirement that allows it to be integrated). The space of possible neural responses is thus not a subset of the Hilbert space $L^2(\mathcal{S})$, but rather of a much larger space $\mathcal{C} = \mathbb{R}^{[0,1]}$ of all set-theoretic functions from the interval $[0, 1]$ to the real numbers. The kernel function is $K(s, s') = 1_{s=s'}$, the function that is 1 if $s = s'$, and 0 otherwise, reflecting the fact that neurons have variance 1, but that population responses to two unequal stimuli are completely uncorrelated, however similar these stimuli are. This kernel function is discontinuous, and the code is completely unable to generalize; a response learned to the representation of one stimulus has absolutely no bearing on the response to other stimuli.

The eigenvalues of the kernel function are all zero, since the function $1_{s_1=s_2}$ is zero everywhere except a set of measure 0. Thus, $\sum_{i=1}^{\infty} \lambda_i \neq \mathbb{E}_{s,c}\left[\phi(c, s)^2\right]$, but theorem 3 is not violated since the requirement on kernel continuity was not satisfied. To understand how the kernel eigenvalues can be zero despite unit population variance, consider the eigenvalues of the sample kernel matrix $K_{i,j} = \mathbb{E}_c\left[\phi(c, s_i)\phi(c, s_j)\right]$. This is an identity matrix, so the eigenvector equation $\frac{1}{N_s}\sum_{j=1}^{N_s} K_{i,j} v_{n,j} = \lambda_n v_{n,i}$ implies that all eigenvalues $\lambda_n = \frac{1}{N_s}$, and thus indeed tend to zero as $N_s \to \infty$. The fact that the eigenvalues approach zero reflects the fact that none of each cell's variance is shared with the rest of the population. A discussion of such representations from the perspective of machine learning can be found in Ref. [31].

Remark. The condition on continuity of K in theorem 3 can actually be relaxed [32]. However, the eigenvalue sum is then equal to the diagonal expectation of a "smoothed" version of the kernel function, that essentially removes points of discontinuity. The smoothed version of $1_{s=s'}$ will be identically zero, so again has only zero eigenvalues.

3.2 Infinite population variance

Theorem 3 describes two conditions that can lead to eigenspectra decaying slower than n^{-1}: a discontinuous kernel function, or infinite population variance. We now consider the second possibility.

As with the Pareto distribution example raised earlier, infinite population variance does not require any neuron to fire at an infinite rate. However it does imply that there must be cells responding to their preferred stimuli at arbitrarily large rates. Although such cells would be increasingly rare as the peak rate become larger, their activity is strong enough to dominate the expected variance of the population. This possibility of course could not actually happen due to physical constraints on neuronal firing, but it is still instructive to consider a counterfactual example. In this example, an infinite set of possible stimuli are represented by dedicated neuronal populations of increasing sparseness. We show that slowly decaying eigenvalues can be obtained if neurons responding to increasingly rarer stimuli do so with increasingly large firing rates. By varying the parameters of the model, we illustrate how eigenspectra decaying as n^{-1} or slower require this increase to be so rapid that population variance diverges to an infinite sum.

Example 5. We consider a discrete, infinite distribution of possible stimuli, each labeled by an positive integer. The stimuli are presented at random, with the probability that stimulus s is presented on any trial following a *zeta* *distribution*:

$$\mathbb{P}(s) = \frac{s^{-\alpha}}{\zeta(\alpha)}.$$

The zeta distribution gets its name from the normalizing factor $\zeta(\alpha) = \sum_{n=1}^{\infty} n^{-\alpha}$, which is known as the Riemann zeta function. The zeta distribution is a discrete power law: the probability of presenting stimulus s is proportional to $s^{-\alpha}$. The distribution is only defined if $\alpha > 1$. It is a "heavy tailed" distribution: while stimulus $s = 1$ is always the most likely to be presented, when $\alpha \approx 1$ the distribution of stimuli will be very wide compared to common probability distributions such as the Gaussian or Poisson, whose probabilities decay as an exponential rather than power law. The zeta distribution has infinite mean if $\alpha \leq 2$, and infinite variance if $\alpha \leq 3$; more generally, for any value of α, all moments above $\alpha - 1$ will be infinite.

