Population codes enable learning from few examples by shaping inductive bias

Blake Bordelon[a,b] and Cengiz Pehlevan[a,b]

[a]John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA
[b]Center for Brain Science, Harvard University, Cambridge, MA, USA

March 13, 2022

Abstract

Learning from a limited number of experiences requires suitable inductive biases. To identify how inductive biases are implemented in and shaped by neural codes, we analyze sample-efficient learning of arbitrary stimulus-response maps from arbitrary neural codes with biologically-plausible readouts. We develop an analytical theory that predicts the generalization error of the readout as a function of the number of observed examples. Our theory illustrates in a mathematically precise way how the structure of population codes shapes inductive bias, and how a match between the code and the task is crucial for sample-efficient learning. We observe that many different codes can support the same inductive bias. By analyzing recordings from the mouse primary visual cortex, we demonstrate that biological codes have lower total activity than other codes with identical bias. Using these mouse primary visual cortex responses, we demonstrate the existence of an efficiency bias towards low frequency orientation discrimination tasks for grating stimuli and low spatial frequency reconstruction tasks for natural images. We reproduce the discrimination bias in a simple model of primary visual cortex, and further show how invariances in the code to certain stimulus variations alter learning performance. We extend our methods to time-dependent neural codes and predict the sample efficiency of readouts from recurrent networks. Finally, we discuss implications of our theory in the context of recent developments in neuroscience and artificial intelligence. Overall, our study provides a concrete method for elucidating inductive biases of the brain and promotes sample-efficient learning as a general normative coding principle.

1 Introduction

The ability to learn quickly is crucial for survival in a complex and an everchanging environment, and the brain effectively supports this capability. Often, only a few experiences are sufficient to learn a task, whether acquiring a new word [1] or recognizing a new face [2]. Despite the importance and ubiquity of sample efficient learning, our understanding of the brain's information encoding strategies that support this faculty remains poor [3, 4, 5].

In particular, when learning and generalizing from past experiences, and especially from few experiences, the brain relies on implicit assumptions it carries about the world, or its inductive biases [6, 5]. Reliance on inductive bias is not a choice: inferring a general rule from finite observations is an ill-posed problem which requires prior assumptions since many hypotheses can explain the same observed experiences [7]. Consider learning a rule that maps photoreceptor responses to a prediction of whether an observed object is a threat or is neutral. Given a limited number of visual experiences of objects and their threat status, many threat-detection rules are consistent with these experiences. By choosing one of these threat-detection rules, the nervous system reveals an inductive bias. Without the right biases that suit the task at hand, successful generalization is impossible [6, 5]. In order to understand why we can quickly learn to perform certain tasks accurately but not others, we must understand the brain's inductive biases [3, 4, 5].

In this paper, we study sample efficient learning and inductive biases in a general neural circuit model which comprises of a population of sensory neurons and a readout neuron learning a stimulus-response map with a biologically-plausible learning rule (Fig 1A). For this circuit and learning rule, inductive bias arises from the nature of the neural code for sensory stimuli, specifically its similarity structure. While different population codes can encode the same stimulus variables and allow learning of the same output with perfect performance given infinitely many samples, learning performance can depend dramatically on the code when restricted to a small number of samples, where the reliance on and the effect of inductive bias are strong (Fig 1B,C,D). Given the same sensory examples and their associated response values, the readout neuron may make drastically different predictions depending on the inductive bias set by the nature of the code, leading to successful or failing generalizations (Fig 1C,D). We say that a code and a learning rule, together, have a good inductive bias for a task if the task can be learned from a small number of examples.

In order to understand how population codes shape inductive bias and allow fast learning of certain tasks over others with a biologically plausible learning rule, we develop an analytical theory of the readout neuron's learning performance as a function of the number of sampled examples, or sample size. We find that the readout's performance is completely determined by the code's kernel, a function which takes in pairs of population response vectors and outputs a representational similarity defined by the inner product of these vectors. We demonstrate that the spectral properties of the kernel introduce an inductive bias toward explaining sampled data with simple stimulus-response maps and determine compatibility of the population code with learning task, and hence the sample-efficiency of learning. We observe that many codes could support the same kernel function, however, by analyzing data from mouse primary visual cortex (V1) [8, 9, 10, 11], we find

that the biological code is metabolically more efficient than others. Further, mouse V1 responses support sample-efficient learning of low frequency orientation discrimination and low spatial frequency reconstruction tasks over high frequency ones. We demonstrate the discrimination bias in a simple model of V1 and show how response nonlinearity, sparsity, and relative proportion of simple and complex cells influence the code's bias and performance on learning tasks, including ones that involve invariances. Finally, we extend our theory to temporal population codes. As an example, we study codes generated by recurrent neural networks learning a delayed response task.

Overall, our results demonstrate that for a fixed learning rule, the neural sensory representation imposes an inductive bias over the space of learning tasks, allowing some tasks to be learned by a downstream neuron more sample-efficiently than others. Our work provides a concrete method for elucidating inductive biases of populations of neurons and suggest sample-efficient learning as a novel functional role for population codes.

2 Results

We consider a population of N neurons whose responses, $\{r_1(\boldsymbol{\theta}), r_2(\boldsymbol{\theta}), ..., r_N(\boldsymbol{\theta})\}$, vary with the input stimuli, which is parameterized by a vector variable $\boldsymbol{\theta}$, such as the orientation and the phase of a grating (Figure 1A). These responses define the population code. A readout neuron learns its weights $\mathbf{w}$ to approximate a stimulus-response map, or a target function $y(\boldsymbol{\theta})$, such as one that classifies stimuli as apetitive ($y = 1$) or aversive ($y = -1$), or a more smooth one that attaches intermediate values of valence. We emphasize that in our model only the readout neuron performs learning, and the population code is assumed to be static. Our theory is general in its assumptions about the structure of the population code and the stimulus-response map considered (Methods 4.2), and can apply to many scenarios.

The readout neuron learns from P stimulus-response examples with the goal of generalizing to previously unseen ones. Example stimuli $\boldsymbol{\theta}^\mu$, ($\mu = 1, \ldots, P$) are sampled from a probability distribution describing stimulus statistics $p(\boldsymbol{\theta})$. This distribution can be natural or artificially created, for example, for a laboratory experiment (Appendix, App. 1). From the set of learning examples, $\mathcal{D} = \{\boldsymbol{\theta}^\mu, y(\boldsymbol{\theta}^\mu)\}_{\mu=1}^P$, the readout weights are learned with the local, biologically-plausible delta-rule, $\Delta w_j = \eta \sum_\mu r_j(\boldsymbol{\theta}^\mu)(y(\boldsymbol{\theta}^\mu) - \mathbf{r}(\boldsymbol{\theta}^\mu) \cdot \mathbf{w})$, where η is a learning rate (Figure 1A). This learning process, when weights are initialized as $\mathbf{w}_0 = 0$, converges to a unique set of weights $\mathbf{w}^*(\mathcal{D})$ (App. 3A). Generalization error with these weights is given by

$$E_g(\mathcal{D}) = \int p(\boldsymbol{\theta})\ (\mathbf{w}^*(\mathcal{D}) \cdot \mathbf{r}(\boldsymbol{\theta}) - y(\boldsymbol{\theta}))^2 d\boldsymbol{\theta}, \tag{1}$$

which quantifies the expected error of the trained readout over the entire stimulus distribution $p(\boldsymbol{\theta})$. This quantity will depend on the population code $\mathbf{r}(\boldsymbol{\theta})$, the target function $y(\boldsymbol{\theta})$ and the set of training examples $\mathcal{D}$. Our theoretical analysis of this model provides insights into how populations of neurons encode information and allow sample-efficient learning.

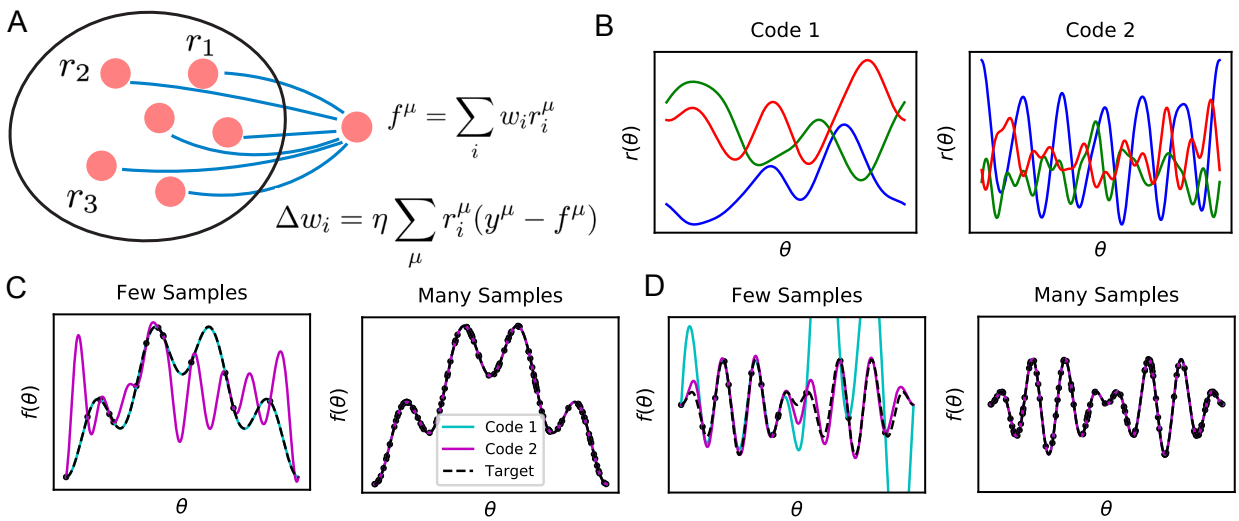

Figure 1: Learning tasks through linear readouts exploit representations of the population code to approximate a target response. **A** The readout weights from the population to a downstream neuron, shown in blue, are updated to fit target values y, using the local, biologically plausible delta rule. **B** Examples of tuning curves for two different population codes: Smooth tuning curves (Code 1) and rapidly varying tuning curves (Code 2). **C** (Left) A target function with low frequency content is approximated through the learning rule shown in **A** using these two codes. The readout from Code 1 (turquoise) fits the target function (black) almost perfectly with only $P = 12$ training examples, while readout from Code 2 (purple) does not accurately approximate the target function. (Right) However, when the number of training examples is sufficiently large ($P = 120$), the target function is estimated perfectly by both codes, indicating that both codes are equally expressive. **D** The same experiment is performed on a task with higher frequency content. (Left) Code 1 fails to perform well with $P = 12$ samples indicating mismatch between inductive bias and the task can prevent sample efficient learning while Code 2 accurately fits the target. (Right) Again, provided enough data $P = 120$, both models can accurately estimate the target function. Details of these simulations are given in Methods 4.1.

Kernel structure of population codes controls learning performance

First, we note that the generalization performance of the learned readout on a given task depends entirely on the inner product kernel, defined by $K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{N}\sum_{i=1}^{N} r_i(\boldsymbol{\theta}) r_i(\boldsymbol{\theta}')$, which quantifies the similarity of population responses to two different stimuli $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. This is because the learning procedure converges to a unique solution $\mathbf{w}^*(\mathcal{D})$ for the training set $\mathcal{D}$ [12, 13] and the readout neuron's learned output has the form

$$f(\boldsymbol{\theta}) = \mathbf{w}^*(\mathcal{D}) \cdot \mathbf{r}(\boldsymbol{\theta}) = \sum_{\mu=1}^{P} \alpha^\mu K(\boldsymbol{\theta}^\mu, \boldsymbol{\theta}), \tag{2}$$

where the coefficient vector $\boldsymbol{\alpha} = \mathbf{K}^+\mathbf{y}$, where $^+$ denotes Moore-Penrose inverse (App. 3A), and the matrix $\mathbf{K}$ has entries $K_{\mu\nu} = K(\boldsymbol{\theta}^\mu, \boldsymbol{\theta}^\nu)$. In these expressions the population code only appears through the kernel K, showing that the kernel alone controls the learned response pattern. This result applied also to nonlinear readouts (App. 3G), showing that the kernel can control the learned solution in a variety of cases.

Biological codes are metabolically more efficient and more selective than other codes with identical kernels

The fact that learning performance depends only on the kernel introduces a large degeneracy in the set of codes which achieve identical desired performance on learning tasks. This is because the kernel is invariant with respect to left-rotations of the population code. An orthogonal transformation $\mathbf{Q}$ applied to a population code $\mathbf{r}(\boldsymbol{\theta})$ generates a new code $\tilde{\mathbf{r}}(\boldsymbol{\theta}) = \mathbf{Q}\mathbf{r}(\boldsymbol{\theta})$ with an identical kernel. Such orthogonal transformations are the most general linear transformations which preserve the kernel for all possible pairs of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$. For rank deficient neural codes, other linear transformations can preserve the kernel, however, for any linear transformation of the code $\tilde{\mathbf{r}}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{r}(\boldsymbol{\theta})$ which preserves the kernel, there exists an orthogonal matrix $\mathbf{Q}$ such that $\tilde{\mathbf{r}}(\boldsymbol{\theta}) = \mathbf{Q}\mathbf{r}(\boldsymbol{\theta})$ (App. 2A). It therefore suffices to consider orthogonal matrices $\mathbf{Q}$. Codes $\mathbf{r}(\boldsymbol{\theta})$ and $\tilde{\mathbf{r}}(\boldsymbol{\theta})$ will have identical readout performance on all possible learning tasks. We illustrate this degeneracy in Figure 2 using a publicly available dataset which consists of activity recorded from $\sim$ 20,000 neurons from the primary visual cortex of a mouse while shown static gratings [8, 9]. An original code $\mathbf{r}(\boldsymbol{\theta})$ is rotated to generate $\tilde{\mathbf{r}}(\boldsymbol{\theta})$ (Figure 2A) which have the same kernels (Figure 2B) and the same performance on a learning task (Figure 2C).

Although, the performance of linear readouts may be invariant to such rotations, metabolic efficiency may favor certain codes over others [14, 15, 16, 17, 18], reducing degeneracy in the space of codes with identical kernels. To formalize this idea, we define $\boldsymbol{\delta}$ to be the vector of spontaneous firing rates of a population of neurons, and $\mathbf{s}^\mu = \mathbf{r}(\boldsymbol{\theta}^\mu) + \boldsymbol{\delta}$ be the spiking rate vector in response to a stimulus $\boldsymbol{\theta}^\mu$. The modulation with respect to the spontaneous activity, $\mathbf{r}(\boldsymbol{\theta}^\mu)$, gives the population code and defines the kernel, $K(\boldsymbol{\theta}^\mu, \boldsymbol{\theta}^\mu) = \frac{1}{N}\mathbf{r}(\boldsymbol{\theta}^\mu) \cdot \mathbf{r}(\boldsymbol{\theta}^\nu)$. To avoid confusion with $\mathbf{r}(\boldsymbol{\theta}^\mu)$, we will refer to $\mathbf{s}^\mu$ as total spiking activity. We propose that population codes prefer smaller spiking activity

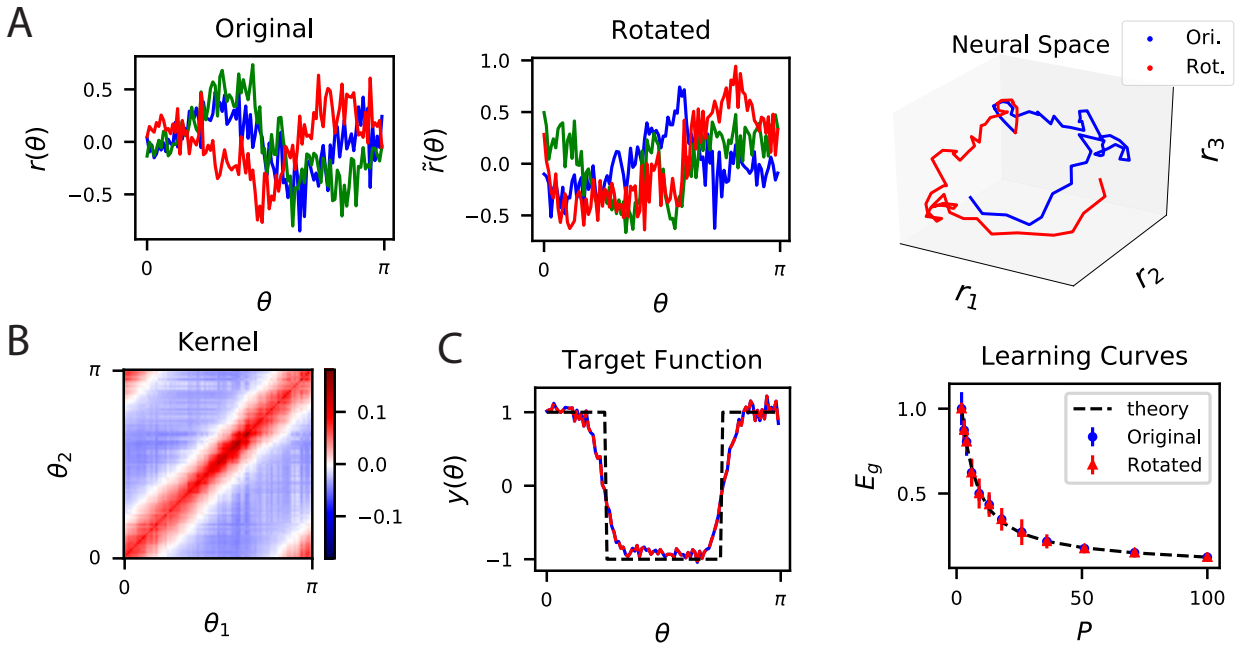

Figure 2: The inner product kernel controls the generalization performance of readouts. **A** Tuning curves $r(\theta)$ for three example recorded Mouse V1 neurons to varying static grating stimuli oriented at angle θ [8, 9] (Left) are compared with a randomly rotated version (Middle) $\tilde{\mathbf{r}}(\theta)$ of the same population code. (Right) These two codes, original (Ori.) and rotated (Rot.) can be visualized as parametric trajectories in neural space. **B** The inner product kernel matrix has elements $K(\theta_1, \theta_2)$. The original V1 code and its rotated counterpart have identical kernels. **C** In a learning task involving uniformly sampled angles, readouts from the two codes perform identically, resulting in identical approximations of the target function (shown on the left as blue and red curves) and consequently identical generalization performance as a function of training set size P (shown on right with blue and red points). The theory curve will be described in the main text.

subject to a fixed kernel. In other words, because the kernel is invariant to any change of the spontaneous firing rates and left rotations of $\mathbf{r}(\boldsymbol{\theta})$, the orientation and shift of the population code $\mathbf{r}(\boldsymbol{\theta})$ should be chosen such that the resulting total spike count $\sum_{i=1}^{N}\sum_{\mu=1}^{P} s_i^\mu$ is small.

We tested whether biological codes exhibit lower total spiking activity than others exhibiting the same kernel on mouse V1 recordings, using deconvolved calcium activity as a proxy for spiking events [8, 9, 19] (Methods 4.4; Figure 3). To compare the experimental total spiking activity to other codes with identical kernels, we computed random rotations of the neural responses around spontaneous activity, $\tilde{\mathbf{r}}(\boldsymbol{\theta}^\mu) = \mathbf{Q}\mathbf{r}(\boldsymbol{\theta}^\mu)$, and added the $\tilde{\delta}$ that minimizes total spiking activity and maintains its nonnegativity (Methods 4.4.2). We refer to such an operation as RROS (random rotation and optimal shift), and a code generated by an RROS operation as an RROS code The matrix $\mathbf{Q}$ is constructed by performing QR decomposition of a random Gaussian matrix, which guarantees sampling from the Haar measure of $O(N)$ [20]. In other words, we compare the true code to the most metabolically efficient realizations of its random rotations. This procedure may

result in an increased or decreased total spike count in the code, and is illustrated in a synthetic dataset in Figure 3A. We conducted this procedure on subsets of various sizes of mouse V1 neuron populations, as our proposal should hold for any subset of neurons (Methods 4.4.2), and found that the true V1 code is much more metabolically efficient than randomly rotated versions of the code (Figure 3B and C). This finding holds for both responses to static gratings and to natural images as we show in Figure 3B and C respectively.

To further explore metabolic efficiency, we posed an optimization problem which identifies the most efficient code with the same kernel as the biological V1 code. This problem searches over rotation matrices $\mathbf{Q}$ and finds the $\mathbf{Q}$ matrix and off-set vector $\boldsymbol{\delta}$ which gives the lowest cost $\sum_{i\mu} s_i^\mu$ (Methods 4.4.3)(Figure 3). Though the local optimum identified with the algorithm is lower in cost than the biological code, both the optimal and biological codes are significantly displaced from the distribution of random codes with same kernel. Our findings do not change when data is preprocessed with an alternative strategy, an upper bound on neural responses is imposed on rotated codes, or subsets of stimuli are considered (Figure App.1). We further verified these results on electrophysiological recordings of mouse visual cortex from the Allen Institute Brain Observatory [21], (Figure App.2). Overall, the large disparity in total spiking activity between the true and randomly generated codes with identical kernels suggests that metabolic constraints may favor the biological code over others that realize the same kernel.

The disparity between the true biological code and the RROS code is not only manifested in terms of total activity level, but also in terms of single neuron and single stimulus sparseness measures, specifically lifetime and population sparseness distributions (Methods 4.4.4) [22, 23, 24, 25]. In Figure 3D, we compare the lifetime and population sparseness distributions of the true biological code with a RROS version of the same code, revealing biological neurons have significantly higher lifetime sparseness. In App. Section 2B, we provide analytical arguments which suggest that tuning curves of optimally sparse non-negative codes with full-rank kernels will have selective tuning.

Code-task alignment governs generalization

We next examine how the population code affects generalization performance of the readout. We calculated analytical expressions of the average generalization error in a task defined by the target response $y(\boldsymbol{\theta})$ after observing P stimuli using methods from statistical physics (Methods 4.2). Because the relevant quantity in learning performance is the kernel, we leveraged results from our previous work studying generalization in kernel regression [26, 27], and approximated the generalization error averaged over all possible realizations of the training dataset of composed of P stimuli, $E_g = \langle E_g(\mathcal{D})\rangle_{\mathcal{D}}$. As P increases, the variance in E_g due to the composition of the dataset falls, and our expressions become descriptive of also the typical case. Our final analytical result is given in Equation (11) in Methods 4.2. We provide details of our calculations in Methods 4.2 and App. Section 3, and focus on their implications here.

One of our main observations is that given a population code $\mathbf{r}(\boldsymbol{\theta})$, the singular value decom-

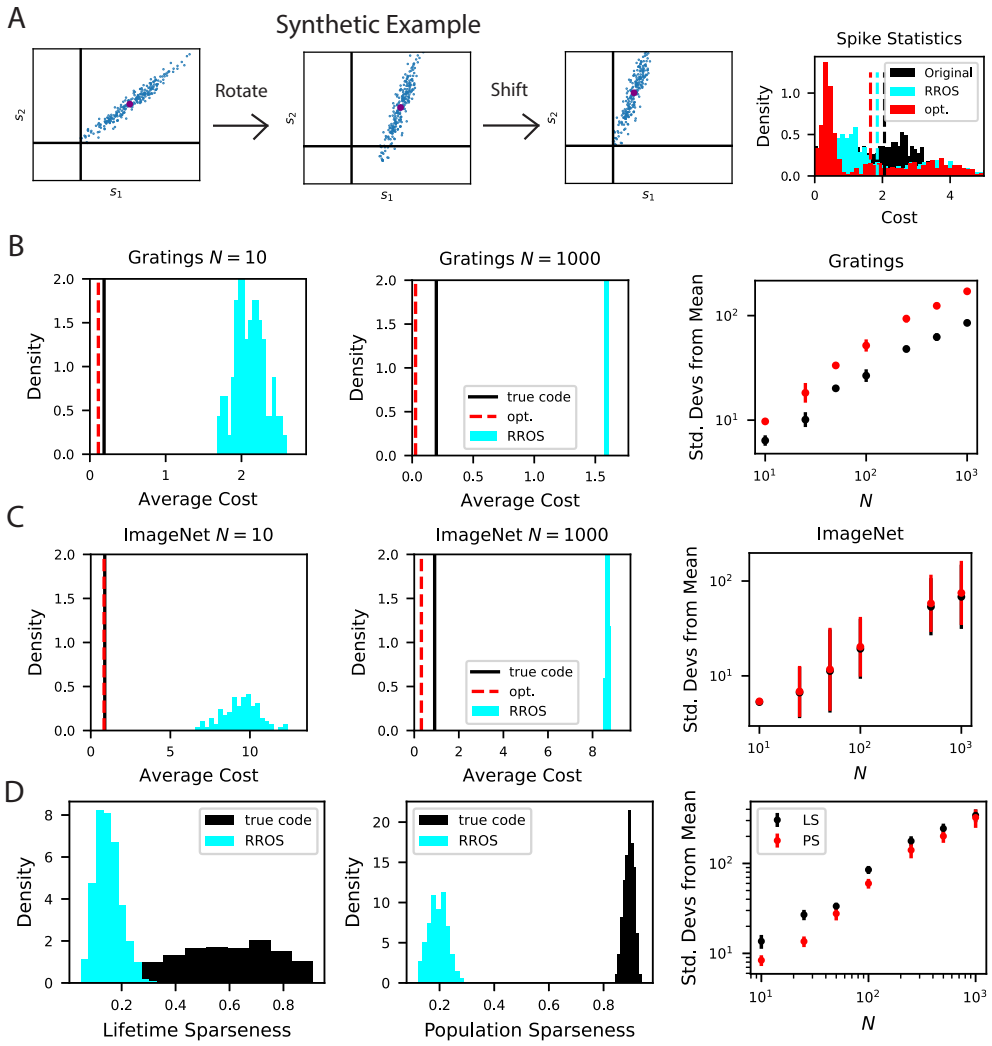

Figure 3: The biological code is more metabolically efficient than random codes with same inductive biases. **A** We illustrate our procedure in a synthetic example. A non-negative population code (left) can be randomly rotated about its spontaneous firing rate (middle), illustrated as a purple dot, and optimally shifted to a new non-negative population code (right). If the kernel is measured about the spontaneous firing rate, these transformations leave the inductive bias of the code invariant but can change the total spiking activity of the neural responses. We refer to such an operation as random rotation + optimal shift (RROS). We also perform gradient descent over rotations and shifts, generating an optimized code (opt). **B** Performing RROS on N neuron subsamples of experimental Mouse V1 recordings [8, 9], shows that the true code has much lower average cost $\frac{1}{NP}\sum_{i\mu} s_i^\mu$ compared to random rotations of the code. The set of possible RROS transformations (Methods 4.4.2,4.4.3) generates a distribution over average cost, which has higher mean than the true code. We also optimize metabolic cost over the space of RROS transformations, which resulted in the red dashed lines. We plot the distance (in units of standard deviations) between the cost of the true and optimal codes and the cost of randomly rotated codes for different neuron subsample sizes N. **C** The same experiment performed on Mouse V1 responses to ImageNet images from 10 relevant classes [11, 10]. **D** The *lifetime* (LS) and *population sparseness* (PS) levels (Methods 4.4.4) are higher for the Mouse V1 code than for a RROS code. The distance between average LS and PS of true code and RROS codes increases with N.

position of the code gives the appropriate basis to analyze the inductive biases of the readouts (Figure 4A). The tuning curves for individual neurons $r_i(\boldsymbol{\theta})$ form an N-by-M matrix $\mathbf{R}$, where M, possibly infinite, is the number of all possible stimuli. We discuss the SVD for continuous stimulus spaces in App. Section 1A. The left-singular vectors (or principal axes) and singular values of this matrix have been used in neuroscience for describing lower dimensional structure in the neural activity and estimating its dimensionality, see e.g. [28, 29, 30, 31, 32, 33, 11, 8, 34, 35, 36]. We found that the function approximation properties of the code are controlled by the singular values, or rather their squares $\{\lambda_k\}$ which give variances along principal axes, indexed in decreasing order, and the corresponding right singular vectors $\{\psi_k(\boldsymbol{\theta})\}$, which are also the kernel eigenfunctions (Methods 4.2 and App. Section 1). This follows from the fact that learned response (Eq. (2)) is only a function of the kernel K, and the eigenvalues λ_k and orthonormal eigenfunctions $\psi_k(\boldsymbol{\theta})$ collectively define the code's inner-product kernel $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ through an eigendecomposition $K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{N}\sum_{i=1}^{N} r_i(\boldsymbol{\theta}) r_i(\boldsymbol{\theta}') = \sum_k \lambda_k \psi_k(\boldsymbol{\theta})\psi_k(\boldsymbol{\theta}')$ [37] (Methods 4.2 and App. 3).

Our analysis shows the existence of a bias in the readout towards learning certain target responses faster than others. The kernel eigenfunctions form a complete basis for square integrable functions, allowing the expansion of the target response $y(\boldsymbol{\theta}) = \sum_k v_k \psi_k(\boldsymbol{\theta})$ and the learned readout response $f(\boldsymbol{\theta}) = \sum_k \hat{v}_k(\mathcal{D})\psi_k(\boldsymbol{\theta})$ in this basis. We found that the readout's generalization is better if the target function $y(\boldsymbol{\theta})$ is aligned with the top eigenfunctions ψ_k, equivalent to v_k^2 decaying rapidly with k (App. Section 3B,C). We formalize this notion by the following metric. Mathematically, generalization error $\langle E_g \rangle$ can be decomposed into normalized estimation errors E_k for the coefficients of these eigenfunctions ψ_k, $\langle E_g \rangle_{\mathcal{D}} = \sum_k v_k^2 E_k$, where $E_k = \left\langle (\hat{v}_k(\mathcal{D}) - v_k)^2 \right\rangle_{\mathcal{D}} / v_k^2$. We found that the ordering of the eigenvalues λ_k controls the rates at which these mode errors E_k decrease as P increases (App. Section 3C, [26]): $\lambda_k > \lambda_\ell \implies E_k < E_\ell$. Hence, larger eigenvalues mean lower generalization error for those normalized mode errors E_k. We term this phenomenon the *spectral bias* of the readout. Based on this observation, we propose *code-task alignment* as a principle for good generalization. To quantify code-task alignment, we use a metric which was introduced in [27] to measure the compatibility of a kernel with a learning task. This is the cumulative power distribution $C(k)$ which measures the total power in of the target function in the top k eigenmodes, normalized by the total power [27]:

$$C(k) = \frac{\sum_{\ell=1}^{k} v_\ell^2}{\sum_{\ell=1}^{\infty} v_\ell^2}. \tag{3}$$

Stimulus-response maps that have high alignment with the population code's kernel will have quickly rising cumulative power distributions $C(k)$, since a large proportion of power is placed in the top modes. Target responses with high $C(k)$ can be learned with fewer training samples than target responses with low $C(k)$ since the mode errors E_k are ordered for all P (App. Section 3C).

Probing learning biases in neural data

Our theory can be used to probe the learning biases of neural populations. Here we provide various examples of this using publicly available calcium imaging recordings from mouse primary visual cortex (V1). Our examples illustrate how our approach can be used to analyze neural population data.

We first analyzed population responses to static grating stimuli oriented at an angle θ [8, 9]. We found that the kernel eigenfunctions have sinusoidal shape with differing frequency. The ordering of the eigenvalues and eigenfunctions in Figure 4A indicates a frequency bias: lower frequency functions of θ are easier to estimate at small sample sizes.

We tested this idea by constructing two different orientation discrimination tasks shown in Figures 4B,C, where we assign static grating orientations to positive or negative valence with different frequency square wave functions of θ. We trained the readout using a subset of the experimentally measured neural responses, and measured the readout's generalization performance. We found that the cumulative power distribution for the low frequency task has a more rapidly rising $C(k)$ (Figure 4B). Using our theory of generalization, we predicted learning curves for these two tasks, which express the generalization error as a function of the number of sampled stimuli P. The error for the low frequency task is lower at all sample sizes than the hard task. The theoretical predictions and numerical experiments show perfect agreement (Figure 4B). More intuition can be gained by visualizing by projection of the neural response along the top principal axes (Figure 4C). For the low frequency task, the two target values are well separated along the top two axes. However, the high frequency task is not well separated along even the top four axes (Figure 4C).

Using the same ideas, we can use our theory to get insight into tasks which the V1 population code is ill-suited to learn. For the task of identifying mice and birds [11, 10] the linear rise in cumulative power indicates that there is roughly equal power along all kernel eigenfunctions, indicative of a representation poorly aligned to this task (Figure 4D).

To illustrate how our approach can be used for different learning problems, we evaluate the ability of linear readouts to reconstruct natural images from neural responses to those images (Figure 5). The ability to reconstruct sensory stimuli from a neural code is an influential normative principle for primary visual cortex [17]. Here we ask which aspects of the presented natural scene stimuli are easiest to learn to reconstruct. Since mouse V1 neurons tend to be selective towards low spatial frequency bands [38, 39, 40], we consider reconstruction of band-pass filtered images with spatial frequency constrained to $|\mathbf{k}| \in [\sqrt{s_{max}^2 - r^2}, s_{max}]$ for $r = 0.2$ (in units of pixels^{-1}) and plot the cumulative power $C(k)$ associated with each choice of the upper limit s_{max} (Figure 5C,D). The frequency cutoffs were chosen in this way to preserve the volume in Fourier space to $V_{\mathbf{k}} = \pi r^2$, which quantifies the dimension of the function space. We see that the lower frequency band-limited images are easier to reconstruct, as evidenced by their cumulative power $C(k)$ and learning curves E_g (Figure 5D,E). This reflects the fact that the population preferentially encodes low spatial frequency content in the image (Figure 5F). Experiments with additional values of r are provided in the Figure App.2 with additional details found in the App. Section 4.

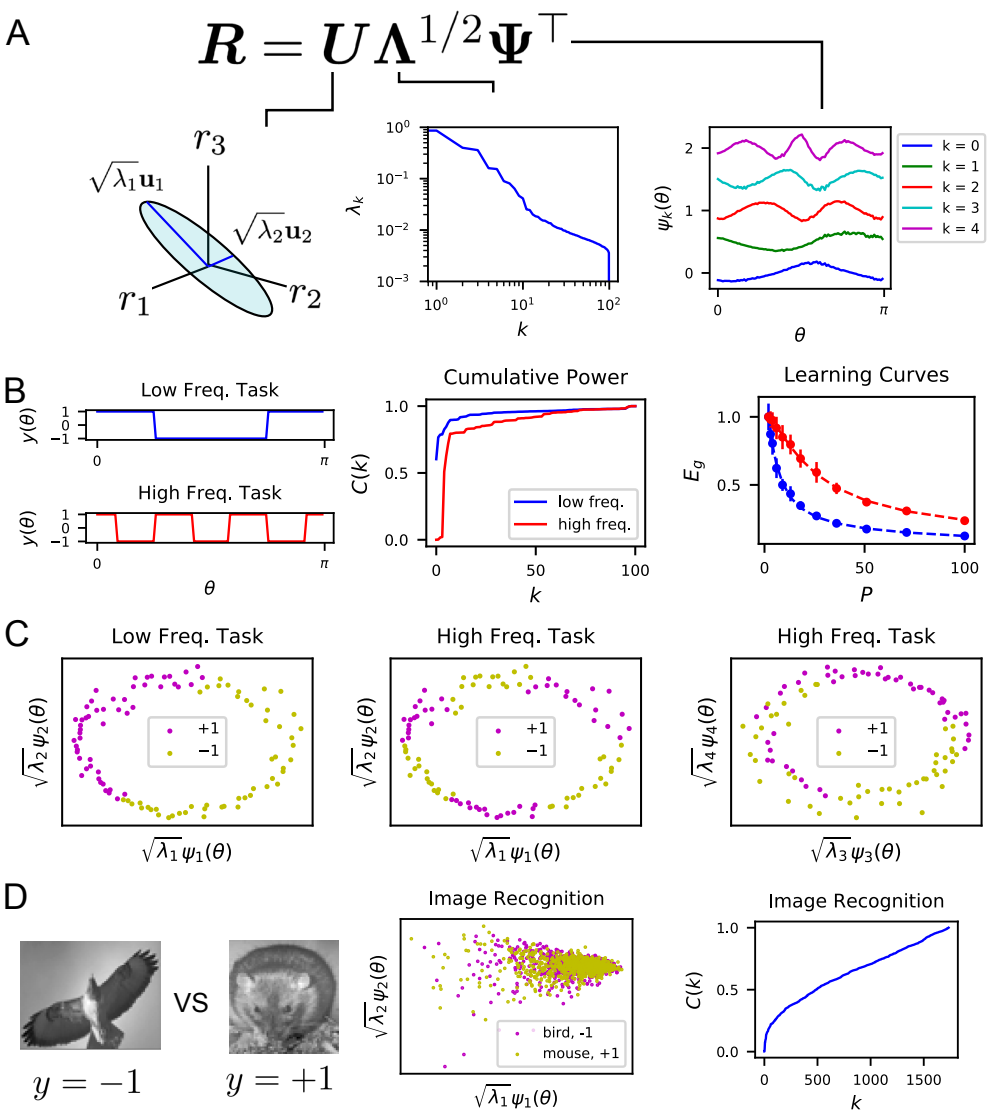

Figure 4: The singular value decomposition of the population code reveals the structure and inductive bias of the code. **A** Singular value decomposition of the response matrix $\mathbf{R}$ gives left singular vectors $\mathbf{u}_k$ (principal axes), kernel eigenvalues λ_k, and kernel eigenfunctions $\psi_k(\boldsymbol{\theta})$. The ordering of eigenvalues provides an ordering of which modes ψ_k can be learned by the code from few training examples. The eigenfunctions were offset by 0.5 for visibility. **B** (Left) Two different learning tasks $y(\theta)$, a low frequency (blue) and high frequency (red) function, are shown. (Middle) The cumulative power distribution rises more rapidly for the low frequency task than the high frequency, indicating better alignment with top kernel eigenfunctions and consequently more sample-efficient learning as shown in the learning curves (right). Dashed lines show theoretical generalization error while dots and solid vertical lines are experimental average and standard deviation over 30 repeats. **C** The feature space representations of the low (left) and high (middle and right) frequeny tasks. Each point represents the embedding of a stimulus response vector along the k-th principal axis $\mathbf{r}^\mu \cdot \mathbf{u}_k = \sqrt{\lambda_k}\psi_k(\boldsymbol{\theta}^\mu)$. The binary target value $\{\pm 1\}$ is indicated with the color of the point. The easy (left), low frequency task is well separated along the top two dimensions, while the hard, high frequency task is not linearly separable in two (middle) or even with four feature dimensions (right). **D** On an image discrimination task (recognizing birds vs mice), V1 has an entangled representation which does not allow good performance of linear readouts. This is evidenced by the top principal components (middle) and the slowly rising $C(k)$ curve (right).

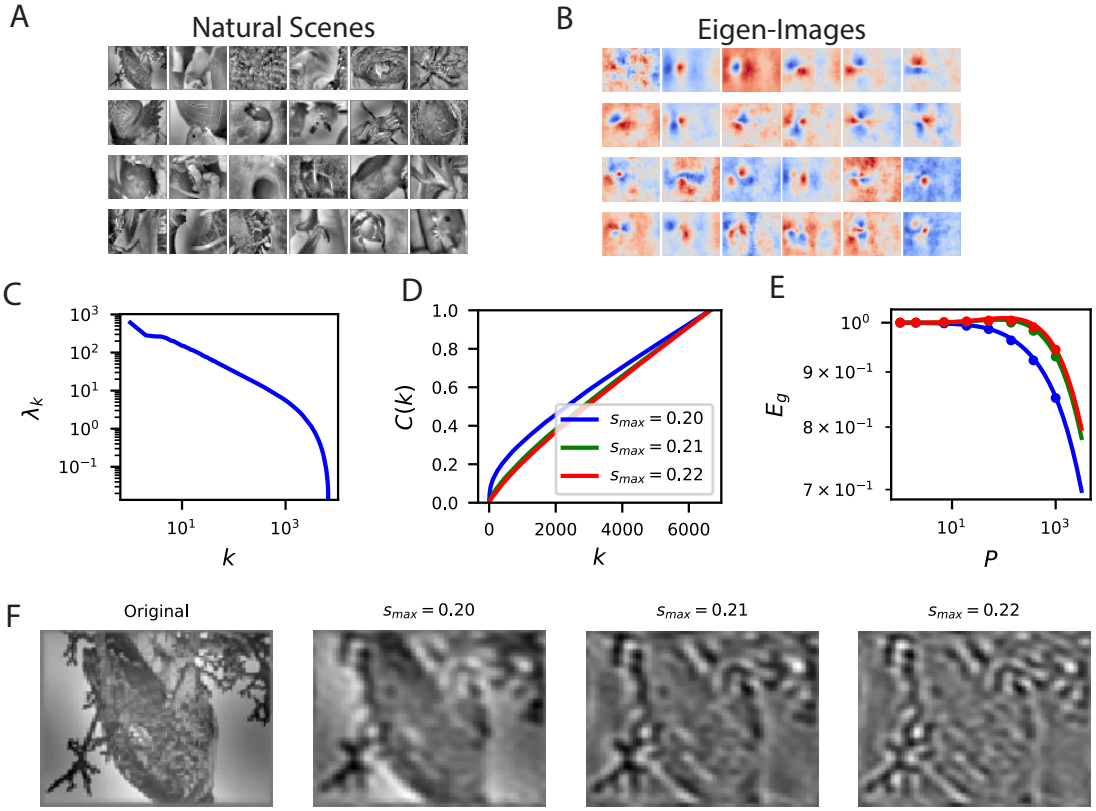

Figure 5: Reconstructing filtered natural images from V1 responses reveals preference for low spatial frequencies. **A** Natural scene stimuli $\boldsymbol{\theta}$ were presented to mice and V1 cells were recorded. **B** The images weighted by the top eigenfunctions $\mathbf{v}_k = \langle \phi_k(\boldsymbol{\theta})\boldsymbol{\theta} \rangle_{\boldsymbol{\theta}}$. These "eigenimages" collectively define the difficulty of reconstructing images through readout. **C** The kernel spectrum of the V1 code for natural images. **D** The cumulative power curves for reconstruction of band-pass filtered images. Filters preserve spatial frequencies in the range $|\mathbf{k}| \in [\sqrt{s_{max}^2 - 0.2^2}, s_{max}]$, chosen to preserve volume in Fourier space as s_{max} is varied. **E** The learning curves obey the ordering of the cumulative power curves. The images filtered with the lowest band-pass cutoff are easiest to reconstruct from the neural responses. **F** Examples of a band-pass filtered image with different preserved frequency bands.

Neural mechanisms of low frequency bias and code-task alignment in a simple model of V1

How do population level inductive biases arise from properties of single neurons? To illustrate that a variety of mechanisms may be involved in a complex manner, we study a simple model of V1 to elucidate neural mechanisms that lead to the low frequency bias at the population level. In particular, we focus on neural nonlinearities and selectivity profiles.

We model responses of V1 neurons as photoreceptor inputs passed through Gabor filters and a subsequent nonlinearity, $g(z)$, modeling a population of orientation selective simple cells (Figure 6A) (see App. 5). In this model, the kernel for static gratings with orientation $\theta \in [0, \pi]$ is of the form $K(\theta, \theta') = \kappa(|\theta - \theta'|)$, and, as a consequence, the eigenfunctions of the kernel in this setting are Fourier modes. The eigenvalues, and hence the strength of the spectral bias, are determined by the nonlinearity.

Motivated by findings in the primary visual cortex [41, 42, 43, 44], we studied the spectral bias induced by rectified power-law nonlinearities of the form $g(z) = \max\{0, z - a\}^q$. From theory side, such a power-law activation function arises in a spiking neuron when firing is driven by input fluctuations [41, 42]. Further, this activation is observed in intracellular recordings over the dynamic range of neurons in primary visual cortex [44]. For example, in cats, the power,q, ranges from 2.7 to 3.9 [43]. We fit parameters of our model to the Mouse V1 kernel and compared to other parameter sets in Figure 6B. Our best fit value of $q = 1.7$ is lower but on par with the estimates from the cat.

Computation of the kernel and its eigenvalues (App. 5A for linear g, App. 5B for threshold-power law g) indicates a low frequency bias: the eigenvalues for low frequency modes are higher than those for high frequency modes, indicating a strong inductive bias to learn functions of low frequency in the orientation. Decreasing sparsity (lower a) leads to a faster decrease in the spectrum (but similar asymptotic scaling at the tail, see App. Section 5) and a stronger bias towards lower frequency functions (Figure 6B; more comparisons in Figure App.3). The effect of the power of nonlinearity q is more nuanced: increasing power may increase spectra at lower frequencies, but may also lead to a faster decay at the tail (Figure 6B; more comparisons in Figure App.3). In general, an exponent q implies a power-law asymptotic spectral decay $\lambda_k \sim k^{-2q-2}$ as $k \to \infty$ (App. Section 5B). The behavior at low frequencies may have significant impact for learning with few samples. We discuss this in more detail in the next section. Overall, our findings show that the spectral bias of a population code can be determined in non-trivial ways by its biophysical parameters, including neural thresholds and nonlinearities.

To further illustrate the importance of code-task alignment, we next study how invariances in the code to stimulus variations may affect the learning performance. We introduce complex cells in addition to simple cells in our model with proportion $s \in [0, 1]$ of simple cells (App. Section 5C; Figure 6A), and allow phase, ϕ, variations in static gratings. We use the energy model [45, 46] to capture the phase invariant complex cell responses (App. 5B, 5C). We reason that in tasks that do not depend on phase information, complex cells should improve sample efficiency.

In this model, the kernel for the V1 population is a convex combination of the kernels for the

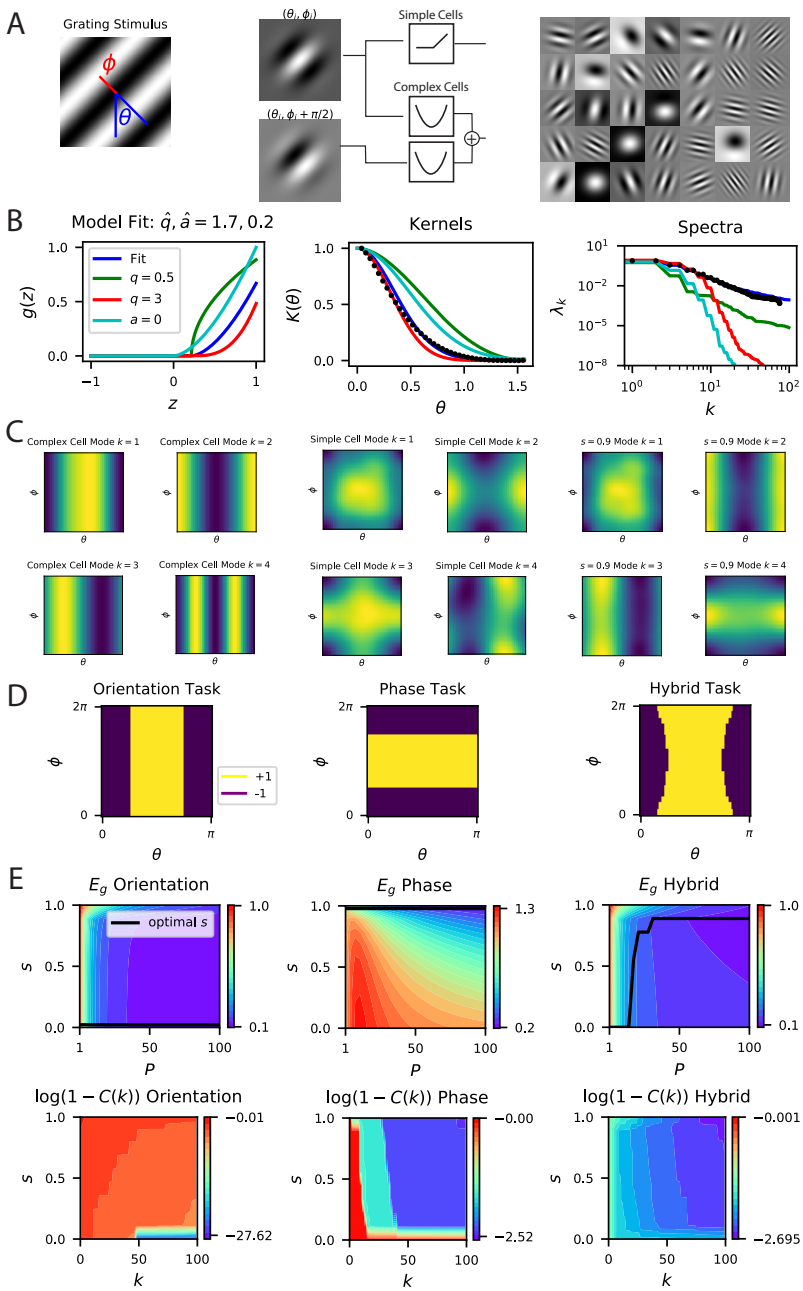

Figure 6: A model of V1 as a bank of Gabor filters recapitulates experimental inductive bias. **A** Gabor filtered inputs are mapped through nonlinearity. A grating stimulus (left) with orientation θ and phase ϕ is mapped through a circuit of simple and complex cells (middle). Some examples of randomly sampled Gabor filters (right) generate preferred orientation tuning of neurons in the population. **B** A threshold-powerlaw nonlinearity $g_{q,a}(z) = \max\{0, z - a\}^q$ is fit to the mouse V1 kernel (black dots). Kernels and spectra for alternative choices of q, a are shown (color code defined in left panel). **C** We plot eigenfunctions ψ_k (modes) for mixtures of sN simple cells and $(1 - s)N$ complex cells. A pure complex cell population has all eigenfunctions independent of phase ϕ. A pure simple cell population $s = 1$ or mixture codes $0 < s < 1$ depend on both orientation phase in a nontrivial way. **D** Three tasks are visualized, where color indicates the binary target value ± 1. The left task only depends on orientation stimulus variable θ, the middle only depends on phase ϕ, the hybrid task (right) depends on both. **E** (top) Generalization error and cumulative power distributions for the three tasks as a function of the simple-complex cell mixture parameter s. In Figure App.3 we provide more comparisons of our theory and numerical experiments.

simple and complex cell populations $K_{V1}(\theta, \theta', \phi, \phi') = sK_s(\theta, \theta', \phi, \phi') + (1-s)K_c(\theta, \theta')$ where K_s is the kernel for a pure simple cell population that depends on both orientation and phase, and K_c is the kernel of a pure complex cell population that is invariant to phase (App. 5C). Figure 6C shows top kernel eigenfunctions for various values of s elucidating inductive bias of the readout.