We consider each stimulus to be represented by a dedicated set of cells, which respond only to that stimulus. These cell sets are therefore also labelled with positive integers. We assume that cells are also drawn according to a zeta distribution, but this time with a different parameter β. Thus, the fraction of cells in the population belonging to cell class c is:

$$\mathbb{P}(c) = \frac{c^{-\beta}}{\zeta(\beta)}.$$

Finally, the response of a cell of class c when stimulus s is presented is modeled as

$$\phi(c, s) = s^{\gamma/2}\delta_{s,c},$$

so cells of class c respond when $s = c$, but not to any other stimuli. Cells with large values of c respond only to rare stimuli, but if $\gamma > 0$, they respond especially strongly to them. The kernel function is

$$K(s, s') = \mathbb{E}_c\left[\phi(c, s)\phi(c, s')\right] = \frac{s^{\gamma-\beta}\delta_{s,s'}}{\zeta(\beta)}.$$

Its eigenfunctions are $v_n(s) = \sqrt{n^{\alpha}\zeta(\alpha)}\delta_{s,n}$, which are easily shown to be orthonormal: $\mathbb{E}_s\left[v_n(s)v_m(s)\right] = \delta_{n,m}$, with eigenvalues

$$\lambda_n = \frac{n^{\gamma-\alpha-\beta}}{\zeta(\alpha)\zeta(\beta)}.$$

The expected squared population activity is

$$\begin{aligned} \mathbb{E}_{c,s}\left[\phi(c, s)^2\right] &= \sum_{c,s} \frac{s^{-\alpha}c^{-\beta}}{\zeta(\alpha)\zeta(\beta)} s^{\gamma}\delta_{c,s} = \sum_{s} \frac{s^{\gamma-\alpha-\beta}}{\zeta(\alpha)\zeta(\beta)} \\ &= \frac{\zeta(\alpha + \beta - \gamma)}{\zeta(\alpha)\zeta(\beta)} = \sum_{n} \lambda_n, \end{aligned} \tag{10}$$

so the eigenspectrum sum equals the expected squared activity, as predicted by theorem 3. (Note that even though the stimulus space is discrete, from a mathematical perspective the requirement of continuous kernel function still holds, as the stimulus space has the *discrete topology*[22].)

The eigenvalues decay as $n^{\gamma-\alpha-\beta}$, and thus will have an infinite sum if and only if $\gamma \geq \alpha + \beta - 1$. Recall that γ determines the extent to which rarer stimuli produce stronger responses, while α and β determine how rare are these stimuli and the cells responding to them. Infinite eigenvalue sum therefore occurs only if these sparsely-firing neurons respond so strongly to their preferred stimuli that they still dominate average population activity.

3.3 Holographic coding for finite-sum eigenspectra

Examples 1 and 5 illustrate that the pathology of neural codes with infinite eigenvalue sums becomes more apparent as the population size gets larger. In example 1, we see this directly: the larger the population of independent cells, the larger the fraction of cells devoted to representing irrelevant stimulus details. With the infinite variance of example 5, the implications of large population size require more careful analysis. As with the Pareto model, the activity variance of a random sample of cells would continue to increase without limit as the sample size grows. Notwithstanding physical limitations on neural firing rates, if an actual neural code had this property then accurately gauging the population's response to stimuli would require sampling the arbitrarily small fraction of cells giving the largest responses. Such a code would be too sparse to be of any use: unless we (or a downstream brain structure) could sample the entire population, we will always miss the most important cells.

Theorem 6 formalizes this concept. It shows that in a neural code with finite population variance, the activity of a sufficiently large random sample of cells suffices to predict the activity of the entire population to arbitrary expected accuracy. We term this the "holographic" property, analogously to how a small fraction of an optical hologram can reconstruct a full image. We conclude that finite-variance codes are suited for structures such as cortex, where downstream neurons can sample only a subset of the population's activity. Conversely, codes not satisfying the n^{-1} bound are pathological in that they require downstream structures to sample every neuron to read out the population code.Examples 4 and 5 illustrate how this fails for codes whose eigenspectra decay too slowly: in the first case, where the eigenspectrum is flat, all neurons are independent so no one can be predicted from any other; in the second case, however many neurons one samples from, there will always be more neurons responding arbitrarily strongly to rare stimuli that did not drive any cells in the training sample.

Theorem 6. *Let $\phi(c, n)$ be a continuous code with finite eigenvalue sum, let $\{c_i\}$ be a sequence of cells drawn independently at random from the cell population $\mathcal{C}$, and let $\boldsymbol{\phi}_n(s)$ be the vector containing the responses of the first n cells to stimulus s. With probability 1, there exists for each possible target cell c a sequence of weight vectors $\mathbf{w}_n(c)$ such that*

$$\lim_{n\to\infty} \mathbb{E}_{c,s}\left[|\mathbf{w}_n(c)\cdot\boldsymbol{\phi}_n(s) - \phi(c,s)|^2\right] = 0.$$

Conversely, if the eigenvalue sum is infinite, no such sequence $\mathbf{w}_n(c)$ exists.

Proof. First note that with probability 1 over c, $\phi(c,\cdot) \in L^2(\mathcal{S})$. Indeed, if this were not the case then $\mathbb{E}_{c,s}\left[\phi(c,s)^2\right]$ would be infinite, contradicting finite eigenvalue sum.