Figures 6D and 6E show generalization performance on tasks with varying levels of dependence on phase and orientation. On pure orientation discrimination tasks, increasing the proportion of complex cells by decreasing s improves generalization. Increasing the sensitivity to the nuisance phase variable, ϕ, only degrades performance. The cumulative power distribution is also maximized at $s = 0$. However, on a task which only depends on the phase, a pure complex cell population cannot generalize, since variation in the target function due to changes in phase cannot be explained in the codes' responses. In this setting, a pure simple cell population attains optimal performance. The cumulative power distribution is maximized at $s = 1$. Lastly, in a nontrivial hybrid task which requires utilization of both variables θ, ϕ, an optimal mixture s exists for each sample budget P which minimizes the generalization error. The cumulative power distribution is maximized at different s values depending on k, the component of the target function. This is consistent with an optimal heterogenous mix, because components of the target are learned successively with increasing sample size. In reality, V1 must code for a variety of possible tasks and we can expect a nontrivial optimal simple cell fraction s. We conclude that the degree of invariance required for the set of natural tasks, and the number of samples determine the optimal simple cell, complex cell mix.

Small and large sample size behaviors of generalization

Recently, Stringer *et al.* [11] argued that the input-output differentiability of the code may be enabling better generalization, which is in turn governed by the asymptotic rate of spectral decay. Our results provide a more nuanced view of the relation between generalization and kernel spectra. First, generalization with low sample sizes crucially depend on the top eigenvalues and eigenfunctions of the code's kernel, not the tail. Second, generalization requires alignment of the code with the task of interest. Non-differentiable codes can generalize well if there is such an alignment. To illustrate these points, here, we provide examples where asymptotic conditions on the kernel spectrum are insufficient to describe generalization performance for small sample sizes (Figure 7 and App.5), and where non-differentiable kernels generalize better than differentiable kernels (Figure App.6).

Our example demonstrates how a code allowing good generalization for large sample sizes can be disadvantageous for small sizes. In Figure 7A, we plot three different populations of neurons with smooth (infinitely differentiable) tuning curves that tile a periodic stimulus variable, such as the direction of a moving grating. The tuning width, σ, of the tuning curves strongly influences the structure of these codes: narrower widths have more high frequency content as we illustrate in a random 3D projection of the population code for $\theta \in [0, 2\pi]$ (Figure 7A). Visualization of the corresponding (von Mises) kernels and their spectra are provided in Figure 7B. The width of the tuning curves control bandwidths of the kernel spectra Figure 7B, with narrower curves

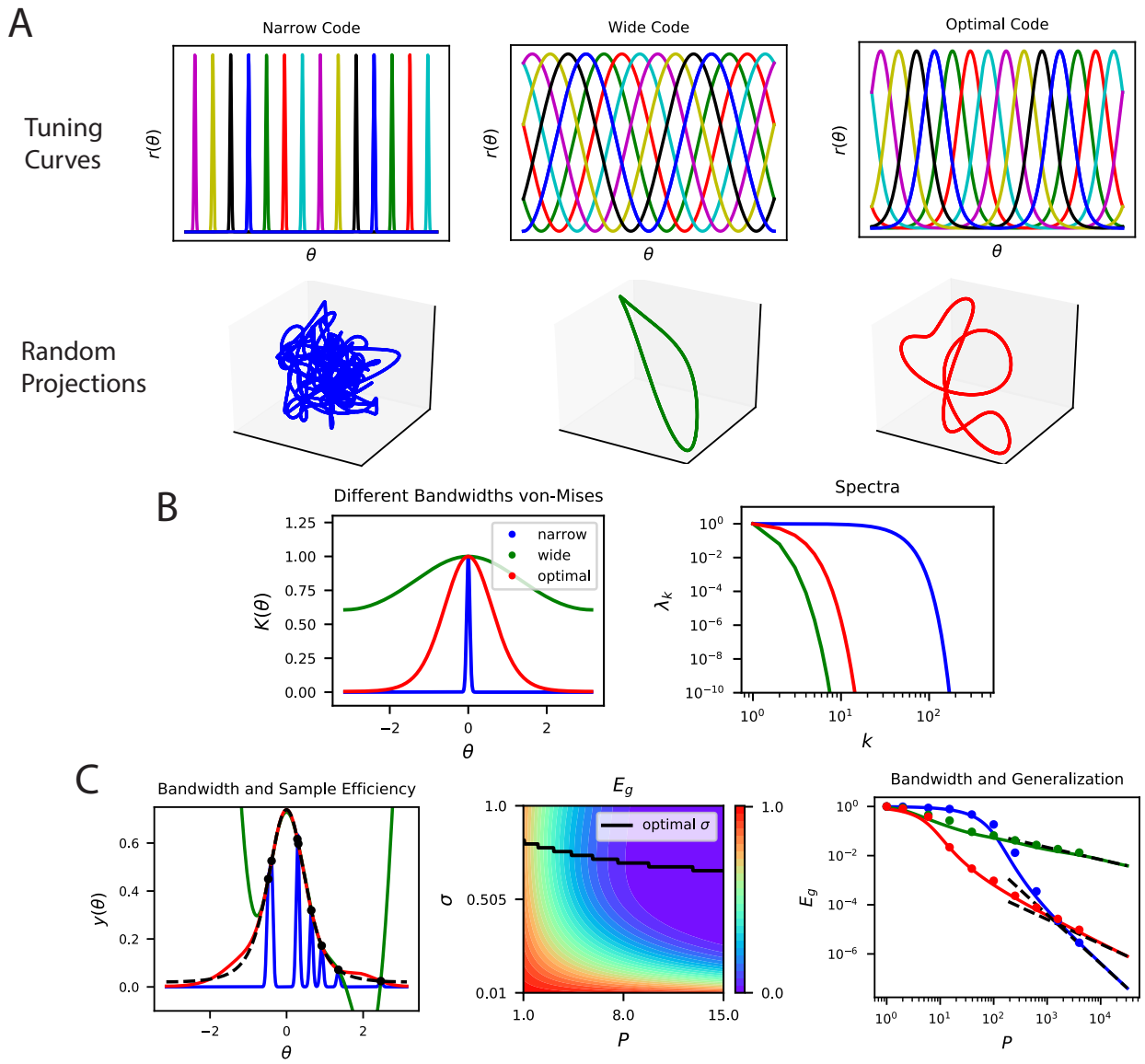

Figure 7: The top eigensystem of a code determines its low-P generalization error. **A** A periodic variable is coded by a population of neurons with tuning curves of different widths (top). Narrow, wide and optimal refers to the example in C. These codes are all smooth (infinitely differentiable) but have very different feature space representations of the stimulus variable θ, as random projections reveal (below). **B** (left) The population codes in the above figure induce von Mises kernels $K(\theta) \propto e^{\cos(\theta)/\sigma^2}$ with different bandwidths σ. (right) Eigenvalues of the three kernels. **C** (left) As an example learning task, we consider estimating a "bump" target function. The optimal kernel (red, chosen as optimal bandwidth for $P = 10$) achieves a better generalization error than either the wide (green) or narrow (blue) kernels. (middle) A contour plot shows generalization error for varying bandwidth σ and sample size P. (right) The large P generalization error scales in a power law. Solid lines are theory, dots are simulations averaged over 15 repeats, dashed lines are asymptotic power law scalings described in main text. Same color code as B and C-left.

having an later decay in the spectrum and higher high frequency eigenvalues. These codes can have dramatically different generalization performance, which we illustrate with a simple "bump" target response (Figure 7C). In this example, for illustration purposes, we let the network learn with a delta-rule with a weight decay, leading to a regularized kernel regression solution (App. 3A.1). For a sample size of $P = 10$, we observe that codes with too wide or too narrow tuning curves (and kernels) do not perform well, and there is a well-performing code with an optimal tuning curve width σ, which is compatible with the width of the target bump, σ_T. We found that optimal σ is different for each P (Figure 7C). In the large-P regime, the ordering of the performance of the three codes are reversed (Figure 7C). In this regime generalization error scales in a power law $E_g \sim P^{-\min(2, \frac{\ln \sigma_T}{\ln \sigma})}$ (App. 3D) and the narrow code, which performed worst for $P \sim 10$, performs the best. This example demonstrates that asymptotic conditions on the tail of the spectra are insufficient to understand generalization in the small sample size limit. The bulk of the kernel's spectrum needs to match the spectral structure of the task to generalize efficiently in the low-sample size regime. However, for large sample sizes, the tail of the eigenvalue spectrum becomes important. We repeat the same exercise and draw the same conclusions for a non-differentiable kernel (Laplace) (Figure App.5) showing that these results are not an artifact of the infinite differentiability of von Mises kernels. We further provide examples where non-differentiable kernels generalizing better than differentiable kernels in Figure App.5.

Time-Dependent Neural Codes

Our framework can directly be extended to learning of arbitrary time-varying functions of time-varying inputs from an arbitrary spatiotemporal population code (Methods 4.3, App. 6). In this setting, the population code $\mathbf{r}(\{\boldsymbol{\theta}(t)\}, t)$ is a function of an input stimulus sequence $\boldsymbol{\theta}(t)$ and possibly its entire history, and time t. A downstream linear readout $f(\{\boldsymbol{\theta}\}, t) = \mathbf{w} \cdot \mathbf{r}(\{\boldsymbol{\theta}\}, t)$ learns a target sequence $y(\{\boldsymbol{\theta}\}, t)$ from a total of $\mathcal{P}$ examples that can come at any time during any sequence. Learning is again achieved through the delta-rule and the learned function can be expressed as a linear combination of the kernel evaluated at the $\mathcal{P}$ examples. The kernel in this case is a more complicated object that computes inner products of neural population vectors at different times t, t' for different input sequences $\{\boldsymbol{\theta}\}, \{\boldsymbol{\theta}'\}$: $K(\{\boldsymbol{\theta}\}, \{\boldsymbol{\theta}'\}, t, t') = \frac{1}{N}\mathbf{r}(\{\boldsymbol{\theta}\}, t) \cdot \mathbf{r}(\{\boldsymbol{\theta}\}, t')$ [47, 48, 49]. Our theory carries over from the static case with appropriate modifications (Methods 4.3). Kernels whose top eigenfunctions have high alignment with the target time-varying response $y(\{\boldsymbol{\theta}\}, t)$ will achieve the best average case generalization performance.

As a concrete example, we focus on readout from a temporal population code generated by a recurrent neural network in a task motivated by a delayed reach task [50] (Figure 8A,B). We consider a randomly connected network with current dynamics $\tau \dot{\mathbf{z}}(t) = -\mathbf{z}(t) + \mathbf{W}_{\mathbf{r}}\mathbf{r}(t) + \mathbf{W}_{\boldsymbol{\theta}}\boldsymbol{\theta}(t)$ where the rates are related to input currents through a tanh nonlinearity $\mathbf{r}(t) = \tanh(\mathbf{z}(t))$. The recurrent weights are drawn from a normal distribution $W^{\mathbf{r}}_{ij} \sim \mathcal{N}(0, g^2/N)$ and the input encoding weights from $W^{\theta}_{ij} \sim \mathcal{N}(0, 1)$ (Methods 4.3). The gain parameter g was set to 1.5 to generate rich dynamics [51]. In this task, the network is presented for a short time an input cue sequence

coding an angular variable which is drawn randomly from a distribution (Figure 8C). The recurrent neural network must remember this angle and reproduce an output sequence which is a simple step function whose height depends on the angle which begins after a time delay from the cessation of input stimulus and lasts for a short time (Figure 8D).

The kernel induced by the spatiotemporal code is shown in Figure 8E. The high dimensional nature of the activity in the recurrent network introduces complex and rich spatiotemporal similarity structure. Figure 8F shows the kernel's eigensystem, which consists of stimulus dependent time-series $\psi_k(\{\boldsymbol{\theta}\};t)$ for each eigenvalue λ_k. An interesting link can be made with this eigensystem and linear low-dimensional manifold dynamics observed in several cortical areas [28, 29, 31, 52, 33, 53, 36, 32, 54, 30]. The kernel eigenfunctions also define the latent variables obtained through a singular value decomposition of the neural activity $\mathbf{r}(\{\boldsymbol{\theta}\};t) = \sum_k \sqrt{\lambda_k}\mathbf{u}_k\psi_k(\{\boldsymbol{\theta}\};t)$ [31].

With enough samples, the readout neuron can learn to output the desired angle with high fidelity (Figure 8G). Unsurprisingly, tasks involving long time delays are more difficult and exhibit lower cumulative power curves. Consequently, the generalization error for small delay tasks drops much more quickly with increasing $\mathcal{P}$.

3 Discussion

Elucidating inductive biases of the brain is fundamentally important for understanding natural intelligence [3, 4, 5, 55]. These biases are coded into the brain by the dynamics of its neurons, the architecture of its networks, its representations and plasticity rules. Finding ways to extract the inductive biases from neuroscience datasets requires a deeper theoretical understanding of how all these factors shape the biases, and is an open problem. In this work, we attempted to take a step towards filling this gap by focusing on how the structure of static neural population codes shape inductive biases for learning of a linear readout neuron under a biologically plausible learning rule. If the readout neuron's output is correlated with behavior, and that correlation is known, then our theory could possibly be modified to predict what behavioral tasks can be learned faster.

Under the delta rule, the generalization performance of the readout is entirely dependent on the code's inner product kernel; the kernel is a determinant of inductive bias. In its finite dimensional form, the kernel is an example of a representational similarity matrix and is a commonly used tool to study neural representations [56, 57, 58, 59, 60, 61]. Our work elucidates a concrete link between this experimentally measurable mathematical object, and sample-efficient learning.

We derived an analytical expression for the generalization error as a function of sample-size under very general conditions, for an arbitrary stimulus distribution, arbitrary population code and an arbitrary target stimulus-response map. We used our findings in both theoretical and experimental analysis of primary visual cortex, and temporal codes in a delayed reach task. This generality of our theory is a particular strength.

Our analysis elucidated two principles that define the inductive bias. The first one is spectral

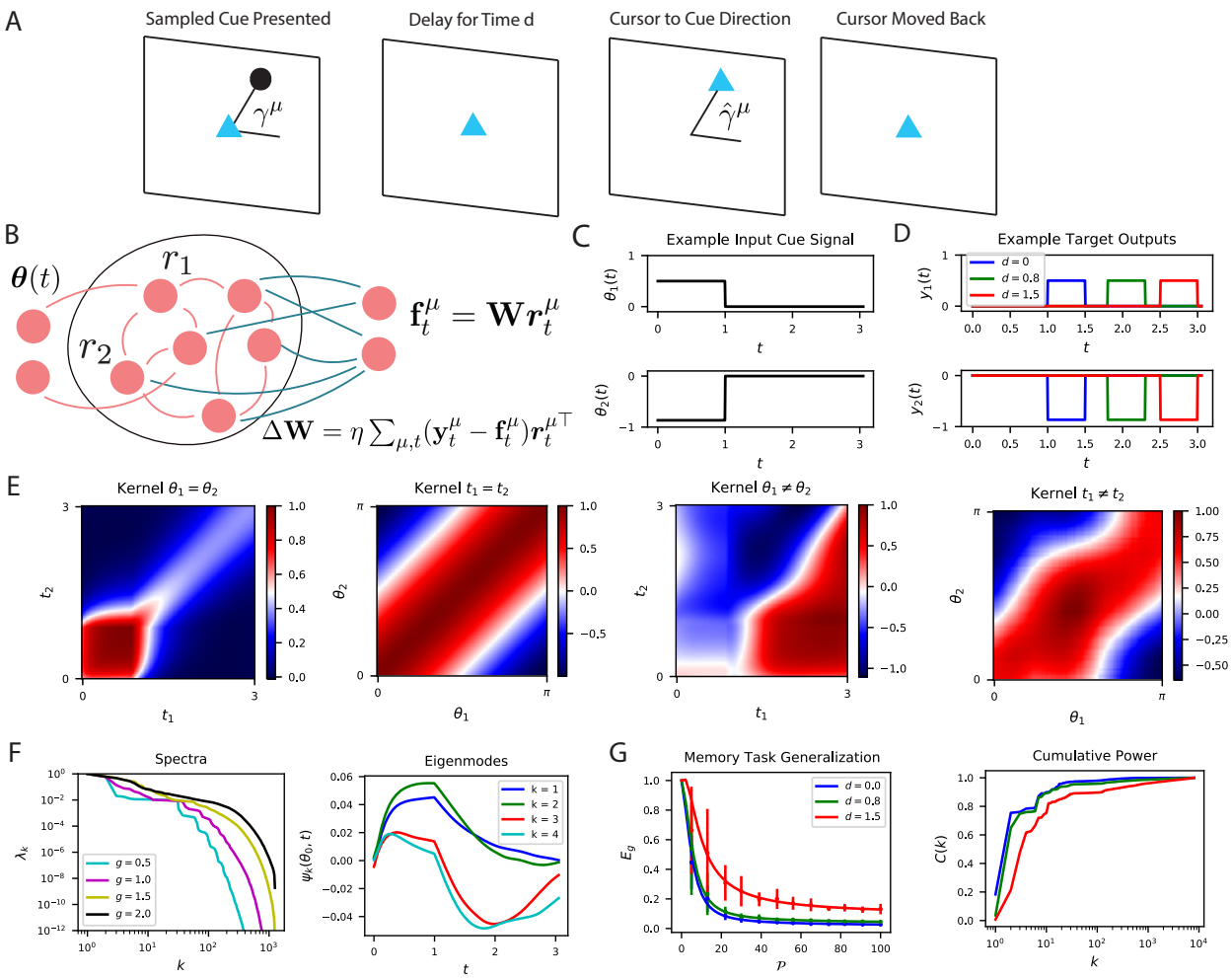

Figure 8: The performance of time-dependent codes when learning dynamical systems can be understood through spectral bias. **A** We study the performance of time dependent codes on a delayed response task which requires memory retrieval. A cue (black dot) is presented at an angle γ^μ. After a delay time d, the cursor position (blue triangle) must be moved to the remembered cue position and then subsequently moved back to the origin after a short time. **B** The readout weights (blue) of a time dependent code can be learned through a modified delta rule. **C** Input is presented to the network as a time series which terminates at $t = 1$. The sequences are generated by drawing an angle $\gamma^\mu \sim \text{Uniform}[0, 2\pi]$ and using two step functions as input time-series that code for the cosine and the sine of the angle (Methods 4.3, App. 6). We show an example of the one of the variables in a input sequence. **D** The target functions for the memory retrieval task are step functions delayed by a time d. **E** The kernel $K_{\mu,\mu',t,t'}$ compares the code for two sequences at two distinct time points. We show the time dependent kernel for identical sequences (left) and the stimulus dependent kernel for equal time points (middle left) as well as for non-equal stimuli (middle right) and non-equal time (right). **F** The kernel can be diagonalized, and the eigenvalues λ_k determine the spectral bias of the reservoir computer (left). We see that higher gain g networks have higher dimensional representations. The "eigensystems" $\psi_k(\theta^\mu, t)$ are functions of time and cue angle. We plot only $\mu = 0$ components of top systems $k = 1, 2, 3, 4$ (right). **G** The readout is trained to approximate a target function $y^\mu(t)$, which requires memory of the presented cue angle. (left) The theoretical (solid) and experimental (vertical errorbar, 100 trials) generalization error E_g are plotted for the three delays d against training sample size $\mathcal{P}$. (right) The ordering of E_g matches the ordering of the $C(k)$ curves as expected.

bias: kernel eigenfunctions with large eigenvalues can be estimated using a smaller number of samples. The second principle is the code-task alignment: Target functions with most of their power in top kernel eigenfunctions can be estimated efficiently and are compatible with a code. The cumulative power distribution, $C(k)$ [27], provides a measure of this alignment. These findings define a notion of "simplicity" bias in learning from examples, and provides a solution to the question of what stimulus-response maps are easier to learn. A similar simplicity bias has been also observed in training deep neural networks [62, 63, 64]. Due to a correspondence between trained neural networks in the infinite-width limit and kernel machines [65], results on the spectral bias of kernel machines may shed light onto these findings [26, 27].

We used our finding in both theoretical and experimental analysis of mouse primary visual cortex. We demonstrated a bias of neural populations towards low frequency orientation discrimination and low spatial frequency reconstruction tasks. The latter finding is consistent with the finding that mouse visual cortex neurons are selective for low spatial frequency [38, 39, 40]. Our modeling showed that these biases can be modified by the neural thresholds and nonlinearities in nontrivial ways. The theoretical model of the visual cortex as a mixture of simple and complex cells demonstrated how invariances, specifically the phase invariance of the complex cells, in the population code can facilitate learning some tasks involving phase invariant responses at the expense of performance on others. Role of invariances in learning with kernel methods have recently been investigated in machine learning literature [66, 67].

A recent proposal considered the possibility that the brain acts as an overparameterized interpolator [68]. Suitable inductive biases are crucial to prevent overfitting and generalize well in such a regime [69]. Our theory can explain these inductive biases since, when the kernel is full-rank, which typically is the case when there are more neurons in the population than the number of learning examples, the delta rule converges to an interpolator of the learning examples. Modern deep learning architectures also operate in an overparameterized regime, but generalize well [70, 69], and an inductive bias towards simple functions has been proposed as an explanation [26, 27, 64, 71].

Our work suggests sample efficiency as a general coding principle for neural populations, relating neural representations to the kinds of problems they are well suited to solve. These codes may be shaped through evolution or themselves be learned through experience [55]. Prior related work demonstrated the dependence of sample-efficient learning of a two-angle estimation task on the width of the individual neural tuning curves [72] and additive function approximation properties of sparsely connected random networks [73].

A sample efficiency approach to population coding differs from the classical efficient coding theories [16, 14, 15, 74, 75, 76, 17, 77], which postulate that populations of neurons optimize information content of their code subject to metabolic constraints or noise. While these theories emphasize different aspect of the code's information content (such as reduced redundancy, predictive power, or sparsity), they do not address sample efficiency demands on learning. Further, recent studies demonstrated hallmarks of redundancy and correlation in population responses [54, 30, 78, 36, 79, 80, 11], violating a generic prediction of efficient coding theories that responses of

different neurons should be uncorrelated across input stimuli in high signal-to-noise regimes to reduce redundancy in the code and maximize information content [14, 15, 74, 75, 81, 82]. In our theory, the structured correlations of neural responses correspond to the decay in the spectrum of the kernel, and play a key role in biasing learned readouts towards simple functions.

In recent related studies, the asymptotic decay rate of the kernel's eigenspectrum was argued to be important for generalization [11] and robustness [83]. Decay rate in the mouse V1 was found to be consistent with a high dimensional (power law) but smooth (differentiable) code, and smoothness was argued to be an enabler of generalization [11]. We show that sample-efficient learning requires more than smoothness conditions in the form of asymptotic decay rates on the kernel's spectrum. The interplay between the stimulus distribution, target response and the code gives rise to sample efficient learning. Because of spectral bias, the top eigenvalues govern the small sample size behavior. The tail of the spectrum becomes important for large sample sizes.

Though the kernel is degenerate with respect to rotations of the code in the neural activity space, we demonstrated that the true V1 code has much lower average activity than random codes with the same kernel, suggesting that evolution and learning may be selecting neural codes with low average spike rates which preserve sample-efficiency demands for downstream learning tasks. We predict that metabolic efficiency may be a determinant in the orientation and placement of the ubiquitously observed low-dimensional coding manifolds [53, 80] in neural activity space in other parts of the brain. The demand of metabolic efficiency is consistent with prior sparse coding theories [84, 17, 18, 85], however, our theory emphasizes sample-efficient learning as the primary normative objective for the code.

Our work constitutes a first step towards understanding inductive biases in neuronal circuits. To achieve this, we focused on a linear, delta-rule readout of a static population code. More work is need to study other factors that affect inductive bias. Importantly, sensory neuron tuning curves can adapt during perceptual learning tasks [86, 87, 88, 89] with the strength of adaptation dependent on brain area [90, 91, 92, 93]. This motivates analysis of learning in multi-layer networks. Our work focused on the effect of signal correlations to coding and inductive bias [94, 95]. Future analysis could study how signal and noise correlations interact to shape inductive bias and determine generalization. Other factors that need to be taken into account are readout nonlinearities, alternative learning rules, initialization schemes, and network architectures.

4 Methods

4.1 Generating example codes (Figure 1)

The two codes in Figure 1 were constructed to produce two different kernels for $\theta \in S^1$:

$$K_1(\theta, \theta') = \exp(0.25 \cos(\theta - \theta')) \ , \ K_2(\theta, \theta') = \sum_{k=1}^{20} \cos(k(\theta - \theta')). \tag{4}$$

An infinite number of codes could generate either of these kernels. After diagonalizing the kernel into its eigenfunctions on a grid of 120 points, $\mathbf{K}_1 = \mathbf{\Psi}_1 \mathbf{\Lambda}_1 \mathbf{\Psi}_1^\top, \mathbf{K}_2 = \mathbf{\Psi}_2 \mathbf{\Lambda}_2 \mathbf{\Psi}_2^\top$, we used a random rotation matrix $\mathbf{Q} \in O(120)$ to generate a valid code

$$\mathbf{R}_1 = \mathbf{Q}\mathbf{\Lambda}_1^{1/2}\mathbf{\Psi}_1 \ , \ \mathbf{R}_2 = \mathbf{Q}\mathbf{\Lambda}_2^{1/2}\mathbf{\Psi}_2. \tag{5}$$

This construction guarantees that $\mathbf{R}_1^\top \mathbf{R}_1 = \mathbf{K}_1$ and $\mathbf{R}_2^\top \mathbf{R}_2 = \mathbf{K}_2$. We plot the tuning curves for the first three neurons. The target function in the first experiment is $y = \cos(\theta) - 0.6\cos(4\theta)$, while the second experiment used $y = \cos(6\theta) - \cos(8\theta)$.

4.2 Theory of Generalization

Recent work has established analytic results that predict the average case generalization error for kernel regression

$$E_g = \langle E_g(\mathcal{D}) \rangle_{\mathcal{D}} = \left\langle [f(\boldsymbol{\theta}, \mathbf{w}_{\mathcal{D}}) - y(\boldsymbol{\theta})]^2 \right\rangle_{\boldsymbol{\theta},\mathcal{D}} \tag{6}$$

where $E_g(\mathcal{D}) = \left\langle [f(\boldsymbol{\theta}, \mathbf{w}_{\mathcal{D}}) - y(\boldsymbol{\theta})]^2 \right\rangle_{\boldsymbol{\theta}}$ is the generalization error for a certain sample $\mathcal{D}$ of size P and $f(\boldsymbol{\theta}, \mathbf{w}_{\mathcal{D}}) = \mathbf{w}_{\mathcal{D}} \cdot \mathbf{r}(\boldsymbol{\theta})$ is the kernel regression solution for $\mathcal{D}$ (App. 3A) [26, 27]. The typical or average case error E_g is obtained by averaging over all possible datasets of size P. This average case generalization error is determined solely by the decomposition of the target function $y(\mathbf{x})$ along the eigenbasis of the kernel and the eigenspectrum of the kernel. This continuous diagonalization again takes the form (App. 1A) [96]

$$\int p(\boldsymbol{\theta}) K(\boldsymbol{\theta}, \boldsymbol{\theta}') \psi_k(\boldsymbol{\theta}) d\boldsymbol{\theta} = \lambda_k \psi_k(\boldsymbol{\theta}'). \tag{7}$$

Our theory is also applicable to discrete stimuli if $p(\boldsymbol{\theta})$ is a Dirac measure as we describe in (App. 1B). Since the eigenfunctions form a complete set of square integrable functions [96], we expand both the target function $y(\boldsymbol{\theta})$ and the learned function $f(\boldsymbol{\theta}, \mathbf{w}_{\mathcal{D}})$ in this basis $y(\boldsymbol{\theta}) = \sum_k v_k \psi_k(\boldsymbol{\theta})$, $f(\boldsymbol{\theta}, \mathbf{w}_{\mathcal{D}}) = \sum_k \hat{v}_k(\mathbf{w}_{\mathcal{D}}) \psi_k(\boldsymbol{\theta})$. Due to the orthonormality of the kernel eigenfunctions $\{\psi_k\}$, the generalization error for any set of coefficients $\hat{\mathbf{v}}$ is

$$E_g(\mathcal{D}) = \left\langle (y(\boldsymbol{\theta}) - f(\boldsymbol{\theta}, \mathbf{w}_{\mathcal{D}}))^2 \right\rangle_{\boldsymbol{\theta}} = ||\hat{\mathbf{v}}(\mathbf{w}_{\mathcal{D}}) - \mathbf{v}||^2 \tag{8}$$

We now introduce the equivalent training error, or empirical loss, written directly in terms of eigenfunction coefficients $\hat{\mathbf{v}}$, which depends on the disorder in the dataset $\mathcal{D} = \{(\boldsymbol{\theta}^\mu, y^\mu)\}_{\mu=1}^P$

$$H(\hat{\mathbf{v}}, \mathcal{D}) = \sum_\mu [(\hat{\mathbf{v}} - \mathbf{v}) \cdot \boldsymbol{\psi}(\boldsymbol{\theta}^\mu)]^2 + \lambda \sum_k \frac{\hat{v}_k^2}{\lambda_k} \tag{9}$$

It is straightforward to verify that the H-minimizing coefficients are $\hat{\mathbf{v}}^*(\mathcal{D}) = (\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda \mathbf{\Lambda}^{-1})^{-1} \mathbf{\Psi}\mathbf{\Psi}^\top \mathbf{v}$, giving the learned function $f(\boldsymbol{\theta}, \mathcal{D}) = \hat{\mathbf{v}}^*(\mathcal{D}) \cdot \boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{k}_{\mathcal{D}}(\boldsymbol{\theta})^\top (\mathbf{K}_{\mathcal{D}} + \lambda \mathbf{I})^{-1} \mathbf{y}_{\mathcal{D}}$, whose $\lambda \to 0$ limit coincides with kernel interpolation. This allows us to characterize generalization without reference

to learned readout weights $\mathbf{w}$. The generalization error for this optimal function is

$$E_g(\mathcal{D}) = ||\hat{\mathbf{v}}^*(\mathcal{D}) - \mathbf{v}||^2 = \mathbf{v}^\top \mathbf{\Lambda}^{-1} \mathbf{G}(\mathcal{D})^2 \mathbf{\Lambda}^{-1} \mathbf{v}$$
$$\mathbf{G}(\mathcal{D}) = \left(\frac{1}{\lambda} \mathbf{\Psi}\mathbf{\Psi}^\top + \mathbf{\Lambda}^{-1} \right)^{-1}. \tag{10}$$

We note that the dependence on the randomly sampled dataset $\mathcal{D}$ only appears through the matrix $\mathbf{G}(\mathcal{D})$. Thus to compute the *typical* generalization error we need to average over this matrix $\langle \mathbf{G}(\mathcal{D}) \rangle_{\mathcal{D}}$. There are multiple strategies to perform such an average and we will study one here based on a partial differential equation which was introduced in [97, 98] and studied further in [26]. We describe in detail one method for performing such an average in App. 3B. After this computation, we find that the generalization error can be approximated at large P as

$$E_g = \frac{\kappa^2}{1-\gamma} \sum_k \frac{v_k^2}{(\lambda_k P + \kappa)^2} \ , \ \kappa = \lambda + \kappa \sum_k \frac{\lambda_k}{\lambda_k P + \kappa}, \tag{11}$$

where $\gamma = P \sum_k \frac{\lambda_k^2}{(\lambda_k p + \kappa)^2}$, giving the desired result. We note that (11) defines the function κ implicitly in terms of the sample size P. Taking $\lambda \to 0$ gives the generalization error of the minimum norm interpolant, which desribes the generalization error of the solution. This result was recently reproduced using the replica method from statistical mechanics in an asymptotic limit where the number of neurons and samples are large [26, 27]. Other recent works have verified our theoretical expressions on a variety of kernels and datasets [99, 100].

4.3 RNN Experiment

For the simulations in Figure 8 we integrated a rate based recurrent network model with $N = 6000$ neurons, time constant $\tau = 0.05$ and gain $g = 1.5$. Each of the $P = 80$ randomly chosen angles γ^μ generates a trajectory over $T = 100$ equally spaced points in $t \in [0, 3]$. The two dimensional input sequence is simply $\boldsymbol{\theta}(t) = H(t)H(1-t)\left[\cos(\gamma^\mu), \sin(\gamma^\mu)\right]^\top \in \mathbb{R}^2$. Target function for a delay d is $\mathbf{y}(\theta^\mu, t) = H(1.5+d-t)H(t-d-1)[\cos(\gamma^\mu), \sin(\gamma^\mu)]^\top$ which is nonzero for times $t \in [1+d, 1.5+d]$. In each simulation, the activity in the network is initialized to $\mathbf{u}(0) = \mathbf{0}$. The kernel gram matrix $\mathbf{K} \in \mathbb{R}^{PT \times PT}$ is computed by taking inner products of the time varying code at for different inputs γ^μ and at different times. Learning curves represent the generalization error obtained by randomly sampling $\mathcal{P}$ time points from the PT total time points generated in the simulation process and training readout weights $\mathbf{w}$ to convergence with gradient descent.

4.4 Data Analysis

4.4.1 Data source and processing

Mouse V1 neuron responses to orientation gratings were obtained from a publicly available dataset [8, 9]. Two-photon calcium microscopy fluorescence traces were deconvolved into spike trains and spikes were counted for each stimulus, as described in [8]. The presented grating angles were distributed uniformly over $[0, 2\pi]$ radians. Data pre-processing, which included z-scoring against the mean and standard deviation of null stimulus responses, utilized the provided code for this experiment, which also publicly available at `https://github.com/MouseLand/stringer-et-al-2019`. This preprocessing technique was used in all Figures in the paper. To reduce corruption of the estimated kernel from neural noise (trial-to-trial variability), we first trial average responses, binning the grating stimuli oriented at different angles θ into a collection of 100 bins over the interval from $[0, 2\pi]$ and averaging over all of the available responses from each bin. Since grating angles were sampled uniformly, there is a roughly even distribution of about 45 responses in each bin. After trial averaging, SVD was performed on the response matrix $\mathbf{R}$, generating the eigenspectrum and kernel eigenfunctions as illustrated in Figure 4. Figures 2, 3, 4, all used this data anytime responses to grating stimuli were mentioned.

In Figures 3C 4D and 5, the responses of mouse V1 neurons to ImageNet images [101] were obtained from a different publicly available dataset [10, 11]. The images were taken from 15 different classes from the Imagenet dataset with ethological relevance to mice (birds, cats, flowers, hamsters, holes, insects, mice, mushrooms, nests, pellets, snakes, wildcats, other animals, other natural, other man made). In the experiment in Figure 4D we use all images from the mice and birds category for which responses were recorded. The preprocessing code and image category information were obtained from the publicly available code base at `https://github.com/MouseLand/stringer-pachitariu-et-al-2018b`. Again, spike counts were obtained from deconvolved and z-scored calcium fluorescence traces. In the reconstruction experiment shown in Figure 5 we use the entire set of images for which neural responses were recorded.

4.4.2 Generating RROS codes

In Figure 3, the randomly rotated codes are generated by sampling a matrix $\mathbf{Q}$ from the Haar measure on the set of N-by-N orthogonal matrices [20], and chosing a $\boldsymbol{\delta}$ by solving the following optimization problem:

$$\min_{\boldsymbol{\delta}\in\mathbb{R}^N} \sum_{i=1}^{N}\sum_{\mu=1}^{P} s_i^\mu \ , \ \text{s.t.} \ \ \mathbf{s}^\mu = \mathbf{Q}\mathbf{r}(\boldsymbol{\theta}^\mu) + \boldsymbol{\delta} \ , \ s_i^\mu \geq 0 \ , \ i \in [N] \ , \ \mu \in [P], \tag{12}$$

which minimizes the total spike count subject to the kernel and nonnegativity of firing rates. The solution to this problem is given by $\delta_i^* = -\min_{\mu=1,\ldots,P}[\mathbf{Q}\mathbf{r}(\boldsymbol{\theta}^\mu)]_i$.

4.4.3 Comparing Sparsity of Population Codes

To explore the metabolic cost among the set of codes with the same inductive biases, we estimate the distribution of average spike counts of codes with the same inner product kernel as the biological code. These codes are generated in the form $\mathbf{s}^\mu = \mathbf{Q}\mathbf{r}^\mu + \boldsymbol{\delta}$ where $\boldsymbol{\delta}$ solves the optimization problem

$$\min_{\boldsymbol{\delta}\in\mathbb{R}^N} \sum_{i,\mu} s_i^\mu \ , \ s.t. \ \mathbf{s}^\mu = \mathbf{Q}\mathbf{r}^\mu + \boldsymbol{\delta} \ , \ s_i^\mu \geq 0 \tag{13}$$

To quantify the distribution of such codes, we randomly sample $\mathbf{Q}$ from the Haar-measure on $O(N)$ and compute the optimal $\boldsymbol{\delta}$ as described above. This generates the aqua colored distribution in Figure 3 B and C.

We also attempt to characterize the most efficient code with the same inner product kernel

$$\min_{\mathbf{Q}\in O(N),\boldsymbol{\delta}} \sum_{i,\mu} s_i^\mu \ , \ s.t. \ \mathbf{s}^\mu = \mathbf{Q}\mathbf{r}^\mu + \boldsymbol{\delta} \ , \ s_i^\mu \geq 0. \tag{14}$$

Since this optimization problem is non-convex in $\mathbf{Q}$, there is no theoretical guarantee that minima are unique. Nonetheless, we attempt to optimize the code by starting $\mathbf{Q}$ at the identity matrix and conduct gradient descent in the tangent space $so(N)$. Such updates take the form

$$\mathbf{Q}_{t+1} = \exp(-\eta\nabla\mathcal{L})\mathbf{Q}_t \ , \ \nabla\mathcal{L} = \frac{\partial\mathcal{L}}{\partial\mathbf{Q}}\mathbf{Q}^\top - \mathbf{Q}\frac{\partial\mathcal{L}}{\partial\mathbf{Q}}^\top \tag{15}$$

where $\exp(\cdot)$ is the matrix exponential. To make the loss function differentiable, we incorporate the non-negativity constraint with a soft-minimum:

$$\mathcal{L} = \sum_{i\mu} \left(\mathbf{q}_i^\top \mathbf{r}^\mu - \text{softmin}_\nu(\mathbf{q}_i^\top \mathbf{r}^\nu, \beta)\right)$$

$$\text{softmin}(a_1, a_2, ..., a_P; \beta) = \frac{1}{Z}\sum_{\mu=1}^{P} a_\nu \exp(-\beta a_\nu), \tag{16}$$

where $Z = \sum_\nu \exp(-\beta a_\nu)$ is a normalizing constant and $\mathbf{Q} = [\mathbf{q}_1, ...\mathbf{q}_N]$. In the $\beta \to \infty$ limit, this cost function converges to the exact optimization problem with non-negativity constraint. Finite β, however, allows learning with gradient descent. Gradients are computed with automatic differentiation in JAX [102]. This optimization routine is run until convergence and the optimal value is plotted as dashed red lines labeled "opt." in Figure 3.

We show that our result is robust to different pre-processing techniques and to imposing bounds on neural firing rates in the Figure App.1. To demonstrate that our result is not an artifact of z-scoring the deconvolved signals against the spontaneous baseline activity level, we also conduct the random rotation experiment on the raw deconvolved signals. In addition, we show that imposing realistic constraints on the upper bound of the each neuron's responses does not change our findings. We used a subset of $N = 100$ neurons and computed random rotations. However, we only accepted

a code as valid if it's maximum value was less than some upper bound u_b. Subsets of $N = 100$ neurons in the biological code achieve maxima in the range between 3.2 and 4.7. We performed this experiment for $u_b \in \{3, 4, 5\}$ so that the artificial codes would have maxima that lie in the same range as the biological code.

4.4.4 Lifetime and Population Sparseness

We compute two more refined measures of sparseness in a population code. For each neuron i we compute the lifetime sparseness LS_i (also known as selectivity) and for each stimulus $\boldsymbol{\theta}$ we compute the population sparseness $PS_{\boldsymbol{\theta}}$ which are defined as the following two ratios [22, 23, 24, 25]

$$LS_i = \frac{1}{1 - \frac{1}{P}} \frac{\mathrm{Var}_{\boldsymbol{\theta}}\, r_i(\boldsymbol{\theta})}{\langle r_i(\boldsymbol{\theta})^2 \rangle_{\boldsymbol{\theta}}} \ , \ PS_{\boldsymbol{\theta}} = \frac{1}{1 - \frac{1}{N}} \frac{\mathrm{Var}_i\, r_i(\boldsymbol{\theta})}{\langle r_i(\boldsymbol{\theta})^2 \rangle_i} \tag{17}$$

The normalization factors ensure that these quantities lie in the interval between $(0, 1)$. Intuitively, lifetime sparseness quantifies the variability of each neuron's responses over the full set of stimuli, whereas population sparseness quantifies the variability of responses in the code for a given stimulus $\boldsymbol{\theta}$.

4.4.5 Fitting a Gabor model to mouse V1 kernel

Under the assumption of translation symmetry in the kernel $K(\theta, \theta')$, we averaged the elements of the over rows of the empirical mouse V1 kernel [9]

$$K(\Delta) = \frac{1}{P} \sum_{\mu=1}^{P} K(\theta^\mu, \theta^\mu + \Delta) \tag{18}$$

where angular addition is taken mod π. This generates the black dots in Figure 6 B. We aimed to fit a threshold-power law nonlinearity of the form $g_{q,a}(z) = \max\{0, z - a\}^q$ to the kernel. Based on the Gabor model discussed above, we parameterized tuning curves as

$$r_{\sigma^2,q,a}(\theta, \theta_i) = g_{q,a}\left(\frac{\cosh(\sigma^{-2}\cos(\theta - \theta_i))}{\cosh(\sigma^{-2})}\right), \tag{19}$$

where θ_i is the preferred angle of the i-th neuron's tuning curve. Rather than attempting to perform a fit of $\sigma^2, a, q, \{\theta_i\}_{i=1}^N$ of this form to the responses of each of the $\sim$ 20-k neurons, we instead simply attempt to fit to the population kernel by optimizing over (σ^2, a, q) under the assumption of uniform tiling of θ_i. However, we noticed that two of these variables σ^2, a are constrained by the sparsity level of the code. If each neuron, on average, fires for only a fraction f of the uniformly sampled angles θ, then the following relationship holds between σ^2 and a

$$a = \frac{\cosh\left(\sigma^{-2}\cos\left(\frac{\pi}{2}f\right)\right)}{\cosh(\sigma^{-2})}. \tag{20}$$

Calculation of the coding level f for the recorded responses allowed us to infer a from β during optimization. This reduced the free parameter set to (σ^2, q). We then solve the following optimization problem

$$\min_{\sigma^2,q} \left\langle \left(\hat{K}_{\sigma^2,q}(\theta) - K(\theta) \right)^2 \right\rangle_\theta \quad \hat{K}_{\sigma^2,q}(\theta) = \left\langle r_{\sigma^2,q}(\theta,\theta') r_{\sigma^2,q}(0,\theta') \right\rangle_{\theta'}, \tag{21}$$

where integration over θ_i is performed numerically. Using the Scipy Trust-Region constrained optimizer, we found $(q, \sigma^{-2}, a) = (1.7, 5.0, 0.2)$ which we use as the fit parameters in Figure 6.

Lead contact

Requests for information should be directed to the lead contact, Cengiz Pehlevan (`cpehlevan@seas.harvard.edu`).

Data and code availability

Mouse V1 neuron responses to orientation gratings and preprocessing code were obtained from a publicly available dataset: `https://github.com/MouseLand/stringer-et-al-2019`, [8, 9].

Responses to ImageNet images and preprocessing code were obtained from another publicly available dataset, `https://github.com/MouseLand/stringer-pachitariu-et-al-2018b` [10, 11].

The code generated by the authors for this paper is also available `https://github.com/Pehlevan-Group/sample_efficient_pop_codes`

Acknowledgements

We would like to thank Jacob Zavatone-Veth and Abdulkadir Canatar for useful comments and discussions about this manuscript. BB acknowledges the support of the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award #1764269) and the Harvard Q-Bio Initiative. CP and BB would also like to acknowledge support from the Harvard Data Science Initiative.

Competing Interests

The authors have no competing interests to declare.

References

[1] Susan Carey and Elsa Bartlett. Acquiring a single new word., 1978.

[2] Matthew F. Peterson, Craig K. Abbey, and Miguel P. Eckstein. The surprisingly high human efficiency at learning to recognize faces. *Vision Research*, 49(3):301–314, 2009.

[3] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

[4] Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017.

[5] Fabian H Sinz, Xaq Pitkow, Jacob Reimer, Matthias Bethge, and Andreas S Tolias. Engineering a less artificial intelligence. *Neuron*, 103(6):967–979, 2019.

[6] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[7] David Hume. *An enquiry concerning human understanding : a critical edition*. Hume, David, 1711-1776. Works. 1998. Clarendon Press ; Oxford University Press, Oxford : New York, 1998.

[8] Carsen Stringer, Michalis Michaelos, Dmitri Tsyboulski, Sarah E. Lindo, and Marius Pachitariu. High-precision coding in visual cortex. *Cell*, 0092-8674, 5 2021.

[9] Marius Pachitariu, Michalis Michaelos, and Carsen Stringer. Recordings of 20,000 neurons from V1 in response to oriented stimuli, 11 2019.

[10] Carsen Stringer, Marius Pachitariu, Matteo Carandini, and Kenneth Harris. Recordings of 10,000 neurons in visual cortex in response to 2,800 natural images, Jul 2018.

[11] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571, 6 2018.

[12] M. Radford Neal. Bayesian learning for neural networks, 1994.

[13] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.

[14] H. Barlow. Possible principles underlying the transformation of sensory messages, 1961.

[15] J. J. Atick and A. N. Redlich. What does the retina know about natural scenes? *Neural Computation*, 4(2):196–210, March 1992.

[16] Fred Attneave. Some informational aspects of visual perception. *Psychological review*, 61(3):183, 1954.

[17] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311 – 3325, 1997.

[18] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24(1):1193–1216, 2001. PMID: 11520932.

[19] Marius Pachitariu, Carsen Stringer, and Kenneth D Harris. Robustness of spike deconvolution for neuronal calcium imaging. *Journal of Neuroscience*, 38(37):7976–7985, 2018.

[20] T. Anderson, I. Olkin, and Leslie Underhill. Generation of random orthogonal matrices. *Siam Journal on Scientific and Statistical Computing*, 8, 07 1987.

[21] Saskia E. J. de Vries, Jérôme A. Lecoq, Michael A. Buice, Peter A. Groblewski, Gabriel Koch Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel J. Millman, Kate Roll, Marina Garrett, Tom Keenan, Leonard Kuan, Stefan Mihalas, Shawn R. Olsen, Carol L. Thompson, Wayne Wakeman, Jack Waters, Derric Williams, Chris Barber, Nathan Berbesque, Brandon Blanchard, Nicholas Bowles, Shiella Caldejon, Linzy Casal, Andrew Cho, Sissy Cross, Chinh Dang, Tim Dolbeare, Mélise Edwards, J. Galbraith, Nathalie Gaudreault, Terri L. Gilbert, Fiona Griffin, Perry Hargrave, Robert E. Howard, Lawrence Huang, Sean Jewell, Nika Keller, Ulf Knoblich, Josh D. Larkin, Rachael Larsen, Christopher Lau, Eric Lee, Felix Lee, Arielle Leon, Lu Li, Fuhui Long, Jennifer A. Luviano, Kyla Mace, Thuyanh V. Nguyen, Jed Perkins, Miranda Robertson, Sam Seid, Eric T. Shea-Brown, Jianghong Shi, Nathan Sjoquist, Cliff Slaughterbeck, David Sullivan, Ryan A. Valenza, Casey White, A. Michael Williford, Daniela M. Witten, Jun Zhuang, Hongkui Zeng, Colin Farrell, Lydia Ng, Amy Bernard, John W. Phillips, R. Clay Reid, and Christof Koch. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, 23:138 – 151, 2019.

[22] Benjamin Willmore and David J Tolhurst. Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems*, 12(3):255, 2001.

[23] Sidney R Lehky, Terrence J Sejnowski, and Robert Desimone. Selectivity and sparseness in the responses of striate complex cells. *Vision research*, 45(1):57–73, 2005.

[24] Alessandro Treves and Edmund T Rolls. What determines the capacity of autoassociative memories in the brain? *Network: Computation in Neural Systems*, 2(4):371, 1991.

[25] Cengiz Pehlevan and Haim Sompolinsky. Selectivity and sparseness in randomly connected balanced networks. *PLOS ONE*, 9(2):1–15, 02 2014.

[26] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020.

[27] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, in press, 2021.

[28] Mark Stopfer, Vivek Jayaraman, and Gilles Laurent. Intensity versus identity coding in an olfactory system. *Neuron*, 39(6):991–1004, 2003.

[29] Saul Kato, Harris S Kaplan, Tina Schrödel, Susanne Skora, Theodore H Lindsay, Eviatar Yemini, Shawn Lockery, and Manuel Zimmer. Global brain dynamics embed the motor command sequence of caenorhabditis elegans. *Cell*, 163(3):656–669, 2015.

[30] Brice Bathellier, Derek L. Buhl, Riccardo Accolla, and Alan Carleton. Dynamic ensemble odor coding in the mammalian olfactory bulb: Sensory information at different timescales. *Neuron*, 57(4):586 – 598, 2008.

[31] Juan A. Gallego, Matthew G. Perich, Lee E. Miller, and Sara A. Solla. Neural manifolds for the control of movement. *Neuron*, 94(5):978 – 984, 2017.