Now for each cell c, define the weight vector $\mathbf{w}_n(c)$ by standard multivariate linear regression, as the weights minimizing $\mathbb{E}_s\left[|\hat{\phi}_n(c,s) - \phi(c,s)|^2\right]$, where $\hat{\phi}_n(c,s) = \mathbf{w}_n(c)\cdot\boldsymbol{\phi}_n(s)$. As usual for linear regression, the prediction $\hat{\phi}(c,s)$ is a projection of the target variable onto the space spanned by the predictor variables: $\hat{\phi}_n(c,\cdot) = \mathscr{P}_n\phi(c,\cdot)$, where $\mathscr{P}_n$ represents the projection operator onto the subspace of $L^2(\mathcal{S})$ spanned by the functions $\phi(c_1,\cdot),\ldots,\phi(c_n,\cdot)$.

The expected error over all possible target cells and stimuli is thus

$$\begin{aligned}\mathbb{E}_{c,s}\left[|\hat{\phi}_n(c,s) - \phi(c,s)|^2\right] &= \mathbb{E}_c\left[\|\phi(c,\cdot) - \mathscr{P}_n\phi(c,\cdot)\|^2\right]\\ &= \mathbb{E}_c\left[\|(\mathscr{I} - \mathscr{P}_n)\phi(c,\cdot)\|^2\right]\\ &= \mathrm{Tr}\big(\mathscr{K}(\mathscr{I} - \mathscr{P}_n)\big), \qquad (11)\end{aligned}$$

where $\mathscr{I}$ represents the identity operator, $\mathscr{K}$ the kernel operator, and $\|\cdot\|$ the Hilbert space norm of $L^2(\mathcal{S})$. Now, define the sample kernel function $K_n(s,s') = \frac{1}{n}\sum_{i=1}^n \phi(c_i,s)\phi(c_i,s')$, and let $\mathscr{K}_n$ be the corresponding integral operator. $\mathscr{K}_n$ has rank n, and projection onto the space spanned by its non-zero eigenfunctions is equivalent to application of the projection operator $\mathscr{P}_n$. By theorems 8.1.2 and 5.1.4 of Ref. [4], we can therefore conclude that the operator $\mathscr{I} - \mathscr{P}_n$ converges to zero on the range of $\mathscr{K}$, but only in a particular sense. It does not converge in *operator norm*, which would mean that for any $\epsilon > 0$, there exists an N such that $\|(\mathscr{I} - \mathscr{P}_n)x\| < \epsilon\|x\|$ for all $n \geq N$ and $x \in \text{range}(\mathscr{K})$; indeed, $\mathscr{P}_n$ is a finite rank operator, so if $\mathscr{K}$ is infinite rank there is always an $x \in \text{range}(\mathscr{K})$ with $(\mathscr{I} - \mathscr{P}_n)x = x$. Instead, for any m and ϵ, there exists an N such that $\|(\mathscr{I} - \mathscr{P}_n)x\| < \epsilon\|x\|$ for all $n \geq N$ and for all x in the space spanned by $v_1,\ldots,v_m$, the first m eigenfunctions of $\mathscr{K}$.

If the eigenvalues of $\mathscr{K}$ have finite sum, this mode of convergence is sufficient for the expected prediction error (11) to converge to zero. Indeed, to obtain error less than ϵ, choose m so that $\sum_{i>m}\lambda_i \leq \epsilon/2$, and choose N so that

$\|(\mathscr{I} - \mathscr{P}_n)x\| \leq \delta \, \|x\|$ for all $n \geq N$ and all x in the space spanned by $v_1, \ldots, v_m$, where $\delta = \frac{\epsilon}{2\sum_{i \leq m} \lambda_i}$. As the v_i form an orthonormal basis for $L^2(\mathcal{S})$,

$$\begin{aligned}
\mathrm{Tr}\big((\mathscr{I} - \mathscr{P}_n)\mathscr{K}\big) &= \sum_{i=1}^{\infty} \langle v_i, (\mathscr{I} - \mathscr{P}_n)\mathscr{K} v_i \rangle \\
&= \sum_{i \leq m} \lambda_i \langle v_i, (\mathscr{I} - \mathscr{P}_n) v_i \rangle + \sum_{i > m} \lambda_i \langle v_i, (\mathscr{I} - \mathscr{P}_n) v_i \rangle \\
&\leq \sum_{i \leq m} \lambda_i \delta + \sum_{i > m} \lambda_i \\
&\leq \epsilon/2 + \epsilon/2 = \epsilon.
\end{aligned}$$

Conversely, if the eigenvalues of $\mathscr{K}$ have infinite sum, then because $\mathscr{P}_n$ is a projection of rank n, the Courant-Fischer theorem (Ref. [4], theorem 4.2.7) tells us that

$$\mathrm{Tr}\big((\mathscr{I} - \mathscr{P}_n)\mathscr{K}\big) \geq \sum_{i > n} \lambda_i = \infty.$$

Therefore, however many neurons n we sample from, our expected prediction error is unlimited. □

3.4 Pathologies of infinite derivative codes

Similar arguments to those of sections 3.2 and 3.3 illustrate the pathologies of codes with eigenspectra decaying slower than $n^{-1-2/d}$, which have infinite expected deriviate magnitude (theorem 5). To have an infinite population derivative does not require the tuning curve of any individual neuron to be nonsmooth. Indeed, as we saw in example 3, it is possible for each neuron's stimulus responses to be differentiable, but the population code not to be. In this scenario, the population contains neurons of ever sharper tuning, so though the total derivative is finite for any finite sample of neurons, it continues to grow without bound as the sample size increases. Fractal population codes are therefore further pathological for reasons analogous to theorem 6, and as illustrated in example 3: the representation of the derivative is dominated by increasingly rare neurons of arbitrarily sharp tuning. Reliably reading out the difference in the representations of two similar stimuli therefore requires sampling essentially the entire population, or these rare cells will be missed.

To see why non-differentiability would be pathological for read-out of a neural code, consider the problem of trying to distinguish two nearby stimuli s and s' which differ by an amount ds. The difference in their firing patterns will be approximately $\frac{\partial \phi(c,s)}{\partial s} ds$, and for a downstream brain to distinguish the two stimuli, it must be able to specifically weight those cells for which $\frac{\partial \phi(c,s)}{\partial s}$ is large. However, if $\mathbb{E}_c \left[\left| \frac{\partial \phi(c,s)}{\partial s_j} \right|^2 \right] = \infty$, the neural representation of the derivative is dominated by a sparse set of cells of arbitrarily high derivative: for any D, there are cells for which $\left| \frac{\partial \phi(c,s)}{\partial s_j} \right|^2 > D$, and these cells dominate the total activity $\mathbb{E}_c \left[\left| \frac{\partial \phi(c,s)}{\partial s_j} \right|^2 \right]$. The difference in population responses to s and s' gets smaller the closer s' is to s, however this difference becomes concentrated in an ever sparser subset of cells. Such a code is pathological, for the same reason as for infinite variance codes: a downstream structure will not be able to accurately respond to the difference between s' and s without sampling all the cells in the population.

3.5 Implications for coding in finite populations

The discussion so far has focused on the limit of how the eigenspectrum λ_n scales as $n \to \infty$. However, we can only present a finite number of stimuli, and the brain only contains a finite number of neurons. This section focuses on interpretation of these results for coding by finite populations.

Our experimental results (Figs. 2g-j, Extended Data Fig. 10, Extended Data Fig. 8) show that the power law holds accurately over a progressively larger range the more stimuli are analyzed and the more neurons are recorded; typically this range extends to approximately half the total number of stimuli presented. Theorem 2 suggests that the the reason eigenvalues fall below the power law at this point is because the remaining eigenvectors correspond to or are corrupted by additive noise, and thus will have cross-validated eigenvalues approaching zero.

Because the number of stimuli we presented was determined by the (scientifically arbitrary) maximum time we could perform our experiments, it is reasonable to infer that the powerlaw would hold over a larger range if more stimuli had been presented. Eventually, however, the powerlaw must stop: the mouse's brain contains only a finite number of neurons, and the number of stimuli presentable on the monitor is finite. We also cannot rule out the possibility that, if we were able to present sufficiently many stimuli, we would observe the eigenspectrum would at some point reliably deviate from a power law, in a manner that did not simply result from measurement noise.

The question of how to interpret a scaling law that holds over a finite range is familiar from physics. Scaling laws abound in statistical mechanics, but almost always over a finite range due to the particulate nature of matter. Particularly relevant to the present case is the example of phase transitions. Phase transitions – abrupt changes of state occurring when a system parameter crosses a particular value – only theoretically occur in infinite systems [33]. Finite systems show continuous dependence on the parameter value, but this dependence grows sharper the larger the system is. Thus, for finite but large systems, the dependence on the parameter is smooth but so rapid as to appear discontinuous to close approximation.

The present case is analogous. Consider an eigenspectrum that decayed as $\lambda_n \propto n^{-\alpha}$ over a finite range $n = 1$ to N. The total population variance is always finite, and an integral approximation gives

$$\sum_{n=1}^{N} n^{-\alpha} \approx \int_1^N n^{-\alpha} dn = \frac{N^{1-\alpha} - 1}{1 - \alpha}.$$

This is a continuous function of α, for any finite value of N (including $\alpha = 1$, where Taylor expansion yields $\log(N)$). The actual value of N is unimportant if $\alpha > 1$, as the sum will always be close to $1/(\alpha - 1)$. However, we observe a rapid increase in total variance when α becomes close to 1, and this increase will be more rapid the larger N is. In the limit $N \to \infty$, we obtain a discontinuous "phase transition" to infinite variance when $\alpha \geq 1$.