[32] Juan A Gallego, Matthew G Perich, Stephanie N Naufel, Christian Ethier, Sara A Solla, and Lee E Miller. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature communications*, 9(1):1–13, 2018.

[33] T. Patrick Sadtler, M. Kristen Quick, D. Matthew Golub, M. Steven Chase, I. Steven Ryu, C. Elizabeth Tyler-Kabara, M. Byron Yu, and P. Aaron Batista. Neural constraints on learning. *Nature*, 512(5500):423–426, 2014.

[34] Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and L.F. Abbott. Optimal degrees of synaptic connectivity, 2017.

[35] Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement, 2017.

[36] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current Opinion in Neurobiology*, 32:148 – 155, 2015. Large-Scale Recording Technology (32).

[37] James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

[38] Cristopher M. Niell and Michael P. Stryker. Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience*, 28(30):7520–7536, 2008.

[39] Vincent Bonin, Mark H. Histed, Sergey Yurgenson, and R. Clay Reid. Local diversity and fine-scale organization of receptive fields in mouse visual cortex. *Journal of Neuroscience*, 31(50):18506–18521, 2011.

[40] Samme Vreysen, Bin Zhang, Yuzo M. Chino, Lutgarde Arckens, and Gert Van den Bergh. Dynamics of spatial frequency tuning in mouse visual cortex. *Journal of Neurophysiology*, 107(11):2937–2949, 2012. PMID: 22402662.

[41] D Hansel and C Van Vreeswijk. How noise contributes to contrast invariance of orientation tuning in cat visual cortex. *Journal of Neuroscience*, 22(12):5118–5128, 2002.

[42] Kenneth D Miller and Todd W Troyer. Neural noise can explain expansive, power-law nonlinearities in neural response functions. *Journal of neurophysiology*, 87(2):653–659, 2002.

[43] Nicholas J Priebe, Ferenc Mechler, Matteo Carandini, and David Ferster. The contribution of spike threshold to the dichotomy of cortical simple and complex cells. *Nature neuroscience*, 7(10):1113–1122, 2004.

[44] Nicholas J Priebe and David Ferster. Inhibition, spike threshold, and stimulus selectivity in primary visual cortex. *Neuron*, 57(4):482–497, 2008.

[45] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A, Optics and image science*, 2 2:284–99, 1985.

[46] Eero P. Simoncelli and David J. Heeger. A model of neuronal responses in visual area mt. *Vision Research*, 38(5):743–761, 1998.

[47] Jonathan Dong, Ruben Ohana, Mushegh Rafayelyan, and Florent Krzakala. Reservoir computing meets recurrent kernels and structured transforms, 2020.

[48] Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes, 2019.

[49] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture, 2020.

[50] K Cora Ames, Stephen I Ryu, and Krishna V Shenoy. Simultaneous motor preparation and execution in a last-moment reach correction task. *Nature communications*, 10(1):1–13, 2019.

[51] H. Sompolinsky, A. Crisanti, and H. J. Sommers. Chaos in random neural networks. *Phys. Rev. Lett.*, 61:259–262, Jul 1988.

[52] Yu BM Cunningham JP. Dimensionality reduction for large-scale neural recordings, 2014.

[53] Peiran Gao and Surya Ganguli. On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155, 2015.

[54] John K Chapin and Miguel A.L Nicolelis. Principal component analysis of neuronal ensemble activity reveals multidimensional somatosensory representations. *Journal of Neuroscience Methods*, 94(1):121 – 140, 1999.

[55] Anthony M Zador. A critique of pure learning and what artificial neural networks can learn from animal brains. *Nature communications*, 10(1):1–7, 2019.

[56] S. Edelman. Representation is representation of similarities. *The Behavioral and brain sciences*, 21 4:449–67; discussion 467–98, 1998.

[57] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008.

[58] Aarre Laakso. Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13, 05 2000.

[59] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019.

[60] Charles F. Cadieu, Ha Hong, Daniel L. K. Yamins, Nicolas Pinto, Diego Ardila, Ethan A. Solomon, Najib J. Majaj, and James J. DiCarlo. Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLOS Computational Biology*, 10(12):1–18, 12 2014.

[61] Cengiz Pehlevan, Anirvan M. Sengupta, and Dmitri B. Chklovskii. Why do similarity matching objectives lead to hebbian/anti-hebbian networks? *Neural Computation*, 30(1):84–124, 2018. PMID: 28957017.

[62] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019.

[63] Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks, 2019.

[64] Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin L Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. {SGD} on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, 2019.

[65] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580. Curran Associates, Inc., 2018.

[66] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Learning with invariances in random features and kernel models, 2021.

[67] Zhiyuan Li, Ruosong Wang, Dingli Yu, Simon S. Du, Wei Hu, Ruslan Salakhutdinov, and Sanjeev Arora. Enhanced convolutional neural tangent kernels, 2019.

[68] Uri Hasson, Samuel A Nastase, and Ariel Goldstein. Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434, 2020.

[69] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[70] C Zhang, S Bengio, M Hardt, B Recht, and O Vinyals. Understanding deep learning requires rethinking generalization. In *5th Int. Conf. on Learning Representations (ICLR 2017)*, 2016.

[71] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *International Conference on Learning Representations*, 2018.

[72] Florian Meier, Raphaël Dang-Nhu, and Angelika Steger. Adaptive tuning curve widths improve sample efficient learning. *Frontiers in Computational Neuroscience*, 14, 2020.

[73] Kameron Decker Harris. Additive function approximation in the brain, 2019.

[74] Mandyam Veerambudi Srinivasan, Simon Barry Laughlin, and Andreas Dubs. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216(1205):427–459, 1982.

[75] Johannes H van Hateren. A theory of maximizing sensory information. *Biological cybernetics*, 68(1):23–29, 1992.

[76] Rajesh PN Rao and Dana H Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87, 1999.

[77] Matthew Chalk, Olivier Marre, and Gasper Tkacik. Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115(1):186–191, 2018.

[78] Xaq Pitkow and Markus Meister. Decorrelation and efficient coding by retinal ganglion cells. *Nature neuroscience*, 15(4):628–635, 2012.

[79] Reza Abbasi-Asl, Cengiz Pehlevan, Bin Yu, and Dmitri Chklovskii. Do retinal ganglion cells project natural scenes to their principal subspace and whiten them? In *2016 50th Asilomar Conference on Signals, Systems and Computers*, pages 1641–1645. IEEE, 2016.

[80] J. Gallego, M. Perich, S. Naufel, C. Ethier, S. Solla, and L. Miller. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature Communications*, 9, 10 2018.

[81] M Haft and J Leo van Hemmen. Theory and implementation of infomax filters for the retina. *Network: Computation in Neural Systems*, 9(1):39–71, 1998.

[82] Yanping Huang and Rajesh PN Rao. Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(5):580–593, 2011.

[83] J Nassar, P Sokol, S Chang, and K Harris. On 1/n neural representation and robustness, 2020.

[84] Jeremy E. Niven and Simon B. Laughlin. Energy limitation as a selective pressure on the evolution of sensory systems. *Journal of Experimental Biology*, 211(11):1792–1804, 2008.

[85] Tomas Hromadka, Michael R DeWeese, and Anthony M Zador. Sparse representation of sounds in the unanesthetized auditory cortex. *PLOS Biology*, 6(1):1–14, 01 2008.

[86] Charles D Gilbert. Early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, 91(4):1195, 1994.

[87] Pieter M Goltstein, Sandra Reinert, Tobias Bonhoeffer, and Mark Hübener. Mouse visual cortex areas represent perceptual and semantic features of learned visual categories. *Nature neuroscience*, 24(10):1441–1451, 2021.

[88] Geoffrey M Ghose, Tianming Yang, and John HR Maunsell. Physiological correlates of perceptual learning in monkey v1 and v2. *Journal of neurophysiology*, 87(4):1867–1888, 2002.

[89] Aniek Schoups, Rufin Vogels, Ning Qian, and Guy Orban. Practising orientation identification improves orientation coding in v1 neurons. *Nature*, 412(6846):549–553, 2001.

[90] Tianming Yang and John HR Maunsell. The effect of perceptual learning on neuronal responses in monkey visual area v4. *Journal of Neuroscience*, 24(7):1617–1626, 2004.

[91] Hamed Zivari Adab, Ivo D Popivanov, Wim Vanduffel, and Rufin Vogels. Perceptual learning of simple stimuli modifies stimulus representations in posterior inferior temporal cortex. *Journal of cognitive neuroscience*, 26(10):2187–2200, 2014.

[92] HP Op de Beeck, Johan Wagemans, and Rufin Vogels. Effects of perceptual learning in visual backward masking on the responses of macaque inferior temporal neurons. *Neuroscience*, 145(2):775–789, 2007.

[93] Chi-Tat Law and Joshua I Gold. Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nature neuroscience*, 11(4):505–513, 2008.

[94] Bruno Averbeck, Peter Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7, 2006.

[95] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature neuroscience*, 14(7):811, 2011.

[96] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[97] Peter Sollich. Learning curves for gaussian processes. In *Neurips*, 1998.

[98] Peter Sollich. Gaussian process regression with mismatched models. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 519–526. MIT Press, 2002.

[99] Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. *CoRR*, abs/2102.08127, 2021.

[100] James B. Simon, Madeline Dickens, and Michael R. DeWeese. Neural tangent kernel eigenvalues accurately predict generalization, 2021.

[101] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[102] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

Appendix: Population codes enable learning from few examples by shaping inductive bias

1 Neural Code Decomposition

1.1 Singular Value Decomposition of Continuous Population Responses

SVD of population responses is usually evaluated with respect to a discrete and finite set of stimuli. In the main paper, we implicitly assumed that a generalization of SVD to a continuum of stimuli. In this section we provide an explicit construction of this generalized SVD using techniques from functional analysis. Our construction is an example of the quasimatrix SVD defined in [1] and justifies our use of SVD in the main text.

For our construction, we note that Mercer's theorem guarantees the existence of an eigendecomposition of any inner product kernel $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ in terms of a complete orthonormal set of functions $\{\psi_k\}_{k=1}^{\infty}$ [2]. In particular, there exist a non-negative (but possibly zero) summable eigenvalues $\{\lambda_k\}_{k=1}^{\infty}$ and a corresponding set of orthonormal eigenfunctions such that

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{k=1}^{\infty} \lambda_k \psi_k(\boldsymbol{\theta}) \psi_k(\boldsymbol{\theta}'). \tag{App.1}$$

For a stimulus distribution $p(\boldsymbol{\theta})$, the set of functions $\{\psi_k\}_{k=1}^{\infty}$ are orthonormal and form a complete basis for square integrable functions L_2 which means

$$\begin{aligned} \left\langle \psi_k(\boldsymbol{\theta}) \psi_\ell(\boldsymbol{\theta}) \right\rangle_{\boldsymbol{\theta}} &= \int p(\boldsymbol{\theta}) \psi_k(\boldsymbol{\theta}) \psi_\ell(\boldsymbol{\theta}) d\boldsymbol{\theta} = \delta_{k\ell}, \\ f(\boldsymbol{\theta}) &= \sum_k \left\langle f(\boldsymbol{\theta}') \psi_k(\boldsymbol{\theta}') \right\rangle_{\boldsymbol{\theta}'} \psi_k(\boldsymbol{\theta}) \ , \ \forall f \in L_2. \end{aligned} \tag{App.2}$$

Next, we use this basis to construct the SVD. Each of the tuning curves $r_i \in L_2$ (assumed to be square integrable) can be expressed in this basis with the top N of the functions in the set $\{\psi_k\}_{k=1}^{\infty}$

$$r_i(\boldsymbol{\theta}) = \sum_{k=1}^{N} A_{ik} \psi_k(\boldsymbol{\theta}), \tag{App.3}$$

where we introduced a matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ of expansion coefficients. Note that $\text{rank}(\mathbf{A}) \leq N$. We

compute the singular value decomposition of the finite matrix $\mathbf{A}$

$$\mathbf{A} = \sqrt{N} \sum_{k=1}^{\text{rank}(\mathbf{A})} \sqrt{\lambda_k} \mathbf{u}_k \mathbf{v}_k^\top. \tag{App.4}$$

We note that the signal correlation matrix for this population code can be computed in closed form

$$\mathbf{\Sigma}_s = \frac{1}{N} \mathbf{A} \left\langle \psi(\boldsymbol{\theta}) \psi(\boldsymbol{\theta})^\top \right\rangle_{\boldsymbol{\theta}} \mathbf{A}^\top = \frac{1}{N} \mathbf{A}\mathbf{A}^\top = \sum_{k=1}^{\text{rank}(\mathbf{A})} \lambda_k \mathbf{u}_k \mathbf{u}_k^\top, \tag{App.5}$$

due to the orthonormality of $\{\psi_k\}$. Thus the principal axes $\mathbf{u}_k$ of the neural correlations are the left singular vectors of $\mathbf{A}$. We may similarly express the inner product kernel in terms of the eigenfunctions

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \frac{1}{N} \mathbf{r}(\boldsymbol{\theta}) \cdot \mathbf{r}(\boldsymbol{\theta}') = \frac{1}{N} \psi(\boldsymbol{\theta})^\top \mathbf{A}^\top \mathbf{A} \psi(\boldsymbol{\theta}'). \tag{App.6}$$

The kernel eigenvalue problem demands [2]

$$\begin{aligned} &\int p(\boldsymbol{\theta}) K(\boldsymbol{\theta}, \boldsymbol{\theta}') \psi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{N} \mathbf{A}^\top \mathbf{A} \psi(\boldsymbol{\theta}') = \mathbf{\Lambda} \psi(\boldsymbol{\theta}') \implies \frac{1}{N} \mathbf{A}^\top \mathbf{A} = \mathbf{\Lambda} \\ &\implies \sum_{k=1}^{\text{rank}(\mathbf{A})} \lambda_k \mathbf{v}_k \mathbf{v}_k^\top = \sum_{k=1}^{\text{rank}(\mathbf{A})} \lambda_k \mathbf{e}_k \mathbf{e}_k^\top. \end{aligned} \tag{App.7}$$

The $\mathbf{v}_k$ vectors must be identical to $\pm \mathbf{e}_k$, the Cartesian unit vectors, if the eigenvalues are non-degenerate. From this exercise, we find that the SVD for $\mathbf{A}$ has the form $\mathbf{A} = \sqrt{N} \sum_{k=1}^{\text{rank}(\mathbf{A})} \sqrt{\lambda_k} \mathbf{u}_k \mathbf{e}_k^\top$. With this choice, the population code admits a singular value decomposition

$$\mathbf{r}(\boldsymbol{\theta}) = \mathbf{A} \psi(\boldsymbol{\theta}) = \sqrt{N} \sum_{k=1}^{\text{rank}(\mathbf{A})} \sqrt{\lambda_k} \mathbf{u}_k \psi_k(\boldsymbol{\theta}). \tag{App.8}$$

This singular value decomposition demonstrates the connection between neural manifold structure (principal axes $\mathbf{u}_k$) and function approximation (kernel eigenfunctions ψ_k). This singular value decomposition can be verified by computing the inner product kernel and the correlation matrix, utilizing the orthonormality of $\{\mathbf{u}_k\}$ and $\{\psi_k\}$. This exercise has important consequences for the space of learnable functions, which is at most $\text{rank}(\mathbf{A})$ dimensional since linear readouts lie in $\text{span}\{r_i(\boldsymbol{\theta})\}_{i=1}^N$.

1.2 Discrete Stimulus Spaces: Finding Eigenfunctions with Matrix Eigendecomposition

In our discussion so far, our notation suggested that $\boldsymbol{\theta}$ take a continuum of values. Here we want to point that our theory still applies if $\boldsymbol{\theta}$ take a discrete set of values. In this case, we can think of a Dirac measure $p(\boldsymbol{\theta}) = \sum_{i=1}^{\tilde{P}} p_i \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^i)$, where i indexes all the $\tilde{P}$ values $\boldsymbol{\theta}$ can take. With this

choice

$$\int p(\boldsymbol{\theta})K(\boldsymbol{\theta},\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta})d\boldsymbol{\theta} = \sum_{i=1}^{\tilde{P}} p_i K(\boldsymbol{\theta}^i,\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta}^i) = \lambda_k \psi_k(\boldsymbol{\theta}'). \tag{App.9}$$

Demanding this equality for $\boldsymbol{\theta}' = \boldsymbol{\theta}^i,\ i = 1, ..., \tilde{P}$ generates a matrix eigenvalue problem

$$\mathbf{KB\Psi} = \mathbf{\Psi\Lambda}, \tag{App.10}$$

where $\mathbf{B}_{ij} = \delta_{ij} p_i$. The eigenfunctions over the stimuli are identified as the columns of $\mathbf{\Psi}$ while the eigenvalues are the diagonal elements of $\mathbf{\Lambda}_{k\ell} = \lambda_k \delta_{k\ell}$.

Experimental considerations: In an experimental setting, a finite number of stimuli are presented and the SVD is calculated over this finite set regardless of the support of $p(\boldsymbol{\theta})$. This raises the question of the interpretation of this SVD and its relation to the inductive bias theory we presented. Here we provide two interpretations.

In the first interpretation, we think of the empirical SVD as providing an estimate of the SVD over the full distribution $p(\boldsymbol{\theta})$. To formalize this notion, we can introduce a Monte-Carlo estimate of the integral eigenvalue problem

$$\int p(\boldsymbol{\theta})K(\boldsymbol{\theta},\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta})d\boldsymbol{\theta} \approx \frac{1}{\tilde{P}} \sum_{\mu=1}^{\tilde{P}} K(\boldsymbol{\theta}^\mu,\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta}^\mu) = \lambda_k \psi_k(\boldsymbol{\theta}'). \tag{App.11}$$

For this interpretation to work, the experimenter must sample the stimuli from $p(\boldsymbol{\theta})$, which could be the natural stimulus distribution. Measuring responses to a larger number of stimuli gives a more accurate approximation of the integral above, which will provide a better estimate of generalization performance on the true distribution $p(\boldsymbol{\theta})$.

In the second interpretation, we construct an empirical measure on $\tilde{P}$ experimental stimulus values $\hat{p}(\boldsymbol{\theta}) = \frac{1}{\tilde{P}} \sum_{\mu=1}^{\tilde{P}} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}^\mu)$, and consider learning and generalization over this distribution. This allows the application of our theory to an experimental setting where $\hat{p}(\boldsymbol{\theta})$ is designed by an experimenter. For example, the experimenter could procure a complicated set of $\tilde{P}$ videos, to which an associated function $y(\theta)$ must be learned. After showing these videos to the animal and measuring neural responses, the experimenter could compute, with our theory, generalization error for a uniform distribution over this full set of $\tilde{P}$ videos. Our theory would predict generalization over this distribution after providing supervisory feedback for only a strict subset of $P < \tilde{P}$ videos. Under this interpretation, the relationship between the integral eigenvalue problem and matrix eigenvalue problem is exact rather than approximate

$$\int \hat{p}(\boldsymbol{\theta})K(\boldsymbol{\theta},\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{\tilde{P}} \sum_{\mu=1}^{\tilde{P}} K(\boldsymbol{\theta}^\mu,\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta}^\mu) = \lambda_k \psi_k(\boldsymbol{\theta}'). \tag{App.12}$$

Demanding either of (App.11) or (App.12) equalities for $\boldsymbol{\theta}' = \boldsymbol{\theta}^\nu,\ \nu = 1, ..., P$ generates a matrix

eigenvalue problem

$$\mathbf{K}\mathbf{\Psi} = P\mathbf{\Psi}\mathbf{\Lambda}. \tag{App.13}$$

The eigenfunctions restricted to $\{\boldsymbol{\theta}^\mu\}$ are identified as the columns of $\mathbf{\Psi}$ while the eigenvalues are the diagonal elements of $\mathbf{\Lambda}_{k\ell} = \lambda_k \delta_{k\ell}$. For the case where N and P are finite, the spectrum obtained through eigendecomposition of the kernel $\mathbf{K}$ is the same as would be obtained through the finite N signal correlation matrix $\mathbf{\Sigma}_s$, since they are inner and outer products of trial averaged population response matrices $\mathbf{R}$.

1.3 Translation Invariant Kernels

For the special case where the data distribution $p(\boldsymbol{\theta}) = \frac{1}{V}$ is uniform over volume V and the kernel is translation invariant $K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \kappa(\boldsymbol{\theta} - \boldsymbol{\theta}')$, the kernel can be diagonalized in the basis of plane waves

$$\int p(\boldsymbol{\theta}) K(\boldsymbol{\theta}, \boldsymbol{\theta}') \psi_{\mathbf{k}}(\boldsymbol{\theta}) d\boldsymbol{\theta} = \frac{1}{V} \int \kappa(\boldsymbol{\theta} - \boldsymbol{\theta}') e^{i\mathbf{k}\cdot\boldsymbol{\theta}} d\boldsymbol{\theta} = \frac{1}{V} \hat{\kappa}(\mathbf{k}) e^{i\mathbf{k}\cdot\boldsymbol{\theta}'} \tag{App.14}$$

The eigenvalues are the Fourier components of the Kernel $\lambda_{\mathbf{k}} = \frac{1}{V}\hat{\kappa}(\mathbf{k}) = \frac{1}{V}\int d\boldsymbol{\theta} e^{i\mathbf{k}\cdot\boldsymbol{\theta}} \kappa(\boldsymbol{\theta})$ while the eigenfunctions are plane waves $\psi_{\mathbf{k}}(\boldsymbol{\theta}) = e^{i\mathbf{k}\cdot\boldsymbol{\theta}}$. The set of admissible momenta $\mathcal{S}_{\mathbf{k}} = \{\mathbf{k}_0, \pm\mathbf{k}_1, \pm\mathbf{k}_2, ...\}$ are determined by the boundary conditions. The diagonalized representation of the kernel is therefore

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \sum_{\mathbf{k}\in\mathcal{S}_{\mathbf{k}}} \lambda_{\mathbf{k}} e^{i\mathbf{k}\cdot(\boldsymbol{\theta}-\boldsymbol{\theta}')} \tag{App.15}$$

For example, if the space is the torus $\mathbb{T}^n = S^1 \times S^1 \times ... \times S^1$, then the space of admissable momenta are the points on the integer lattice $\mathcal{S}_{\mathbf{k}} = \mathbb{Z}^n = \{\mathbf{k} \in \mathbb{R}^n | k_i \in \mathbb{Z}\ \forall i = 1, ..., n\}$. Reality and symmetry of the kernel demand that $\mathrm{Im}(\lambda_{\mathbf{k}}) = 0$ and $\lambda_{-\mathbf{k}} = \lambda_{\mathbf{k}} \geq 0$. Most of the models in this paper consider $\theta \sim \mathrm{Unif}\left(S^1\right)$, where the kernel has the following Fourier/Mercer decomposition

$$\begin{aligned} K(\theta - \theta') &= \sum_{k=-\infty}^{\infty} \lambda_k e^{ik(\theta-\theta')} = 2\sum_{k=0}^{\infty} \lambda_k \cos\big(k(\theta - \theta')\big) \\ &= \sum_{k=0}^{\infty} \lambda_k \left[\sqrt{2}\cos(k\theta)\sqrt{2}\cos\big(k\theta'\big) + \sqrt{2}\sin(k\theta)\sqrt{2}\sin\big(k\theta'\big)\right] \end{aligned} \tag{App.16}$$

where we invoked the simple trigonometric identity $\cos(a - b) = \cos(a)\cos(b) + \sin(a)\sin(b)$. By recognizing that $\{\sqrt{2}\cos(k\theta), \sqrt{2}\sin(k\theta)\}_{k=0}^{\infty}$ form a complete orthonormal set of functions with respect to $\mathrm{Unif}\left(S^1\right)$, we have identified this as the collection of kernel eigenfunctions.

1.4 Invariant Kernels Possess Invariant Eigenfunctions

Suppose the kernel $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ is invariant to some set of transformations $t \in \mathcal{T}$, by which we mean that

$$K(t\boldsymbol{\theta}, \boldsymbol{\theta}') = K(\boldsymbol{\theta}, t\boldsymbol{\theta}') = K(\boldsymbol{\theta}, \boldsymbol{\theta}')\ ,\ \forall t \in \mathcal{T} \tag{App.17}$$

We will now show that any eigenfunction of such a kernel with nonzero eigenvalue must be an invariant function. Let $\psi_k(\boldsymbol{\theta})$ be an eigenfunction with eigenvalue $\lambda_k > 0$, then

$$\psi_k(t\boldsymbol{\theta}) = \frac{1}{\lambda_k}\int p(\boldsymbol{\theta}')K(\boldsymbol{\theta}', t\boldsymbol{\theta})d\boldsymbol{\theta}' = \frac{1}{\lambda_k}\int p(\boldsymbol{\theta}')K(\boldsymbol{\theta}', \boldsymbol{\theta})d\boldsymbol{\theta}' = \psi_k(\boldsymbol{\theta}) \tag{App.18}$$

This establishes that all functions which depend on $\mathcal{T}$ transformations must necessarily lie in the null-space of K.

2 Alternative Neural Codes with Same Kernel

2.1 Orthogonal Transformations are Sufficient for Linear Kernel-Preserving Transformations

We will now show that for any linear transformation $\tilde{\mathbf{r}} = \mathbf{A}\mathbf{r}$ which preserves the inner product kernel $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$, there exists an orthogonal matrix $\mathbf{Q}$ such that $\tilde{\mathbf{r}} = \mathbf{Q}\mathbf{r}$.