A neural code whose eigenspectrum decayed more slowly than n^{-1} over a large but finite range would still be pathological, for the same reasons as described in theorems 3 and 6 and examples 1 and 5. The great majority of population variance would come from the tail eigenfunctions, devoted to encoding tiny details of stimuli; this failure would not occur discontinuously at $\alpha \geq 1$, but its dependence on α would grow sharper the larger N is. Analogous arguments show that for a code whose eigenspectrum decayed more slowly than $n^{-1-2/d}$, the great majority of the derivative's variance would come from the tail eigenfunctions, resulting in a poor code for the reasons given in the discussion of theorems 4 and 5, section 3.4, and examples 2 and 3.

Finally, we note that these arguments apply only when the number of neurons in the population is substantially greater than the manifold dimension of the stimulus set. As described in example 1, encoding a stimulus of manifold dimension d to an accuracy of 1 in 1024 will take $10d$ neurons. If $d \ll N$ then nearly all the population variance is devoted to encoding irrelevant stimulus details; but if $d \approx N$ this need not be the case. We therefore hypothesize that the eigenspectrum bounds we observed in visual cortex are likely to hold in any brain structure where the number of neurons substantially exceeds the manifold dimension of the stimulus being encoded.

3.6 cvPCA in the limit $N_c \to \infty$

We can use the random-sampling framework to analyze the cvPCA algorithm in the limit that the number of both cells and stimuli grows large. We will extend theorems 1 and 2 to this limit, showing that the eigenspectrum estimated from a finite sample converges to the eigenspectrum of the continuous kernel function; and that unbiased estimation of the population eigenspectrum does not require all cells to have equal uncorrelated noise variance in this limit.

In the $N_c \to \infty$ limit, we can define both a kernel function $K(s, s') = \mathbb{E}_c [\phi(c, s)\phi(c, s')]$, where s and s' range over the entire population of possible stimuli, and a correlation function $G(c, c') = \mathbb{E}_s [\phi(c, s)\phi(c', s)]$ where c and c' range over the entire population of potential cells. The kernel function and correlation function have the same eigenvalues: if the kernel eigenfunctions satisfy $\mathbb{E}_{s'} [K(s, s')v_n(s')] = \lambda_n v_n(s)$, then correlation eigenfunctions $u_n(c)$ are related to the kernel eigenfunctions by $u_n(c) = \lambda_n^{1/2} \mathbb{E}_s [\phi(c, s)v(s)]$, and have the same eigenvalues. Thus $\mathbb{E}_{c'} [G(c, c')u_n(c')] = \lambda_n u_n(c)$, and $\mathbb{E}_c [u_n(c)u_m(c)] = \delta_{n,m}$. The fundamental result underpinning the current study is that the eigenspectrum of a finite kernel matrix, defined by a random sample of stimuli and neurons, converges to the eigenspectrum of the full kernel function as the size of the sample increases. This convergence is guaranteed to occur if population variance is finite, $\mathbb{E}_{c,s} \left[\phi(c, s)^2\right] < \infty$, and can also happen under less restrictive conditions [34]. Thus, by observing how a sufficiently large number of neurons respond to a sufficiently large number of stimuli, we can make inferences about the entire population code, defined by the responses of any neurons we might have recorded in that mouse, to any natural image stimuli we might have presented with similar properties.

The generalization of theorem 1 to an infinite population closely mirrors the original:

Theorem 7. *Let $\{s_i : i \in \mathbb{N}\}$ be an infinite sequence of cells drawn from $\mathbb{P}(\mathcal{S})$, and $\{c_j : j \in \mathbb{N}\}$ be an infinite sequence of cells drawn from $\mathbb{P}(\mathcal{C})$. Let $\hat{\lambda}_n$ denote the n^{th} eigenvalue estimated by cvPCA from a sample of the first N_s stimuli and N_c cells in this sequence, and let λ_n represent the n^{th} eigenvalue of the full kernel function. Then with probability 1,*

$$\lim_{N_s, N_c \to \infty} \mathbb{E}_{\nu_1,\nu_2} \left[\sum_{n=1}^{N} \hat{\lambda}_n\right] \leq \sum_{n=1}^{N} \lambda_n. \tag{12}$$

Proof. To show convergence of the estimated eigenspectrum as the number of cells and stimuli grows larger, we will make use of versions of the *strong law of large numbers* (SLLN). The simplest version of the SLLN concerns a series independent random variables X_i drawn from a common distribution, and says that with probability 1 the sample means $\frac{1}{N}\sum_{i \leq N} X_i$ converge to $\mathbb{E}[X]$ as $N \to \infty$, provided $\mathbb{E}[|X|]$ is finite. Note that the SLLN applies to single draws of the sequence X_i: it says that there is no chance that the sample means will not converge, for any single draw. Here, we will apply this philosophy to the stimuli and neurons drawn from the population distributions, showing that with probability 1 we expect convergence as we analyze larger sets of stimuli and neurons.

Specifically, we will make use of two specialized forms of SLLN. The first (Ref. [4], theorem 8.1.2) concerns covariance functions, and says if $\mathbb{E}_{c,s} \left[f(c, s)^2\right] < \infty$, then with probability 1, $\lim_{N_s \to \infty} \frac{1}{N_s} \sum_{i \leq N_s} \mathrm{f}_1(s_i, c)\mathrm{f}_1(s_i, c')$ converges to the total covariance function $\tilde{G}(c, c')$ in Hilbert-Schmidt norm, i.e. $\mathbb{E}_{c,c'} \left[\left(\sum_{i \leq N_s} \mathrm{f}_1(s_i, c)\mathrm{f}_1(s_i, c') - \tilde{G}(c, c')\right)^2\right] \to 0$. This is sufficient for its eigenfunctions to converge to the eigenfunctions $\tilde{u}_n(c)$ of the covariance function $\tilde{G}(c, c')$.

The second form of SLLN concerns the eigenvectors of the sample correlation matrices[35]. It says that under the same finite-variance assumption, the eigenvectors of the sample covariance matrix will converge with probability 1 to the population eigenfunctions, in the mean-square sense. Putting the two results together, we can conclude that

$$\lim_{N_c, N_s \to \infty} \frac{1}{N_c} \sum_{i \leq N_c} (\tilde{\mathbf{u}}_{n,i} - \tilde{u}_n(c_i))^2 = 0. \tag{13}$$

Now let us consider the n^{th} cross-validated eigenvalue, $\hat{\lambda}_n = (\mathbf{F}_1\tilde{\mathbf{u}}_n) \cdot (\mathbf{F}_2\tilde{\mathbf{u}}_n)/N_s N_c^2$. Taking expectations over the noise on both repeats, we obtain

$$\mathbb{E}_{\nu_1,\nu_2} \left[\hat{\lambda}_n\right] = \frac{1}{N_s N_c^2} \sum_{k \leq N_s} \mathbb{E}_{\nu_1,\nu_2} \left[\Big(\sum_{i \leq N_c} \big(\phi(c_i, s_k) + \nu_1(c_i, s_k)\big)\tilde{\mathbf{u}}_{n,i}\Big)\Big(\sum_{j \leq N_c} \big(\phi(c_j, s_k) + \nu_2(c_j, s_k)\big)\tilde{\mathbf{u}}_{n,j}\Big)\right]$$

Now, because ν_2 is independent of ν_1 and $\tilde{\mathbf{u}}_n$, it contributes nothing to the expectation. Thus,

$$\mathbb{E}_{\nu_1,\nu_2} \left[\hat{\lambda}_n\right] = \frac{1}{N_s N_c^2} \sum_{k \leq N_s, i \leq N_c, j \leq N_c} \mathbb{E}_{\nu_1} \left[\phi(c_i, s_k)\phi(c_j, s_k)\tilde{\mathbf{u}}_{n,i}\tilde{\mathbf{u}}_{n,j} + \nu_1(c_i, s_k)\phi(c_j, s_k)\tilde{\mathbf{u}}_{n,i}\tilde{\mathbf{u}}_{n,j}\right]$$

Because the sample eigenvector $\tilde{\mathbf{u}}_n$ is computed from repeat 1, it is correlated with ν_1, so the expectation of the second term is not zero. Nevertheless, the size of this expectation converges to zero for large enough numbers of cells and stimuli. Indeed, equation (13) shows that $\tilde{\mathbf{u}}_{n,i}$ approaches the population eigenfunction $\tilde{u}_n(i)$ in mean-square, and $\tilde{u}_n(i)$ is independent of ν_1. Applying the Cauchy-Schwartz inequality and the fact that ν_1 has finite variance, we see that $\frac{1}{N_c}\sum_{i\leq N_c}\nu_1(c_i,s_k)(\tilde{\mathbf{u}}_{n,i}-\tilde{u}_n(i))$ tends to zero, so we may replace $\tilde{\mathbf{u}}_{n,i}$ by its limit, obtaining

$$\begin{aligned}\lim_{N_c,N_s\to\infty}\mathbb{E}_{\nu_1,\nu_2}\left[\hat{\lambda}_n\right] &= \lim_{N_c,N_s\to\infty}\frac{1}{N_sN_c^2}\sum_{k\leq N_s,i\leq N_c,j\leq N_c}\phi(c_i,s_k)\phi(c_j,s_k)\tilde{u}_n(c_i)\tilde{u}_n(c_j)\\ &= \mathbb{E}_{s,c,c'}\left[\phi(c,s)\phi(c',s)\tilde{u}_n(c)\tilde{u}_n(c')\right]\\ &= \mathbb{E}_{c,c'}\left[G(c,c')\tilde{u}_n(c)\tilde{u}_n(c')\right] \end{aligned}\tag{14}$$