Proof. Let $\tilde{\mathbf{r}}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{r}(\boldsymbol{\theta})$ for all stimuli $\boldsymbol{\theta}$. To preserve the kernel, we must have

$$K(\boldsymbol{\theta}, \boldsymbol{\theta}') = \tilde{\mathbf{r}}(\boldsymbol{\theta}) \cdot \tilde{\mathbf{r}}(\boldsymbol{\theta}') = \mathbf{r}(\boldsymbol{\theta}) \cdot \mathbf{r}(\boldsymbol{\theta}') \implies \mathbf{r}(\boldsymbol{\theta})\mathbf{A}^\top\mathbf{A}\mathbf{r}(\boldsymbol{\theta}') = \mathbf{r}(\boldsymbol{\theta}) \cdot \mathbf{r}(\boldsymbol{\theta}'). \tag{App.19}$$

Taking projections against each of the orthonormal eigenfunctions $\psi_\ell(\theta)$ (see Appendix 1A), we define vectors $\mathbf{u}_k$ as $\sqrt{\lambda_k}\mathbf{u}_k = \left\langle \mathbf{r}(\boldsymbol{\theta})\psi_k(\boldsymbol{\theta})\right\rangle_{\boldsymbol{\theta}}$, allowing us to express the SVD of the population code $\mathbf{r}(\boldsymbol{\theta}) = \sum_k \sqrt{\lambda_k}\mathbf{u}_k\psi_k(\boldsymbol{\theta})$. These vectors $\{\mathbf{u}_k\}$ are orthonormal $\mathbf{u}_k \cdot \mathbf{u}_\ell = \delta_{k\ell}$ since, by the definition of the kernel eigenfunctions ψ_k,

$$\begin{aligned}\sqrt{\lambda_k\lambda_\ell}\mathbf{u}_k \cdot \mathbf{u}_\ell &= \left\langle \mathbf{r}(\boldsymbol{\theta}) \cdot \mathbf{r}(\boldsymbol{\theta}')\psi_k(\boldsymbol{\theta})\psi_\ell(\boldsymbol{\theta}')\right\rangle_{\boldsymbol{\theta},\boldsymbol{\theta}'} \\ &= \left\langle \psi_k(\boldsymbol{\theta})\left\langle K(\boldsymbol{\theta}, \boldsymbol{\theta}')\psi_\ell(\boldsymbol{\theta}')\right\rangle_{\boldsymbol{\theta}'}\right\rangle_{\boldsymbol{\theta}} \\ &= \lambda_\ell\left\langle \psi_k(\boldsymbol{\theta})\psi_\ell(\boldsymbol{\theta})\right\rangle_{\boldsymbol{\theta}} = \lambda_k\delta_{k,\ell}.\end{aligned} \tag{App.20}$$

Since $\mathbf{r}(\boldsymbol{\theta})$ and $\tilde{\mathbf{r}}(\boldsymbol{\theta})$ have the same inner product kernel, they must posess the same kernel eigenfunctions ψ_k and kernel eigenvalues λ_k, which are identified through the eigenvalue problem $\int p(\boldsymbol{\theta})K(\boldsymbol{\theta}, \boldsymbol{\theta}')\psi_k(\boldsymbol{\theta})d\boldsymbol{\theta} = \lambda_k\psi_k(\boldsymbol{\theta})$. We therefore have the following two singular value decompositions for $\mathbf{r}$ and $\tilde{\mathbf{r}}$

$$\mathbf{r}(\boldsymbol{\theta}) = \sum_{k=1}^{N}\sqrt{\lambda_k}\mathbf{u}_k\psi_k(\boldsymbol{\theta})\ ,\ \tilde{\mathbf{r}}(\boldsymbol{\theta}) = \sum_{k=1}^{N}\sqrt{\lambda_k}\tilde{\mathbf{u}}_k\psi_k(\boldsymbol{\theta}). \tag{App.21}$$

where $\{\mathbf{u}_k\}_{k=1}^N$ and $\{\tilde{\mathbf{u}}_k\}_{k=1}^N$ are both complete sets of orthonormal vectors (the sums above run over possible zero eigenvalues). Taking the equation $\tilde{\mathbf{r}}(\boldsymbol{\theta}) = \mathbf{A}\mathbf{r}(\boldsymbol{\theta})$, we multiply both sides of the

equation by $\psi_k(\boldsymbol{\theta})$ and average over $\boldsymbol{\theta}$ giving

$$\left\langle \tilde{\mathbf{r}}(\boldsymbol{\theta})\psi_k(\boldsymbol{\theta})\right\rangle = \sqrt{\lambda_k}\tilde{\mathbf{u}}_k = \mathbf{A}\left\langle \mathbf{r}(\boldsymbol{\theta})\psi_k(\boldsymbol{\theta})\right\rangle_{\boldsymbol{\theta}} = \sqrt{\lambda_k}\mathbf{A}\mathbf{u}_k \tag{App.22}$$

For an eigenmode k with positive eigenvalue $\lambda_k > 0$, this implies $\tilde{\mathbf{u}}_k = \mathbf{A}\mathbf{u}_k$, while there is no corresponding constraint for the null modes with $\lambda_k = 0$. However, the action of $\mathbf{A}$ on the nullspace of the code has no influence on $\tilde{\mathbf{r}}$ so there is no loss in generality to restrict consideration to transformations $\mathbf{A}$ which satisfy $\tilde{\mathbf{u}}_k = \mathbf{A}\mathbf{u}_k$ for all $k \in [N]$ (rather than just the $\lambda_k > 0$ modes). This choice gives $\mathbf{A} = \sum_{k=1}^{N} \tilde{\mathbf{u}}_k \mathbf{u}_k^\top \in O(N)$. Thus, the space of codes $\tilde{\mathbf{r}}(\boldsymbol{\theta})$ with equivalent kernels to $\mathbf{r}(\boldsymbol{\theta}) \cdot \mathbf{r}(\boldsymbol{\theta}')$ generated through linear transformations is equivalent to all possible orthogonal transformations of the original code $\{\mathbf{Q}\mathbf{r}(\boldsymbol{\theta}) : \mathbf{Q} \in O(N)\}$. □

2.2 Necessary Conditions for Optimally Sparse Codes

Next we argue why optimally sparse codes should be lifetime and population selective. We consider the following optimization problem: find a non-negative neural responses $\mathbf{S} \in \mathbb{R}^{N\times P}$ and baseline vector $\boldsymbol{\delta} \in \mathbb{R}^N$ so that baseline subtracted responses $\mathbf{R} = \mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top$ realize a desired inner product kernel $\mathbf{K} \in \mathbb{R}^{P\times P}$ and have minimal total firing. This is equivalent to finding the most metabolically efficient code among the space of codes with equivalent inductive bias. Mathematically, we formulate this problem as

$$\min_{\mathbf{S}\in\mathbb{R}^{N\times P},\, \boldsymbol{\delta}\in\mathbb{R}^N} \sum_{i\mu} S_{i\mu}, \quad \text{s.t. } \left(\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top\right)^\top \left(\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top\right) = \mathbf{K}, \quad S_{i\mu} \geq 0 \;\; \forall i \in [N], \mu \in [P]. \tag{App.23}$$

To enforce the constraints for the definition of the kernel and the non-negativity of the responses, we introduce the following Lagrangian

$$\mathcal{L}(\mathbf{S}, \boldsymbol{\delta}, \mathbf{A}, \mathbf{V}) = \mathbf{1}^\top\mathbf{S}\mathbf{1} - \mathrm{Tr}\left(\left[(\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top)^\top(\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top) - \mathbf{K}\right]\mathbf{A}\right) - \mathrm{Tr}\mathbf{V}^\top\mathbf{S} \tag{App.24}$$

where $\mathbf{1}$ is the vector containing all ones, the Lagrange multiplier matrix $\mathbf{A}$ enforces the definition of the kernel and the KKT multiplier matrix $\mathbf{V}$ enforces the non-negativity constraints for each element of $\mathbf{S}$. The KKT conditions require that any local optimum of the objective would have to satisfy the following equations [3]

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{S}} &= \mathbf{1}\mathbf{1}^\top - (\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top)\mathbf{A} - \mathbf{V} = \mathbf{0} \\
\frac{\partial \mathcal{L}}{\partial \boldsymbol{\delta}} &= -(\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top)\mathbf{A}\mathbf{1} = \mathbf{0} \\
\frac{\partial \mathcal{L}}{\partial \mathbf{A}} &= (\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top)^\top(\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top) - \mathbf{K} = \mathbf{0} \\
\mathbf{V} \odot \mathbf{S} &= \mathbf{0},
\end{aligned}
\tag{App.25}
$$

where $\odot$ denotes the element-wise Hadamard product. Using the complementary slackness condition $\mathbf{S} \odot \mathbf{V} = \mathbf{0}$, and the first optimality condition $\frac{\partial \mathcal{L}}{\partial \mathbf{S}} = \mathbf{0}$, we have

$$
\mathbf{S} = \mathbf{S} \odot (\mathbf{S} - \boldsymbol{\delta}\mathbf{1}^\top)\mathbf{A} \tag{App.26}
$$

Therefore, for any neuron-stimulus pair (i, μ), either $S_{i\mu} = 0$ or $\sum_{\nu \in [P]} (S_{i\nu} - \delta_i) A_{\nu\mu} = 1$. Further, under the condition that $\mathbf{K}$ is full rank, we conclude that for any stimulus μ, $\sum_{\nu \in [P]} A_{\mu\nu} = 0$ from the equation $\frac{\partial \mathcal{L}}{\partial \boldsymbol{\delta}} = \mathbf{0}$. Let $\mathcal{I}_i = \{\mu \in [P] : S_{i\mu} > 0\}$ represent the set of stimuli for which neuron i fires. We will call this the *receptive field set* for neuron i. Let $\mathbf{B}_{(i)} \in \mathbb{R}^{P \times P}$ have entries

$$
[\mathbf{B}_{(i)}]_{\mu\nu} = \begin{cases} [\mathbf{A}^+_{(i)}]_{\mu\nu} & \mu, \nu \in \mathcal{I}_i \\ 0 & \mu \notin \mathcal{I}_i \text{ or } \nu \notin \mathcal{I}_i \end{cases} \tag{App.27}
$$

where the matrix $\mathbf{A}_{(i)}$ is the $|\mathcal{I}_i| \times |\mathcal{I}_i|$ minor of $\mathbf{A}$ obtained by taking all rows and columns with indices $\mu, \nu \in \mathcal{I}_i$, and $\mathbf{A}^+$ denotes pseudo-inverse of $\mathbf{A}$. Then the i-th neuron's tuning curve is a function of the index set $\mathcal{I}_i$ the baseline δ_i and the neuron-independent $P \times P$ matrix $\mathbf{A}$: $\mathbf{s}(\mathcal{I}_i, \delta_i, \mathbf{A}) = \mathbf{B}_{(i)} [\delta_i \mathbf{A} + \mathbf{I}] \mathbf{1} \in \mathbb{R}^P$. The non-negativity constraint for neuron i's tuning curve implies that $S_{i\mu} = \sum_{\nu \in \mathcal{I}_i} B_{(i),\mu\nu} [\delta_i \sum_{\gamma \in [P]} A_{\nu\gamma} + 1] > 0$ for all $\mu \in \mathcal{I}_i$. To satisfy the definition of the kernel, we have the following constraint on the matrix $\mathbf{A}$, the index sets $\mathcal{I}_i$ and baselines δ_i

$$
\mathbf{K} = \sum_{i=1}^{N} (\mathbf{s}(\mathcal{I}_i, \delta_i, \mathbf{A}) - \delta_i \mathbf{1})(\mathbf{s}(\mathcal{I}_i, \delta_i, \mathbf{A}) - \delta_i \mathbf{1})^\top \tag{App.28}
$$

This equation implictly defines the index sets $\mathcal{I}_i$ the baselines δ_i and the KKT matrix $\mathbf{A}$. We see that, in order to fit an arbitrary kernel, the receptive field sets $\{\mathcal{I}_i\}$ and baselines δ_i for each neuron must be sufficiently diverse since otherwise only a low rank kernel matrix can be achieved from the optimally sparse code. As a concrete example, suppose that $\mathcal{I}_i = \mathcal{I}$ so that $\mathbf{B}_{(i)} = \mathbf{B}$ and $\delta_i = \delta$ for all i. For example, this could occur if each neuron fired for every possible stimulus. In this case, the kernel would be rank one: $\mathbf{K} = N(\mathbf{s}(\mathcal{I}, \delta, \mathbf{A}) - \delta\mathbf{1})(\mathbf{s}(\mathcal{I}, \delta, \mathbf{A}) - \delta\mathbf{1})^\top$. In order to achieve a higher rank code there must be sufficient diversity of the receptive fields $\mathcal{I}_i$. Thus the only way for optimally sparse codes to realize high rank kernels $\mathbf{K}$ is to have neurons to have different receptive field sets $\mathcal{I}_i$. The necessary optimality conditions thus reveal a preference for sparse neural tuning

curves to have high *lifetime sparseness*; to achieve diverse index sets $\mathcal{I}_i$, any given neuron will fire only for a unique subset of the possible stimuli.

3 Theory of Generalization

3.1 Convergence of the delta-rule

Gradient descent training of readout weights $\mathbf{w}$ on a finite sample of size P converges to the kernel regression solution [4, 5, 6]. Let $\mathcal{D} = \{\boldsymbol{\theta}^\mu, y^\mu\}_{\mu=1}^P$ be the dataset with samples $\boldsymbol{\theta}^\mu$ and target values y^μ. We introduce a shorthand $\mathbf{r}^\mu = \mathbf{r}(\boldsymbol{\theta}^\mu)$ for convenience. The empirical loss we aim to minimize is a sum of the squared losses of each data point in the training set

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{\mu=1}^{P}(\mathbf{r}^\mu \cdot \mathbf{w} - y^\mu)^2. \tag{App.29}$$

Performing gradient descent updates

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta\frac{\partial \mathcal{L}}{\partial \mathbf{w}_t} = \mathbf{w}_t - \eta\sum_{\mu=1}^{P}\mathbf{r}^\mu(\mathbf{r}^\mu \cdot \mathbf{w}_t - y^\mu), \tag{App.30}$$

recovers the delta rule that we discussed in the main text [7, 8]. Letting the empirical response matrix $\mathbf{R} = [\mathbf{r}^1, ..., \mathbf{r}^P] \in \mathbb{R}^{N\times P}$ have a SVD $\mathbf{R} = \sum_k \sqrt{\hat{\lambda}_k}\hat{\mathbf{u}}_k\hat{\boldsymbol{\psi}}_k^\top$, and expanding the weights $\mathbf{w}_t = \sum_k w_{t,k}\hat{\mathbf{u}}_k$ and labels $\mathbf{y} = \sum_k \hat{v}_k\hat{\boldsymbol{\psi}}_k$ in their respective SVD bases, we find

$$w_{t+1,k} = w_{t,k} - \eta\hat{\lambda}_k w_{t,k} + \eta\sqrt{\hat{\lambda}_k}\hat{v}_k \tag{App.31}$$

For all directions with $\hat{\lambda}_k > 0$, the dynamics converge to the unique fixed point $w_k^* = \frac{\hat{v}_k}{\sqrt{\hat{\lambda}_k}}$, while for all modes with $\hat{\lambda}_k = 0$, the weights remain at $w_k^* = 0$. Thus

$$\mathbf{w}^* = \left[\sum_{k:\hat{\lambda}_k>0}\frac{\hat{\mathbf{u}}_k\hat{\boldsymbol{\psi}}_k^\top}{\sqrt{\hat{\lambda}_k}}\right]\mathbf{y} = \mathbf{R}\left[\sum_{k:\hat{\lambda}_k>0}\frac{\hat{\boldsymbol{\psi}}_k\hat{\boldsymbol{\psi}}_k^\top}{\hat{\lambda}_k}\right]\mathbf{y} = \mathbf{R}\mathbf{K}^+\mathbf{y} \tag{App.32}$$

where $\mathbf{K}^+$ is the Moore-Penrose inverse of the kernel matrix $K_{\mu,\nu} = K(\boldsymbol{\theta}^\mu, \boldsymbol{\theta}^\nu)$. The predictions of the learned function are given by $f = \mathbf{w}^* \cdot \mathbf{r}(\boldsymbol{\theta})$ which can be expressed as

$$f(\boldsymbol{\theta}) = \mathbf{k}(\boldsymbol{\theta})^\top\mathbf{K}^+\mathbf{y} \tag{App.33}$$

The fact that the solution can be written in terms of a linear combination of $\{K(\boldsymbol{\theta}, \boldsymbol{\theta}^\mu)\}_{\mu=1}^P$ is known as the representer theorem [9, 2]. A similar analysis for nonlinear readouts where $f(\boldsymbol{\theta}) = g\left(\mathbf{w}\cdot\mathbf{r}(\boldsymbol{\theta})\right)$ is provided in Appendix 3G.

3.1.1 Weight Decay and Ridge Regression

We can introduce a regularization term in our learning problem which penalizes the size of the readout weights. This leads to a modified learning objective of the form

$$\mathcal{L}(\mathbf{w}) = \sum_{\mu} (\mathbf{r}^{\mu} \cdot \mathbf{w} - y^{\mu})^2 + \lambda ||\mathbf{w}||^2. \tag{App.34}$$

Inclusion of this regularization alters the learning rule through *weight decay* $\mathbf{w}_{t+1} = (1 - \eta\lambda)\mathbf{w}_t + \eta \sum_{\mu} \mathbf{r}^{\mu}(\mathbf{r}^{\mu} \cdot \mathbf{w}_t - y^{\mu})$, which multiplies the existing weight value by a factor of $1 - \eta\lambda$ before adding the data dependent update. This learning problem and gradient descent dynamics have a closed form solution

$$f(\boldsymbol{\theta}) = \mathbf{r}(\boldsymbol{\theta}) \cdot \mathbf{w}^* = \sum_{\mu=1}^{P} \alpha^{\mu} K(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mu}) \ , \ \boldsymbol{\alpha} = (\mathbf{K} + \lambda \mathbf{I})^{-1} \mathbf{y}. \tag{App.35}$$

The generalization benefits of explicit regularization through weight decay is known to be related to the noise statistics in the learning problem [10]. We simulate weight decay only in Figure 6C, where we use $\lambda = 0.01 \sum_k \lambda_k$ to improve numerical stability at large P.

3.2 Computation of Learning Curves

Recent work has established analytic results that predict the average case generalization error for kernel regression

$$E_g = \left\langle E_g(\mathcal{D}) \right\rangle_{\mathcal{D}} = \left\langle (f(\boldsymbol{\theta}, \mathcal{D}) - y(\boldsymbol{\theta}))^2 \right\rangle_{\boldsymbol{\theta}, \mathcal{D}} \tag{App.36}$$

where $E_g(\mathcal{D}) = \left\langle (f(\boldsymbol{\theta}, \mathcal{D}) - y(\boldsymbol{\theta}))^2 \right\rangle_{\boldsymbol{\theta}}$ is the generalization error for a certain sample $\mathcal{D}$ of size P and $f(\boldsymbol{\theta}, \mathcal{D})$ is the kernel regression solution for $\mathcal{D}$ [11, 10]. The typical or average case error E_g is obtained by averaging over all possible datasets of size P. This average case generalization error is determined solely by the decomposition of the target function $y(\mathbf{x})$ along the eigenbasis of the kernel and the eigenspectrum of the kernel. This diagonalization takes the form

$$\int p(\boldsymbol{\theta}) K(\boldsymbol{\theta}, \boldsymbol{\theta}') \psi_k(\boldsymbol{\theta}) d\boldsymbol{\theta} = \lambda_k \psi_k(\boldsymbol{\theta}') \tag{App.37}$$

Since the eigenfunctions form a complete set of square integrable functions, we expand both the target function $y(\boldsymbol{\theta})$ and the learned function $f(\boldsymbol{\theta})$ in this basis

$$y(\boldsymbol{\theta}) = \sum_k v_k \psi_k(\boldsymbol{\theta}) \ , \ f(\boldsymbol{\theta}) = \sum_k w_k \psi_k(\boldsymbol{\theta}) \tag{App.38}$$

Due to the orthonormality of the kernel eigenfunctions $\{\psi_k\}$, the generalization error for any set of coefficients $\mathbf{w}$ is

$$E_g(\mathbf{w}) = \left\langle (y(\boldsymbol{\theta}) - f(\boldsymbol{\theta}))^2 \right\rangle_{\boldsymbol{\theta}} = \sum_k (w_k - v_k)^2 = ||\mathbf{w} - \mathbf{v}||^2 \tag{App.39}$$

We now introduce training error, or empirical loss, which depends on the disorder in the dataset $\mathcal{D} = \{(\boldsymbol{\theta}^\mu, y^\mu)\}_{\mu=1}^P$

$$H(\hat{\mathbf{v}}, \mathcal{D}) = \sum_\mu (\mathbf{w} \cdot \psi(\boldsymbol{\theta}^\mu) - \mathbf{v} \cdot \psi(\boldsymbol{\theta}^\mu))^2 + \lambda \sum_k \frac{w_k^2}{\lambda_k} \tag{App.40}$$

It is straightforward to verify that the optimal $\hat{\mathbf{v}}^*$ which minimizes $H(\hat{\mathbf{v}}, \mathcal{D})$ is the kernel regression solution for kernel with eigenvalues $\{\lambda_k\}$ when $\lambda \to 0$. The optimal weights $\hat{\mathbf{v}}$ can be identified through the first order condition $\nabla H(\hat{\mathbf{v}}, \mathcal{D}) = 0$ which gives

$$\mathbf{w}^* = (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \lambda \boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top \mathbf{v} = \mathbf{v} - \lambda (\boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \lambda \boldsymbol{\Lambda}^{-1})^{-1} \boldsymbol{\Lambda}^{-1} \mathbf{v} \tag{App.41}$$

where $\Psi_{k,\mu} = \psi_k(\boldsymbol{\theta}^\mu)$ are the eigenfunctions evaluated on the training data and $\Lambda_{k,\ell} = \delta_{k,\ell}\lambda_k$ is a a diagonal matrix containing the kernel eigenvalues. The generalization error for this optimal solution is

$$E_g(\mathcal{D}) = ||\mathbf{w}^* - \mathbf{v}||^2 = \mathbf{v}^\top \boldsymbol{\Lambda}^{-1} \mathbf{G}(\mathcal{D})^2 \boldsymbol{\Lambda}^{-1} \mathbf{v} \ , \ \mathbf{G}(\mathcal{D}) = \left(\frac{1}{\lambda} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \boldsymbol{\Lambda}^{-1} \right)^{-1} \tag{App.42}$$

We note that the dependence on the randomly sampled dataset $\mathcal{D}$ only appears through the matrix $\mathbf{G}(\mathcal{D})$. Thus to compute the *typical* generalization error we need to average over this matrix $\left\langle \mathbf{G}(\mathcal{D}) \right\rangle_{\mathcal{D}}$. There are multiple strategies to perform such an average and we will study one here based on a partial differential equation which was introduced in [12, 13] and studied further in [11, 10]. In this setting, we denote the average matrix $\mathbf{G}(P) = \left\langle \mathbf{G}(\mathcal{D}) \right\rangle_{|\mathcal{D}|=P}$ for a dataset of size P. We first will derive a recursion relationship using the Sherman Morrison formula for a rank-1 update to an inverse matrix. We imagine adding a new sampled feature vector ϕ to a dataset ψ with size P. The average matrix $\mathbf{G}(P+1)$ at $P+1$ samples can be related to $\mathbf{G}(P)$ through the Sherman Morrison rule

$$\begin{aligned} \mathbf{G}(P+1) &= \left\langle \left(\frac{1}{\lambda} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \frac{1}{\lambda} \psi\psi^\top + \boldsymbol{\Lambda}^{-1} \right)^{-1} \right\rangle_{\psi,\mathcal{D}} = \mathbf{G}(P) - \left\langle \frac{\mathbf{G}(\mathcal{D}) \psi\psi^\top \mathbf{G}(\mathcal{D})}{\lambda + \psi^\top \mathbf{G}(\mathcal{D}) \psi} \right\rangle_{\psi,\mathcal{D}} \\ &\approx \mathbf{G}(P) - \frac{\left\langle \mathbf{G}(\mathcal{D}) \left\langle \psi\psi^\top \right\rangle_\psi \mathbf{G}(\mathcal{D}) \right\rangle_{\mathcal{D}}}{\lambda + \left\langle \psi^\top \mathbf{G}(\mathcal{D}) \psi \right\rangle_{\psi,\mathcal{D}}} \end{aligned} \tag{App.43}$$

where in the last step we approximated the average of the ratio with the ratio of averages. This operation, is of course, unjustified theoretically, but has been shown to produce accurate learning curves [11, 13]. Since the chosen basis of kernel eigenfunctions are orthonormal, the average over the new sample is trivial $\left\langle \psi\psi^\top \right\rangle_\phi = \mathbf{I}$. We thus arrive at the following recursion relationship for $\mathbf{G}$

$$\mathbf{G}(P+1) = \mathbf{G}(P) - \frac{\left\langle \mathbf{G}(\mathcal{D})^2 \right\rangle_{\mathcal{D}}}{\lambda + \mathrm{Tr}\ \mathbf{G}(P)} \tag{App.44}$$