If the functions $\tilde{u}_n$ were eigenfunctions of the signal covariance operator $K(c,c')$, then this last term would equal the signal population eigenvalues λ_n. However, $\tilde{u}_n$ are the eigenfunctions of the total covariance operator $\tilde{K}(c,c')$. As we discuss below in Theorem 8, these are likely to coincide for situations such as the current recording. However, even if they do not coincide, we may obtain a one-sided bound. Indeed, the Courant-Fisher minimax principle (Ref. [4], theorem 4.2.7) implies that for any set of orthonormal functions $\tilde{u}_n(c)$, this expectation cannot exceed what would be obtained with the true eigenfunctions: $\sum_{n\leq N}\mathbb{E}_{c,c'}\left[K(c,c')\tilde{u}_n(c)\tilde{u}_n(c')\right]\leq\sum_{n\leq N}\lambda_n$. □

The generalization of theorem 2 closely mirrors the original, but without the requirement for identical noise variances.

Theorem 8. *Consider a noise model*

$$\nu_r(c,s)=\alpha_r(s)\phi(c,s)+\beta_r(c,s)+\gamma_r(c,s)$$

where $\alpha_r(s)$ represents multiplicative noise scaling the entire population's response to stimulus s on repeat r; $\beta_r(c,s)$ is additive noise in dimensions orthogonal to the stimulus; and $\gamma_r(c,s)$ is independent between neurons and stimuli. Assume that with α, β and γ statistically independent of each other and of $\phi(c,s)$, but let the variance of γ vary between cells. Then the eigenvalue estimates $\mathbb{E}_{\nu_1,\nu_2}\left[\hat{\lambda}_n\right]$ converge to the population eigenvalues λ_n together with additional zero eigenvalues corresponding to the additive noise dimensions.

Proof. From equation (14), we see that if an eigenfunction $\tilde{u}_n(c)$ of the total covariance function $\tilde{G}(c,c')$ is equal to an eigenfunction $u_m(c)$ of the signal covariance function $G(c,c')$, then $\lim_{N_s,N_c\to\infty}\mathbb{E}_{\nu_1,\nu_2}\left[\hat{\lambda}_n\right]=\lambda_m$. It therefore suffices to show that the eigenfunctions of $\tilde{G}(c,c')$ are the same as those of $G(c,c')$, together with additional orthogonal dimensions. To do so, note that the total response

$$f_r(c,s)=\left(1+\alpha_r(s)\right)\phi(c,s)+\beta_r(c,s)+\gamma_r(c,s)$$

Thus,

$$\begin{aligned}\tilde{G}(c,c') &= \mathbb{E}_{s,\alpha,\beta,\gamma}\left[f_r(c,s)f_r(c',s)\right]\\ &= \mathbb{E}_{s,\alpha}\left[(1+\alpha_r(s))^2\phi(c,s)\phi(c',s)\right]+\mathbb{E}_{s,\beta}\left[\beta_r(c,s)\beta_r(c',s)\right]+\mathbb{E}_{s,\gamma}\left[\gamma_r(c,s)\gamma_r(c',s)\right]\\ &= V_\alpha G(c,c')+B(c,c')+V_\gamma(c)1_{c=c'}.\end{aligned}$$

Here, $V_\alpha=\mathbb{E}_{\alpha,s}\left[(1+\alpha_r(s))^2\right]$ is a scale factor resulting from multiplicative modulation, $B(c,c')$ represents the correlation of the additive noise, $V_\gamma(c)=\mathbb{E}_{s,\gamma}\left[\gamma_r(c,s)^2\right]$ is the independent noise variance, and $1_{c=c'}$ represents the function that is 1 if $c=c'$, zero otherwise. Note that $\mathbb{E}_{c'}\left[B(c,c')u(c')\right]=0$ as we have assumed the additive noise dimensions orthogonal to the signal dimensions, and $\mathbb{E}_{c'}\left[1_{c=c'}u(c')\right]=0$ since $\mathbb{P}(c=c')=0$. Thus, the signal eigenfunctions $u_n(c)$ are eigenfunctions of $\tilde{K}$:

$$\mathbb{E}_{c'}\left[\tilde{K}(c,c')u_n(c')\right]=V_\alpha\lambda_nu_n(c)$$

Although their eigenvalues for function $\tilde{G}$ have been inflated by multiplicative noise, the eigenvalues estimated by cross-validated PCA are those of G as the noise has mean 0. Thus, $\mathbb{E}\left[\hat{\lambda}_n\right]=\lambda_n$. In addition to these eigenfunctions,

$\tilde{G}$ has a new set of eigenfunctions $b_m(c)$ corresponding to the eigenfunctions of the additive noise covariance B. These are orthogonal to the directions of signal covariance, and thus contribute estimated eigenvalues of 0. Thus, under the assumed noise model, cvPCA provides an asymptotically unbiased estimate of population eigenvalues, together with an additional set of zero eigenvalues resulting from orthogonal additive noise. □

References

[1] Harris, K. D. Neural signatures of cell assembly organization. *Nature Reviews Neuroscience* **6**, 399 (2005).