By introducing an additional source J so that $\mathbf{G}(\mathcal{D}, J)^{-1} = \frac{1}{\lambda} \boldsymbol{\Psi}\boldsymbol{\Psi}^\top + \boldsymbol{\Lambda}^{-1} + J\mathbf{I}$, we can relate

$\mathbf{G}(\mathcal{D}, J)$'s first and second moments through differentiation

$$\frac{\partial}{\partial J}\mathbf{G}(P,J) = \frac{\partial}{\partial J}\left\langle \left(\frac{1}{\lambda}\mathbf{\Psi}\mathbf{\Psi}^\top + J\mathbf{I} + \mathbf{\Lambda}^{-1}\right)^{-1}\right\rangle_{\mathcal{D}} = -\left\langle \mathbf{G}(\mathcal{D},J)^2\right\rangle_{\mathcal{D}}. \tag{App.45}$$

Thus the recursion relation simplifies to

$$\mathbf{G}(P+1,J) - \mathbf{G}(P,J) \approx \frac{\partial}{\partial p}\mathbf{G}(p,J) = \frac{1}{\lambda + \mathrm{Tr}\mathbf{G}(P,J)}\frac{\partial}{\partial J}\mathbf{G}(P,J), \tag{App.46}$$

where we approximated the finite difference in P as a derivative, treating P as a continuous variable. Taking the trace of both sides and defining $\kappa(P,J) = \lambda + \mathrm{Tr}\mathbf{G}(P,J)$ we arrive at the following quasilinear PDE

$$\frac{\partial}{\partial P}\kappa(P,J) = \frac{1}{\kappa(P,J)}\frac{\partial}{\partial J}\kappa(P,J) \tag{App.47}$$

with the initial condition $\kappa(0,J) = \lambda + \mathrm{Tr}(\mathbf{\Lambda}^{-1} + J\mathbf{I})^{-1}$. Using the method of characteristics, we arrive at the solution $\kappa(P,J) = \lambda + \mathrm{Tr}\left(\mathbf{\Lambda}^{-1} + (v + \frac{P}{\kappa(P,J)})\mathbf{I}\right)^{-1}$. Using this solution to κ, we can identify the solution to $\mathbf{G}$

$$\mathbf{G}(P,J)_{k,\ell} = \left(\frac{P}{\kappa} + J + \lambda_k^{-1}\right)^{-1}\delta_{k,\ell} = \frac{\kappa\lambda_k}{\lambda_k P + \kappa + J\kappa\lambda_k}\delta_{k,\ell}. \tag{App.48}$$

The generalization error, therefore can be written as

$$E_g = \mathbf{v}^\top\mathbf{\Lambda}^{-1}\left\langle \mathbf{G}(\mathcal{D})^2\right\rangle_{\mathcal{D}}\mathbf{\Lambda}^{-1}\mathbf{v} = -\frac{\partial}{\partial J}\mathbf{v}^\top\mathbf{\Lambda}^{-1}\mathbf{G}(P,J)\mathbf{\Lambda}^{-1}\mathbf{v} \tag{App.49}$$

$$= -\sum_k \frac{v_k^2}{\lambda_k^2}\frac{\partial}{\partial J}\left(\frac{P}{\kappa} + J + \lambda_k^{-1}\right)^{-1} = \frac{\kappa^2}{1-\gamma}\sum_k \frac{v_k^2}{(\lambda_k P + \kappa)^2}, \tag{App.50}$$

where $\gamma = P\sum_k \frac{\lambda_k^2}{(\lambda_k P + \kappa)^2}$, giving the desired result. Note that κ depends on J implicitly, which is the source of the $\frac{1}{1-\gamma}$ factor. This result was recently reproduced using techniques from statistical mechanics [11, 10].

3.3 Spectral Bias and Code-Task Alignment

Through implicit differentiation it is straightforward to verify that the ordering of the mode errors $E_k = \frac{\kappa^2}{1-\gamma}(\lambda_k P + \kappa)^{-2}$ matches the ordering of the eigenvalues [10]. Let $\lambda_k > \lambda_\ell$, then we have

$$\begin{aligned}\frac{d}{dP}\log\left(\frac{E_k}{E_\ell}\right) =& 2\left[\frac{\lambda_\ell}{\lambda_\ell P + \kappa} - \frac{\lambda_k}{\lambda_k P + \kappa}\right] \\ &+ 2\kappa'(P)\left[\frac{1}{\lambda_\ell P + \kappa} - \frac{1}{\lambda_k P + \kappa}\right].\end{aligned} \tag{App.51}$$

Since $\lambda_\ell < \lambda_k$, the first bracket must be negative and the second bracket must be positive. Further, it is straightforward to compute that $\kappa'(P) = -\frac{\kappa\gamma}{P(1+\gamma)} < 0$. Therefore $\lambda_k > \lambda_\ell$ implies

$\frac{d}{dP}\log\left(\frac{E_k}{E_\ell}\right) < 0$ for all P. Since $\log\left(\frac{E_k}{E_\ell}\right) = 0$ at $P = 0$ we therefore have that $\log\big(E_k/E_\ell\big) < 0$ for all P and consequently $E_k < E_\ell$. Modes with larger eigenvalues λ_k have lower normalized mode errors E_k. This observation can be used to prove that target functions acting on the same data distribution with higher cumulative power distributions $C(k)$ for all k will have lower generalization error normalized by total target power, $E_g(P)/E_g(0)$, for all P. Proof can be found in [10].

3.4 Asymptotic power law scaling of learning curves

Exponential Spectral Decays: First, we will study the setting relevant to the von-Mises kernel where $\lambda_k \sim \beta^k$ and $v_k^2 \sim \alpha^k$ where $\alpha, \beta < 1$. This exponential behavior accounts for differences in bandwidth between kernels which modulates the base β of the exponential scaling of λ_k with k. We will approximate the sum over all mode errors with an integral

$$E_g = \frac{\kappa^2}{1-\gamma}\sum_{k=0}^{\infty}\frac{v_k^2}{(\lambda_k P+\kappa)^2} \sim \kappa^2\int_0^\infty \frac{\alpha^k}{(\beta^k P+\kappa)^2}dk. \tag{App.52}$$

If we include a regularization parameter λ, then $\kappa \sim \lambda$ as $P \to \infty$. With this fact, we can therefore approximate the integral at large P by splitting it up into all $k < k^* = \ln\big(P/\lambda\big)/\ln\big(1/\beta\big)$ and $k > k^*$.

$$E_g \sim \frac{\lambda^2}{P^2}\int_0^{k^*}\frac{\alpha^k}{\beta^{2k}\left[1+\frac{\lambda}{P\beta^k}\right]^2}dk + \int_{k^*}^\infty \frac{\alpha^k}{\left[1+\frac{\beta^k P}{\lambda}\right]^2}dk = AP^{-\frac{\log(1/\alpha)}{\log(1/\beta)}} + \sum_{n=0}^{\infty} A_n P^{-n-2} \tag{App.53}$$

for P-independent constants A and A_n. Thus, we obtain a power law scaling of the learning curve E_g which is dominated at large P by $E_g \sim P^{-\min\left(2,\frac{\ln(1/\alpha)}{\ln(1/\beta)}\right)}$. For the von-Mises kernel we can approximate the spectra with $\lambda_k \sim \sigma^{-2k}$ and $v_k^2 \sim \sigma_T^{-2k}$ giving rise to a generalization scaling scaling $E_g \sim P^{-\min\left(2,\frac{\ln\sigma_T}{\ln\sigma}\right)}$.

Power Law Spectral Decays: The same arguments can be applied for power law kernels $\lambda_k \sim k^{-b}$ and power law targets $v_k^2 \sim k^{-a}$, which is of interest due to its connection to nonlinear rectified neural populations. In this setting, the generalization error is

$$\begin{aligned} E_g &\approx \int_1^\infty \frac{k^{-a}}{(k^{-b}P+\kappa)^2}dk \approx \frac{\kappa^2}{P^2}\int_1^{P^{1/b}} k^{-a+2b}dk + \int_{P^{1/b}}^\infty k^{-a}dk \\ &= \frac{1}{P^2(1-a+2b)}\left[P^{(1-a)/b+2}-1\right] + \frac{1}{a-1}P^{(1-a)/b}. \end{aligned} \tag{App.54}$$

We see that there are two possible power law scalings for E_g with the exponents $(a-1)/b$ and 2. At large P this formula will be dominated by the term with minimum exponent so $E_g \sim P^{-\min(a-1,2b)/b}$.

3.5 Laplace Kernel Generalization

We calculate similar learning curves as we did for the von-Mises kernel but with Laplace kernels to show that our results is not an artifact of the infinite differentiability of the Von Mises kernel. Each of these Laplace kernels has the same asymptotic power law spectrum $\lambda_k \sim o(k^{-2})$, exhibiting a discontinuous first derivative (Figure App.5 A). Despite having the same spectral scaling at large k, these kernels can give dramatically different performance in learning tasks, again indicating the influence of the top eigenvalues on generalization at small P (Figure App.5). Again, the trend for which kernels perform best at low P can be reversed at large P. In this case, all generalization errors scale with $E_g \sim P^{-2}$ (Figure App.5B). More generally, our theory shows that if the task power spectrum and kernel eigenspectrum are both falling as power laws with exponents a and b respectively, then the generalization error asymptotically falls with a power law, $E_g \sim P^{-\min(a-1,2b)/b}$ (Methods) [11]. This decay is fastest when $b \geq \frac{a-1}{2}$ for which $E_g \sim P^{-2}$. Therefore, the tail of the kernel's eigenvalue spectrum determines the large sample size behavior of the generalization error for power law kernels. Small sample size limit is still governed by the bulk of the spectrum.

3.6 Learning with Multiple Output Channels

Our theory is not limited to scalar target functions but rather can be easily extended to multiple output functions $y_1, ..., y_C$ from the same data, if for example the task requires computing class membership for C categories. In this setting, each data point has the form $(\boldsymbol{\theta}^\mu, \mathbf{y}^\mu)$ with $\mathbf{y}^\mu \in \mathbb{R}^C$. For these C classes, the generalization error takes the form

$$E_g = \left\langle ||\mathbf{f}(\boldsymbol{\theta}) - \mathbf{y}(\boldsymbol{\theta})||^2 \right\rangle = \sum_{c=1}^{C} \left\langle \left(f_c(\boldsymbol{\theta}) - y_c(\boldsymbol{\theta})\right)^2 \right\rangle = \sum_k \left[\sum_c \left\langle y_c(\boldsymbol{\theta}) \phi_k(\boldsymbol{\theta}) \right\rangle^2 \right] E_k \ . \tag{App.55}$$

We therefore find that the generalization error in the multi-class setting is the same as the E_g obtained for a single scalar target function with power spectrum $v_k^2 = \sum_c \left\langle y_c(\boldsymbol{\theta}) \phi_k(\boldsymbol{\theta}) \right\rangle^2$ [11, 10]. The relevant cumulative power distribution measures the fraction of total output variance captured by the first k eigenfunctions of the population code

$$C(k) = \frac{\sum_c \sum_{\ell=1}^{k} \left\langle y_c(\boldsymbol{\theta}) \phi_k(\boldsymbol{\theta}) \right\rangle^2}{\sum_c \sum_{\ell=1}^{\infty} \left\langle y_c(\boldsymbol{\theta}) \phi_k(\boldsymbol{\theta}) \right\rangle^2} \ . \tag{App.56}$$

3.7 Convergence of Delta-Rule for Nonlinear Readouts

In this section, we consider gradient descent dynamics on a least squares cost with nonlinear readout function. Let $\mathcal{D} = \{\boldsymbol{\theta}^\mu, y^\mu\}_{\mu=1}^P$ be the dataset with samples $\boldsymbol{\theta}^\mu$ and target values y^μ. We introduce a shorthand $\mathbf{r}^\mu = \mathbf{r}(\boldsymbol{\theta}^\mu)$ for convenience. We consider output neurons which produce activity $f(\boldsymbol{\theta}) = g(\mathbf{w} \cdot \mathbf{r}(\boldsymbol{\theta}))$ for invertible nonlinear function g with non-vanishing gradient. The empirical

loss we aim to minimize is a sum of the squared losses of each data point in the training set

$$\mathcal{L}(\mathbf{w}) = \sum_{\mu=1}^{P} (g(\mathbf{r}^\mu \cdot \mathbf{w}) - y^\mu)^2. \tag{App.57}$$

Performing gradient descent updates

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \mathbf{w}_t - \eta \sum_{\mu=1}^{P} \mathbf{r}^\mu g'(\mathbf{w} \cdot \mathbf{r}^\mu)(g(\mathbf{r}^\mu \cdot \mathbf{w}_t) - y^\mu), \tag{App.58}$$

recovers the delta rule that we discussed in the main text for $g(z) = z$ [7, 8]. We define the variables $\mathbf{z}_t = \mathbf{R}^\top \mathbf{w}_t$ which satisfy the dynamics

$$\mathbf{z}_{t+1} - \mathbf{z}_t = \eta \mathbf{K} \text{ diag}(g'(\mathbf{z}_t)) \, (\mathbf{y} - g(\mathbf{z}_t)) \tag{App.59}$$

The possible fixed points of this system are the $\mathbf{z}$ which satisfy $\mathbf{K} \text{ diag}(g'(\mathbf{z})) \, (\mathbf{y} - g(\mathbf{z})) = 0$. We can make further progress under the condition that $\mathbf{K}$ is full rank (generally true in overparameterized setting $N > P$), the activation function is monotonic $g'(z) > 0$ for all z and is invertible (some examples include softplus $g(z) = \ln(1 + e^z)$, sigmoids such as $g(z) = \tanh(z), \text{erf}(z)$, etc.). If y^μ are in the range of g, then we obtain the simple fixed point condition $\mathbf{z}^* = g^{-1}(\mathbf{y})$. From this condition, we can infer the learned function. First note that $\mathbf{w}_t \in \text{span}\{\mathbf{r}^\mu\}_{\mu=1}^P$ for all t so that $\mathbf{w}^* = \mathbf{R}\boldsymbol{\alpha}^*$. Multiplying both sides of this last expression by $\mathbf{R}^\top$, we find

$$\mathbf{R}^\top \mathbf{R} \boldsymbol{\alpha}^* = \mathbf{R}^\top \mathbf{w}^* = \mathbf{K} \boldsymbol{\alpha}^* = \mathbf{z}^* \implies \boldsymbol{\alpha}^* = \mathbf{K}^{-1} \mathbf{z}^* = \mathbf{K}^{-1} g^{-1}(\mathbf{y}) \tag{App.60}$$

which is legitimate under the assumption that $\mathbf{K}$ was full rank. Now, the final learned function is merely $f(\boldsymbol{\theta}) = g(\mathbf{w}^* \cdot \mathbf{r}(\boldsymbol{\theta})) = g\left(\mathbf{r}(\boldsymbol{\theta})^\top \mathbf{R} \mathbf{K}^{-1} g^{-1}(\mathbf{y})\right) = g\left(\mathbf{k}(\boldsymbol{\theta})^\top \mathbf{K}^{-1} g^{-1}(\mathbf{y})\right)$.

3.8 Comparison of our Theory with Recursive Least Squares Adaptive Filter

In the under-parameterized regime where $P > N$, the asymptotic error of recursive least squares adaptive filter has been obtained in prior works [14, 15]. Let $\mathbf{w}$ represent the weights and $\mathbf{r}_\mu$ represent the μ-th presented example. We aim to reproduce a noisy ground truth $y_\mu = \mathbf{w}^* \cdot \mathbf{r}_\mu + \epsilon_\mu$. Analysis of the recursive least squares filter in the under-parameterized regime $P > N + 1$ can be performed exactly

$$\mathbf{w}(P) = \text{argmin}_{\mathbf{w} \in \mathbb{R}^N} \sum_{\mu=1}^{P} [y_\mu - \mathbf{w} \cdot \mathbf{r}_\mu]^2 \tag{App.61}$$

This has a unique minimizer when $P > N$. Letting $\mathbf{C}(P) = \sum_{\mu=1}^{P} \mathbf{r}_\mu \mathbf{r}_\mu^\top$

$$\mathbf{w}(P) = \mathbf{C}(P)^{-1}\left[\sum_{\mu=1}^{P} y_\mu \mathbf{r}_\mu\right] = \mathbf{C}(P)^{-1}\left[\mathbf{C}(P)\mathbf{w}^* + \sum_{\mu=1}^{P}\mathbf{r}_\mu\epsilon_\mu\right]$$
$$= \mathbf{w}^* + \mathbf{C}(P)^{-1}\left[\sum_{\mu}\mathbf{r}_\mu\epsilon_\mu\right] \tag{App.62}$$

This filter can be computed efficiently in an online manner using the Recursive Least Squares (RLS) algorithm. This recursion is obtained through the formula for the rank-one update to an inverse matrix

$$\mathbf{w}(P+1) = \left[\mathbf{C}(P) + \mathbf{r}_{P+1}\mathbf{r}_{P+1}^\top\right]^{-1}\left[\sum_{\mu=1}^{P+1}\mathbf{r}_\mu\epsilon_\mu\right]$$
$$= \mathbf{C}(P)^{-1}\left[\sum_{\mu=1}^{P+1}\mathbf{r}_\mu\epsilon_\mu\right] - \frac{\mathbf{C}(P)^{-1}\mathbf{r}_{P+1}\mathbf{r}_{P+1}^\top\mathbf{C}(P)^{-1}}{1+\mathbf{r}_{P+1}^\top\mathbf{C}(P)^{-1}\mathbf{r}_{P+1}}\left[\sum_{\mu=1}^{P+1}\mathbf{r}_\mu\epsilon_\mu\right]$$
$$= \left[\mathbf{I} - \frac{\mathbf{C}(P)^{-1}\mathbf{r}_{P+1}\mathbf{r}_{P+1}^\top}{1+\mathbf{r}_{P+1}^\top\mathbf{C}(P)^{-1}\mathbf{r}_{P+1}}\right]\left[\mathbf{w}(P) + \mathbf{r}_{P+1}\epsilon_{P+1}\right]. \tag{App.63}$$

Under the assumption that the noise has covariance $\left\langle \epsilon\epsilon^\top \right\rangle = \sigma^2\mathbf{I}$ and the neural responses are Gaussian $\mathbf{r}_\mu \sim \mathcal{N}(0, \mathbf{\Sigma})$, the excess generalization error can be obtained exactly using the formula for the average of an inverse Wishart matrix $\left\langle \mathbf{C}(P)^{-1}\right\rangle = \frac{1}{P-N-1}\mathbf{\Sigma}^{-1}$ giving

$$E_g - \sigma^2 = \mathrm{Tr}\mathbf{\Sigma}\left\langle (\mathbf{w}(P)-\mathbf{w}^*)(\mathbf{w}(P)-\mathbf{w}^*)^\top\right\rangle_{\{\mathbf{r}_\mu,\epsilon_\mu\}}$$
$$= \mathrm{Tr}\mathbf{\Sigma}\left\langle \mathbf{C}(P)^{-1}\mathbf{R}(P)\epsilon\epsilon^\top\mathbf{R}(P)^\top\mathbf{C}(P)^{-1}\right\rangle$$
$$= \sigma^2\mathrm{Tr}\mathbf{\Sigma}\left\langle \mathbf{C}(P)^{-1}\right\rangle = \sigma^2\frac{\mathrm{Tr}[\mathbf{\Sigma}\mathbf{\Sigma}^{-1}]}{P-N-1} = \sigma^2\frac{N}{P-N-1}. \tag{App.64}$$

This analysis shows that RLS is consistent in the sense that as $P \to \infty$ with N held constant the generalization error reach its theoretical minimum σ^2.

Though our main paper focuses on noise free estimation ($\sigma^2 = 0$) and the over-parameterized ($N > P$) rather than under parameterized case (which is what is discussed above), the above result agrees with what is obtained by our theory in the under-parameterized case since, in the $P > N$ limit, $\kappa = 0$ and $\gamma = \frac{N}{P}$. From our previous work, the error in the underparameterized noisy case for large P is [10]

$$E_g - \sigma^2 = \sigma^2 \frac{\gamma}{1-\gamma} = \sigma^2 \frac{N/P}{1-N/P} = \sigma^2 \frac{N}{P-N} \tag{App.65}$$

which is agrees with the exact result up to an error which is negligible in the large N, P limit. In this over-parameterized setting, the only contribution to the error comes from the explicit noise in the target function σ^2.

We note however, that the above technique does not generalize to the *overparameterized* setting where the system is trying to learn from a small number of examples $P < N$. In this case, when $\sigma = 0$, the delta-rule converges to the minimum norm solution which has the form

$$\mathbf{w}(P) = \mathbf{R}(P)\mathbf{K}(P)^{-1}\mathbf{y}(P) = \mathbf{R}(P)\mathbf{K}(P)^{-1}\mathbf{R}(P)^\top \mathbf{w}^* \tag{App.66}$$

where $\mathbf{K}(P) = \mathbf{R}(P)^\top \mathbf{R}(P)$. We see that the learned weight vector $\mathbf{w}(P)$ is a random rank-P projection of the optimal weight vector $\mathbf{w}^*$. The challenging task is to compute an average over the random data $\mathbf{R}$ to get generalization error

$$E_g = \mathbf{w}^{*\top} \left\langle \left(\mathbf{I} - \mathbf{R}(P)\mathbf{K}(P)^{-1}\mathbf{R}(P)^\top\right) \mathbf{\Sigma} \left(\mathbf{I} - \mathbf{R}(P)\mathbf{K}(P)^{-1}\mathbf{R}(P)^\top\right) \right\rangle \mathbf{w}^* \tag{App.67}$$

We solve this problem by first making a relaxation to the $\lambda \to 0$ limit of ridge regression

$$\begin{aligned}
\mathbf{R}(P)\mathbf{K}^{-1}\mathbf{R}(P) &= \lim_{\lambda\to 0} \mathbf{R}(P)\left(\mathbf{K}(P) + \lambda \mathbf{I}\right)^{-1}\mathbf{R}(P)^\top = \lim_{\lambda\to 0}\left(\mathbf{C}(P) + \lambda\mathbf{I}\right)^{-1}\mathbf{C}(P) \\
&= \lim_{\lambda\to 0}\left[\mathbf{I} - \lambda\left(\mathbf{C}(P) + \lambda\mathbf{I}\right)^{-1}\right] \\
\implies E_g &= \lim_{\lambda\to 0} \mathbf{v}^\top \mathbf{\Lambda}^{-1} \left\langle \left(\frac{1}{\lambda}\mathbf{\Psi}\mathbf{\Psi}^\top + \mathbf{\Lambda}^{-1}\right)^{-2} \right\rangle \mathbf{\Lambda}^{-1}\mathbf{v} \\
&= -\lim_{\lambda\to 0}\lim_{z\to 0} \frac{\partial}{\partial z} \mathrm{Tr} \left\langle \left(\frac{1}{\lambda}\mathbf{\Phi}\mathbf{\Phi}^\top + \mathbf{\Lambda}^{-1} + z\mathbf{\Lambda}^{-1}\mathbf{v}\mathbf{v}^\top\mathbf{\Lambda}^{-1}\right)^{-1} \right\rangle
\end{aligned} \tag{App.68}$$

where $\mathbf{v} = \mathbf{\Lambda}^{1/2}\mathbf{U}^\top \mathbf{w}^*$. This average can be characterized exactly for arbitrary $\lambda, \mathbf{\Lambda}, \mathbf{w}^*$ only in asymptotic limits where the number of samples P and the number of features N are both large. Though in this present work, we resort to the PDE approximation, our theory can also be derived with the replica method from disordered systems [11, 10].