[2] Stringer, C. *et al.* Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **364**, 255–255 (2019). URL https://science.sciencemag.org/content/364/6437/255. https://science.sciencemag.org/content/364/6437/255.full.pdf.

[3] Musall, S., Kaufman, M. T., Gluf, S. & Churchland, A. Movement-related activity dominates cortex during sensory-guided decision making. *bioRxiv* 308288 (2018).

[4] Hsing, T. & Eubank, R. *Theoretical foundations of functional data analysis, with an introduction to linear operators* (John Wiley & Sons, 2015).

[5] Goris, R. L. T., Movshon, J. A. & Simoncelli, E. P. Partitioning neuronal variability. *Nature Neuroscience* **17**, 858–65 (2014).

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

[7] Mandelbrot, B. How long is the coast of britain? statistical self-similarity and fractional dimension. *Science* **156**, 636–638 (1967).

[8] Mattila, P. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability* (Cambridge university press, 1999).

[9] Falconer, K. *Fractal geometry: mathematical foundations and applications* (John Wiley & Sons, 2004).

[10] Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation* **15**, 1373–1396 (2003).

[11] DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).

[12] Anselmi, F. & Poggio, T. Representation learning in sensory cortex: a theory. Tech. Rep., Center for Brains, Minds and Machines (CBMM) (2014).

[13] Hertz, J., Krogh, A. & Palmer, R. G. *Introduction to the theory of neural computation.* (Addison-Wesley/Addison Wesley Longman, 1991).

[14] Poggio, T. & Smale, S. The mathematics of learning: Dealing with data. *Notices of the AMS* **50**, 537–544 (2003).

[15] Poggio, T. & Girosi, F. Networks for approximation and learning. *Proceedings of the IEEE* **78**, 1481–1497 (1990).

[16] Saxe, A. M., McClelland, J. L. & Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120* (2013).

[17] Haykin, S. S. *et al.* *Neural networks and learning machines* (New York: Prentice Hall,, 2009).

[18] Liu, W., Principe, J. C. & Haykin, S. *Kernel adaptive filtering: a comprehensive introduction* (John Wiley & Sons, 2011).

[19] Scholkopf, B. & Smola, A. J. *Learning with kernels: support vector machines, regularization, optimization, and beyond* (MIT press, 2001).

[20] Steinwart, I. & Christmann, A. *Support vector machines* (Springer Science & Business Media, 2008).

[21] Rasmussen, C. E. & Williams, C. K. *Gaussian process for machine learning* (MIT press, 2006).

[22] Tao, T. *An Epsilon of Room, I: Real Analysis* (American Mathematical Society, Providence, RI, 2010).

[23] Tao, T. *An introduction to measure theory* (American Mathematical Society Providence, RI, 2011).

[24] Bérard, P. H. *Spectral geometry: direct and inverse problems* (Springer, 2006).

[25] Cohn, P. M. *Algebra* (J. Wiley,, 1980).

[26] Reed, M. & Simon, B. Methods of modern mathematical physics, vol. III: Scattering theory. *New York, San Francisoco, London* (1979).

[27] Weierstraß, K. Über continuirliche functionen eines reellen arguments, die für keinen werth des letzteren einen bestimmten differentialquotienten besitzen. In *Ausgewählte Kapitel aus der Funktionenlehre*, 190–193 (Springer, 1988).

[28] Wood, D. The computation of polylogarithms. *University of Kent Technical Report* (1992).

[29] Ramsay, J. O. & Silverman, B. W. *Applied functional data analysis: methods and case studies* (Springer, 2007).

[30] Pareto, V. *Cours d'économie politique: professé à l'Université de Lausanne*, vol. 1 (F. Rouge, 1896).

[31] Steinwart, I. & Scovel, C. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constructive Approximation* **35**, 363–417 (2012).

[32] Brislawn, C. Traceable integral kernels on countably generated measure spaces. *Pacific Journal of Mathematics* **150**, 229–240 (1991).

[33] Mainwood, P. Phase transitions in finite systems. *PhilSci Archive* (2005).

[34] Koltchinskii, V., Giné, E. *et al.* Random matrix approximation of spectra of integral operators. *Bernoulli* **6**, 113–167 (2000).

[35] Koltchinskii, V. I. Asymptotics of spectral projections of some random matrices approximating integral operators. In *High dimensional probability*, 191–227 (Springer, 1998).