3.9 Comparison of our Theory with Wiener Filter and Equivalent Kernel

We will provide a discussion of how the theory of generalization used in our paper differs from these older results, such as those discussed in Chapter 7 of [2]. Many of the older asymptotic results for

kernel regression with a ridge λ can be obtained through an averaging argument (see for example Rasmussen and Williams Chapter 7.1 and 7.2 [2]). In the limit where the number of independent samples is large, we have by the strong law of large numbers, a convergence of the empirical loss to an expectation integral

$$\mathcal{L}[f] = \frac{1}{P}\sum_{\mu}\left(f(\boldsymbol{\theta}^{\mu}) - y(\boldsymbol{\theta}^{\mu})\right)^2 + \frac{\lambda}{P}||f||^2_{\mathcal{H}} \sim \int p(\boldsymbol{\theta})\left(y(\boldsymbol{\theta}) - f(\boldsymbol{\theta})\right)^2 d\boldsymbol{\theta} + \frac{\lambda}{P}\sum_k \frac{\left\langle \psi_k(\boldsymbol{\theta}) f(\boldsymbol{\theta})\right\rangle^2_{\boldsymbol{\theta}}}{\lambda_k} \tag{App.69}$$

for a complete orthonormal family of eigenfunctions ψ_k. The minimizer to the above functional f and its corresponding generalization error E_g have the form

$$f(x) = \sum_k \frac{P\lambda_k}{P\lambda_k + \lambda} v_k \psi_k(x) \ , \ E_g = \lambda^2 \sum_k \frac{v_k^2}{(P\lambda_k + \lambda)^2} \tag{App.70}$$

for target coefficients $v_k = \left\langle y(x)\psi_k(x)\right\rangle$. This equation (which is equivalent to the Wiener filter when ψ_k are plane waves; see [2] Ch 7.1) is somewhat sensible in that this function approaches $y(x)$ as $P \to \infty$, and agrees with our theory in the limit of large ridge parameter $\lambda \to \infty$. Further, the large eigenvalue modes are learned at smaller values of P: it takes roughly $P \approx \lambda/\lambda_k$ samples to learn mode k. Note, however, that this theory is essentially useless in the interpolating limit $\lambda \to 0^+$ where it predicts that the learning curve has no dependence on sample size. This occurs because the fluctuations in the loss function from sample to sample (quenched disorder in statistical physics language) have been disregarded. Our analysis does not suffer from this issue and predicts the effects of sampling induced fluctuations in the loss function. Such sample fluctuations cannot typically be disregarded in thermodynamic limits where both dimension and dataset size are taken to infinity simultaneously [16]. A more complete discussion of the error due to sampling noise can be found in [10]. Our theory, in contrast to the Wiener filter described above, gives the following average predictor.

$$\left\langle f(x)\right\rangle_{\mathcal{D}} = \sum_k \frac{P\lambda_k}{P\lambda_k + \kappa} v_k \psi_k(x) \ , \ \kappa = \lambda + \kappa \sum_k \frac{\lambda_k}{P\lambda_k + \kappa} \tag{App.71}$$

We see that even in the $\lambda \to 0$ limit, the mean predictor is not equivalent to $y(x)$ at finite P. Rather the sample complexity of each eigenmode ψ_k is set by the evolution of the self-consistent function κ which can be nonzero even in the $\lambda \to 0$ limit. The full generalization error can be decomposed into bias and variance components, where bias component corresponds to the error of the average predictor $\left\langle f(\mathbf{x})\right\rangle_{\mathcal{D}}$.

$$E_g = \left\langle \left(\langle f(\mathbf{x}) \rangle_{\mathcal{D}} - y(\mathbf{x}) \right)^2 \right\rangle_{\mathbf{x}} + \left\langle \left(f(\mathbf{x}) - \langle f(\mathbf{x}) \rangle_{\mathcal{D}} \right)^2 \right\rangle_{\mathbf{x},\mathcal{D}} \tag{App.72}$$

where the bias and variance have the forms

$$\text{Bias} = \left\langle \left(\langle f(\mathbf{x}) \rangle_{\mathcal{D}} - y(\mathbf{x}) \right)^2 \right\rangle_{\mathbf{x}} = \kappa^2 \sum_k \frac{v_k^2}{(P\lambda_k + \kappa)^2} \tag{App.73}$$

$$\text{Variance} = \left\langle \left(f(\mathbf{x}) - \langle f(\mathbf{x}) \rangle_{\mathcal{D}} \right)^2 \right\rangle_{\mathbf{x},\mathcal{D}} = \frac{\kappa^2 \gamma}{1-\gamma} \sum_k \frac{v_k^2}{(P\lambda_k + \kappa)^2} \tag{App.74}$$

$$\gamma = P \sum_k \frac{\lambda_k^2}{(P\lambda_k + \kappa)^2} \tag{App.75}$$

We see that the bias is identical to the generalization error predicted by the equivalent kernel provided that the self-consistent function κ is substituted for the ridge parameter λ. This substitution allows for nonvanishing bias in the $\lambda \to 0$ limit since $\kappa > 0$ for P smaller than the number of nonzero eigenvalues. We find that the sample-to-sample variance of the learned predictor is nonzero $\langle f(x)^2 \rangle_{\mathcal{D}} - \langle f(x) \rangle_{\mathcal{D}}^2 > 0$ and varies with P. This is unlike the Wiener filter theory which treats the variance as 0 for all P due to concentration of the training loss.

We note a correspondence between the two theories in the heavily regularized $\lambda \to \infty$ limit. In this limit, $\kappa \sim \lambda$ and the bias approaches the risk predicted by the equivalent kernel theory. Further, the variance goes to zero in the $\lambda \to \infty$ limit since $\gamma \sim \lambda^{-2} \to 0$. A comparison of the two theories for linear ridge regression on random Gaussian design is provided in Figure App.7.

4 Visual Scene Reconstruction Task

4.1 Reconstruction of Natural Scenes from Neural Responses

Using the mouse V1 responses to natural scenes, we attempt to reconstruct original images from the neural codes using different numbers of images. The presented natural scenes are taken from ten classes of imagenet which can be downloaded from `https://github.com/MouseLand/stringer-pachitariu-et-al-2018b`. Let $\boldsymbol{\theta}^\mu \in \mathbb{R}^D$ be a D-dimensional flattened vector containing the pixel values of the μ-th image and let $\mathbf{r}^\mu \in \mathbb{R}^N$ represent the neural response to the μ-th image. The goal in the problem is to learn a collection of weights $\mathbf{W} \in \mathbb{R}^{D \times N}$ which map neural responses $\mathbf{r}^\mu$ to images $\boldsymbol{\theta}^\mu$

$$\boldsymbol{\theta}^\mu \approx \mathbf{W}\mathbf{r}^\mu. \tag{App.76}$$

The generalization error E_g again measures the average error on all points, averaged over all possible datasets $\mathcal{D} = \{(\boldsymbol{\theta}^\mu, \mathbf{r}^\mu)\}_{\mu=1}^P$ of size P. If the optimal weights for dataset $\mathcal{D}$ is $\mathbf{W}(\mathcal{D})$ then

the generalization error is

$$E_g = \left\langle ||\mathbf{W}(\mathcal{D})\mathbf{r}(\boldsymbol{\theta}) - \boldsymbol{\theta}||^2 \right\rangle_{\boldsymbol{\theta},\mathcal{D}} . \tag{App.77}$$

After identifying eigenfunctions $\phi_k(\boldsymbol{\theta})$, we expand the images in this basis $\boldsymbol{\theta} = \sum_k \mathbf{v}_k \phi_k(\boldsymbol{\theta})$ where $\mathbf{v}_k \in \mathbb{R}^D$. The generalization error is therefore $E_g = \sum_k |\mathbf{v}_k|^2 E_k(P)$ and the cumulative power is $C(k) = \frac{\sum_{\ell \leq k} |\mathbf{v}_\ell|^2}{\sum_{\ell=1}^{\infty} |\mathbf{v}_\ell|^2}$. We perform this reconstruction task on many filtered versions of the natural scenes. To construct a filter, we first compute the Fourier transform of the image. Let $\mathbf{M}(\boldsymbol{\theta}) \in \mathbb{R}^{\sqrt{D}\times\sqrt{D}}$ represent the non-flattened image and let $\hat{\mathbf{M}}(\boldsymbol{\theta}) \in \mathbb{R}^{\sqrt{D}\times\sqrt{D}}$ represent the Fourier transform of the image, computed explicitly as

$$\hat{M}_{k,k'}(\boldsymbol{\theta}) = D^{-1/4} \sum_{n,m} M_{n,m}(\boldsymbol{\theta}) \exp\Big(2\pi i(nk + mk')/\sqrt{D}\Big) \tag{App.78}$$

To develop the band-pass filter, we calculate $|\mathbf{k}| = \sqrt{k^2 + (k')^2}$ for each of the indices in the matrix. For a band-pass filter with parameters s_{max}, r we simply zero out the entries in $\hat{M}$ which correspond to states with frequencies outside the appropriate band: for any k, k' with $|\mathbf{k}| \notin [\sqrt{s_{max}^2 - r^2}, s_{max}^2]$ then $\hat{M}_{k,k'} \to 0$. We then perform the inverse Fourier transform on $\hat{M}$ to obtain a filtered version of the original image.

5 A Simple Feedforward Model of V1

5.1 Linear Neurons

We consider a simplified but instructive model of the V1 population code as a linear-nonlinear map from photoreceptor responses through Gabor filters and then nonlinearity [17, 18, 19]. Let $\mathbf{x} \in \mathbb{R}^2$ represent the two-dimensional retinotopic position of photoreceptors. The firing rates of the photoreceptor at position $\mathbf{x}$ to a static grating stimulus oriented at angle θ is

$$h(\mathbf{x}, \theta) = \cos\big(\mathbf{k}(\theta) \cdot \mathbf{x}\big) \ , \ \mathbf{k} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \in \mathbb{R}^2 \ , \ \theta \in [0, 2\pi]. \tag{App.79}$$

We model each V1 neuron's receptive field as a Gabor filter of the receptor responses $h(\mathbf{x}, \theta)$. The i-th V1 neuron has preferred wavevector $\mathbf{k}_i$, generating the following set of weights between photoreceptors and the i-th V1 neuron

$$\mathcal{F}(\mathbf{x}, \theta_i) = \frac{\sigma^2}{2\pi} e^{-\frac{\sigma^2}{2}|\mathbf{x}|^2} \cos\big(\mathbf{k}(\theta_i) \cdot \mathbf{x}\big). \tag{App.80}$$

The V1 population code is obtained by filtering the photoreceptor responses. By approximating the resulting sum over all retinal photoreceptors with an integral, we find the response of neuron i

to grating stimulus with wavenumber $\mathbf{k}$ is

$$\mathbf{w}(\theta_i) \cdot \mathcal{F}(\theta) = \int \mathcal{F}(\mathbf{x}, \theta_i) h(\mathbf{x}, \theta) d\mathbf{x} = \frac{1}{2} e^{-\frac{1}{2\sigma^2}|\mathbf{k}+\mathbf{k}_i|^2} + \frac{1}{2} e^{-\frac{1}{2\sigma^2}|\mathbf{k}-\mathbf{k}_i|^2}. \tag{App.81}$$

The response of neuron i is computed through nonlinear rectification of this input current $r_i(\theta) = g(\mathbf{w}(\theta_i) \cdot h(\theta))$. For a linear neuron $g(z) = z$, the kernel has the following form

$$K(\theta, \theta') = \frac{\cosh\Big(\cos\big(\theta - \theta'\big)/\sigma^2\Big)}{\cosh\big(\sigma^{-2}\big)}, \tag{App.82}$$

where the kernel is normalized to have maximum value of 1. Note that this normalization of the kernel is completely legitimate since it merely rescales each eigenvalue by a constant and does not change the learning curves.

Since the kernel only depends on the difference between angles $\theta - \theta'$, it is said to posess translation invariance. Such translation invariant kernels admit a Mercer decomposition in terms of Fourier modes $K(\theta) = \sum_n \lambda_n \cos(n\theta)$ since the Fourier modes diagonalize shift invariant integral operators on $\mathbb{S}^1$. For the linear neuron, the kernel eigenvalues scale like $\lambda_n \sim \frac{\beta^n}{2^n n!}$, indicating infinite differentiability of the tuning curves. Since λ_n decays rapidly with n, we find that this Gabor code has an inductive bias that favors low frequency functions of orientation θ.

5.2 Nonlinear Simple Cells

Introducing nonlinear functions $g(z)$ that map input currents z into the V1 population into firing rates, we can obtain a non-linear kernel $K_g(\theta)$ which has the following definition

$$K_g(\mathbf{k}, \mathbf{k}') = \int p(\mathbf{k}_i) g(\mathcal{F}(\mathbf{k}_i) \cdot \mathbf{h}(\mathbf{k})) g(\mathcal{F}(\mathbf{k}_i) \cdot \mathbf{h}(\mathbf{k}')) d\mathbf{k}_i. \tag{App.83}$$

In this setting, it is convenient to restrict $\mathbf{k}_i, \mathbf{k}, \mathbf{k}' \in \mathbb{S}^1$ and assume that the preferred wavevectors $\mathbf{k}_i$ are uniformly distributed over the circle. In this case, it suffices to identify a decomposition of the composed function $g(\mathbf{w}_i \cdot \mathbf{h}(\theta))$ in the basis of Chebyshev polynomials $T_n(z)$ which satisfy $T_n(\cos(\theta)) = \cos(n\theta)$

$$\begin{aligned} a_n &= \frac{1}{2\pi} \int_0^{2\pi} g\left(e^{-\frac{1}{\sigma^2}} \cosh\left(\frac{1}{\sigma^2}\cos(\theta)\right)\right) \cos(n\theta) d\theta \\ &= \frac{1}{2\pi} \int_{-1}^{1} \frac{1}{\sqrt{1-z^2}} g\left(e^{-\frac{1}{\sigma^2}} \cosh\left(z/\sigma^2\right)\right) T_n(z) dz, \end{aligned} \tag{App.84}$$

which can be computed efficiently with an appropriate quadrature scheme. Once the coefficients a_n are determined, we can compute the kernel by first letting θ_i to be the angle between $\mathbf{k}$ and $\mathbf{k}_i$

and letting θ be the angle between $\mathbf{k}$ and $\mathbf{k}'$

$$K_g(\theta) = \int_0^{2\pi} \frac{d\theta_i}{2\pi} \sum_{n,n'} a_n a_{n'} T_n(\cos(\theta_i)) T_{n'}(\cos(\theta_i + \theta)) d\theta_i$$
$$= \frac{1}{2} \sum_n a_n^2 \cos(n\theta). \tag{App.85}$$

Thus the kernel eigenvalues are $\lambda_n = \frac{1}{2} a_n^2(\psi)$.

Asymptotic scaling of spectra: Activation functions that encourage sparsity have slower eigenvalue decays. If the nonlinear activation function has the form $g_{q,t}(z) = \max\{0, z - a\}^q$, then the spectrum decays like $\lambda_n \sim n^{-2q-2}$. A simple argument justifies this scaling: if the function $g(e^{-\sigma^2} \cosh(\sigma^2 z))$ is only $q-1$ times differentiable then $a_n n^q \sim n^{-1}$ since $\sum_n a_n n^q$ must diverge. Therefore $\lambda_n = a_n^2 \sim n^{-2q-2}$. Note that this scaling is independent of the threshold.

5.3 Phase Variation, Complex Cells and Invariance

We can consider a slightly more complicated model where Gabors and stimuli have phase shifts

$$h(\mathbf{x}, \theta, \phi) = \cos\big(\mathbf{k}(\theta) \cdot \mathbf{x} - \phi\big) \ , \ \mathcal{F}(\mathbf{x}, \theta_i, \phi_i) = \frac{\sigma^2}{2\pi} e^{-\frac{\sigma^2}{2}|\mathbf{x}|^2} \cos(\mathbf{k}_i \cdot \mathbf{x} - \phi_i). \tag{App.86}$$

The simple cells are generated by nonlinearity

$$r_i(\theta, \phi) = g(\mathcal{F}(\theta_i, \phi_i) \cdot \mathbf{h}(\theta, \phi)). \tag{App.87}$$

The input currents into the simple V1 cells can be computed exactly

$$\mathbf{w}(\theta_i, \phi_i) \cdot \mathcal{F}(\theta, \phi) = \big\langle \cos(\mathbf{k}_i \cdot \mathbf{x} - \phi_i) \cos(\mathbf{k} \cdot \mathbf{x} - \phi) \big\rangle_{\mathbf{x} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \cdot \tag{App.88}$$
$$= \frac{1}{2} \cos(\phi + \phi_i) e^{-\frac{1}{2\sigma^2}|\mathbf{k} + \mathbf{k}_i|^2} + \frac{1}{2} \cos(\phi - \phi_i) e^{-\frac{1}{2\sigma^2}|\mathbf{k} - \mathbf{k}_i|^2}. \tag{App.89}$$

When $|\mathbf{k}| = |\mathbf{k}_i| = 1$, the simple cell tuning curves $r_i = g(\mathbf{w}_i \cdot \mathbf{h})$ only depend on $\cos(\theta - \theta_i)$ and ϕ, allowing a Fourier decomposition

$$r_i(\theta, \phi) = \sum_n a_n(\phi, \phi_i) \cos\big(n(\theta - \theta_i)\big)). \tag{App.90}$$

The simple cell kernel K_s, therefore decomposes into Fourier modes over θ

$$K_s(\theta, \theta', \phi, \phi') = \sum_n b_n(\phi, \phi') \cos\big(n(\theta - \theta')\big), \tag{App.91}$$

where $b_n(\phi, \phi') = \big\langle a_n(\phi, \phi_i) a_n(\phi', \phi_i) \big\rangle_{\phi_i}$. It therefore suffices to solve the infinite sequence of

integral eigenvalue problems over ϕ

$$\frac{1}{2\pi}\int_0^{2\pi} b_n(\phi,\phi')v_{n,k}(\phi)d\phi = \lambda_{n,k}v_{n,k}(\phi') \implies K_s(\theta,\theta',\phi,\phi') = \sum_{n,k}\lambda_{n,k}\cos\big(n(\theta-\theta')\big)v_{n,k}(\phi)v_{n,k}(\phi'). \tag{App.92}$$

With this choice it is straightforward to verify that the kernel eigenfunctions are $v_{n,k}(\theta,\phi) = e^{in\theta}v_{n,k}(\phi)$ with corresponding eigenvalue $\lambda_{n,k}$. Since b_n is not translation invariant in $\phi-\phi'$, the eigenfunctions $v_{n,k}$ are not necessarily Fourier modes. These eigenvalue problems for b_n must be solved numerically when using arbitrary nonlinearity ψ. The top eigenfunctions of the simple cell kernel depend heavily on the phase of the two grating stimuli ϕ. Thus, a pure orientation discrimination task which is independent of phase requires a large number of samples to learn with the simple cell population.

5.4 Complex Cells Populations are Phase Invariant

V1 also contains complex cells which possess invariance to the phase ϕ of the stimulus. Again using Gabor filters

$$\mathcal{F}(\mathbf{x},\theta_i,\phi_i) = \frac{\sigma^2}{2\pi}e^{-\frac{\sigma^2}{2}|\mathbf{x}|^2}\cos\big(\mathbf{k}(\theta_i)\cdot\mathbf{x}-\phi_i\big), \tag{App.93}$$

we model the complex cell responses with a quadratic nonlinearity and sum over two squared filters which are phase shifted by $\pi/2$

$$r_i(\theta,\phi) = (\mathcal{F}(\theta_i,\phi_i)\cdot\mathbf{h}(\theta,\phi))^2+(\mathcal{F}(\theta_i,\phi_i-\pi/2)\cdot\mathbf{h}(\theta,\phi))^2 = \frac{1}{4}e^{-\frac{1}{\sigma^2}|\mathbf{k}+\mathbf{k}_i|^2}+\frac{1}{4}e^{-\frac{1}{\sigma^2}|\mathbf{k}-\mathbf{k}_i|^2}+\frac{1}{2}e^{-\sigma^2}\cos(2\phi_i), \tag{App.94}$$

which we see is independent of the phase ϕ of the grating stimulus. Integrating over the set of possible Gabor filters $(\mathbf{k}_i,\phi_i)$ again gives the following kernel for the complex cells

$$K_c(\theta) = \frac{1}{\cosh(2\beta)}\cosh\big(2\beta\cos(\theta)\big). \tag{App.95}$$

Remarkably, this kernel is independent of the phase ϕ of the grating stimulus. Thus, complex cell populations possess good inductive bias for vision tasks where the target function only depends on the orientation of the stimulus rather than it's phase. In reality, V1 is a mixture of simple and complex cells. Let $s\in[0,1]$ represent the relative proportion of neurons which are simple cells and $(1-s)$ the relative proportion of complex cells. The kernel for the mixed V1 population is given by a simple convex combination of the simple and complex cell kernels

$$\begin{aligned}K_{V1}(\theta,\theta',\phi,\phi') &= \frac{1}{N}\sum_{i=1}^{N} r_i(\theta,\phi)r_i(\theta',\phi') \to \big\langle r(\theta,\phi,c)r(\theta',\phi',n)\big\rangle_{n\sim p_{V1}(n)}\\ &= s\big\langle r(\theta,\phi,n)r(\theta',\phi',n)\big\rangle_{n\sim p_s(n)} + (1-s)\big\langle r(\theta,\phi,n)r(\theta',\phi',n)\big\rangle_{n\sim p_c(n)}\\ &= sK_s(\theta,\theta',\phi,\phi') + (1-s)K_c(\theta,\theta'),\end{aligned} \tag{App.96}$$

where n denotes neuron type (simple vs complex, tuning etc) and $P_{V1}(n), p_s(n), p_c(n)$ are probability distributions over the V1 neuron identities, the simple cell identities and the complex cell identities respectively. Increasing s increases the phase dependence of the code by giving greater weight to the simple cell population. Decreasing s gives weight to the complex cell population, encouraging phase invariance of readouts.

5.5 Visualization of Feedforward Gabor V1 Model and Induced Kernels

Examples of the induced kernels for the Gabor-bank V1 model are provided in Figure App.4. We show how choice of rectifying nonlinearity $g(z)$ and sparsifying threshold a influence the kernel and their spectra. Learning curves for simple orientation tasks are provided.

6 Time Dependent Neural Codes

6.1 RNN Model and Decomposition

In this setting, the population code $\mathbf{r}(\{\boldsymbol{\theta}(t)\}, t)$ is a function of an input stimulus sequence $\boldsymbol{\theta}(t)$ and time t. In general the neural code $\mathbf{r}$ at time t can depend on the entire history of the stimulus input $\boldsymbol{\theta}(t')$ for $t' \leq t$, as is the case for recurrent neural networks. We denote dependence of a function f on $\boldsymbol{\theta}(t)$ in this causal manner with the notation $f(\{\boldsymbol{\theta}\}, t)$. In a learning task, a set of readout weights $\mathbf{w}$ are chosen so that a downstream linear readout $f(\{\boldsymbol{\theta}\}, t) = \mathbf{w} \cdot \mathbf{r}(\{\boldsymbol{\theta}\}, t)$ approximates a target sequence $y(\{\boldsymbol{\theta}\}, t)$ which maps input stimulus sequences to output scalar sequences. The quantity of interest is the generalization E_g, which in this case is an average over both input sequences and time, $E_g = \left\langle (y(\{\boldsymbol{\theta}\}, t) - f(\{\boldsymbol{\theta}\}, t))^2 \right\rangle_{\boldsymbol{\theta}(t), t}$. The average is computed over a distribution of input stimulus sequences $p(\boldsymbol{\theta}(t))$. To train the readout, $\mathbf{w}$, the network is given a sample of P stimulus sequences $\boldsymbol{\theta}^\mu(t), \mu = 1, ..., P$. For the μ-th training input sequence, the target system y is evaluated at a set of discrete time points $\mathcal{T}_\mu = \{t_1, t_2, ..., t_{|\mathcal{T}_\mu|}\}$ giving a collection of target values $\{y_t^\mu\}_{t \in \mathcal{T}_\mu}$ and a total dataset of size $\mathcal{P} = \sum_{\mu=1}^{P} |\mathcal{T}_\mu|$. The *average case generalization* computes a further average of the generalization error E_g over randomly sampled datasets of size $\mathcal{P}$.

Learning is again achieved through iterated weight updates with delta-rule form, but now have contributions from both sequence index and time $\Delta \mathbf{w} = \eta \sum_\mu \sum_{t \in \mathcal{T}_\mu} \mathbf{r}_t^\mu (y_t^\mu - f_t^\mu)$. As before, optimization of the readout weights is equivalent to kernel regression with a kernel that computes inner products of neural population vectors at different times t, t' for different input sequences $\{\boldsymbol{\theta}\}, \{\boldsymbol{\theta}'\}$: $K(\{\boldsymbol{\theta}\}, \{\boldsymbol{\theta}'\}, t, t') = \frac{1}{N} \mathbf{r}(\{\boldsymbol{\theta}\}, t) \cdot \mathbf{r}(\{\boldsymbol{\theta}\}, t')$. This kernel depends on details of the time varying population code including its recurrent intrinsic dynamics as well as its encoding of the time-varying input stimuli. The optimization problem and delta rule described above converge to the kernel regression solution for kernel gram matrix $K_{t,t'}^{\mu,\mu'} = \frac{1}{N} \mathbf{r}_t^\mu \cdot \mathbf{r}_{t'}^{\mu'}$ [20, 21, 22]. The learned function has the form $f(\{\boldsymbol{\theta}\}, t) = \sum_{\mu, t' \in \mathcal{T}_\mu} \alpha_t^\mu K(\{\boldsymbol{\theta}\}, \{\boldsymbol{\theta}\}^\mu, t, t')$, where $\boldsymbol{\alpha} = \mathbf{K}^+ \mathbf{y}$ for kernel gram matrix $\mathbf{K} \in \mathbb{R}^{\mathcal{P} \times \mathcal{P}}$ which is computed for the entire set of training sequences, and the vector $\mathbf{y} \in \mathbb{R}^{\mathcal{P}}$ is the vector containing the desired target outputs for each sequence. Assuming a probability distribution

over sequences $\boldsymbol{\theta}(t)$, the kernel can be diagonalized with orthonormal eigenfunctions $\psi_k(\{\boldsymbol{\theta}\}, t)$. Our theory carries over from the static case: kernels whose top eigenfunctions have high alignment with the target dynamical system $y(\{\boldsymbol{\theta}\}, t)$ will achieve the best average case generalization performance.

Appendix Figures

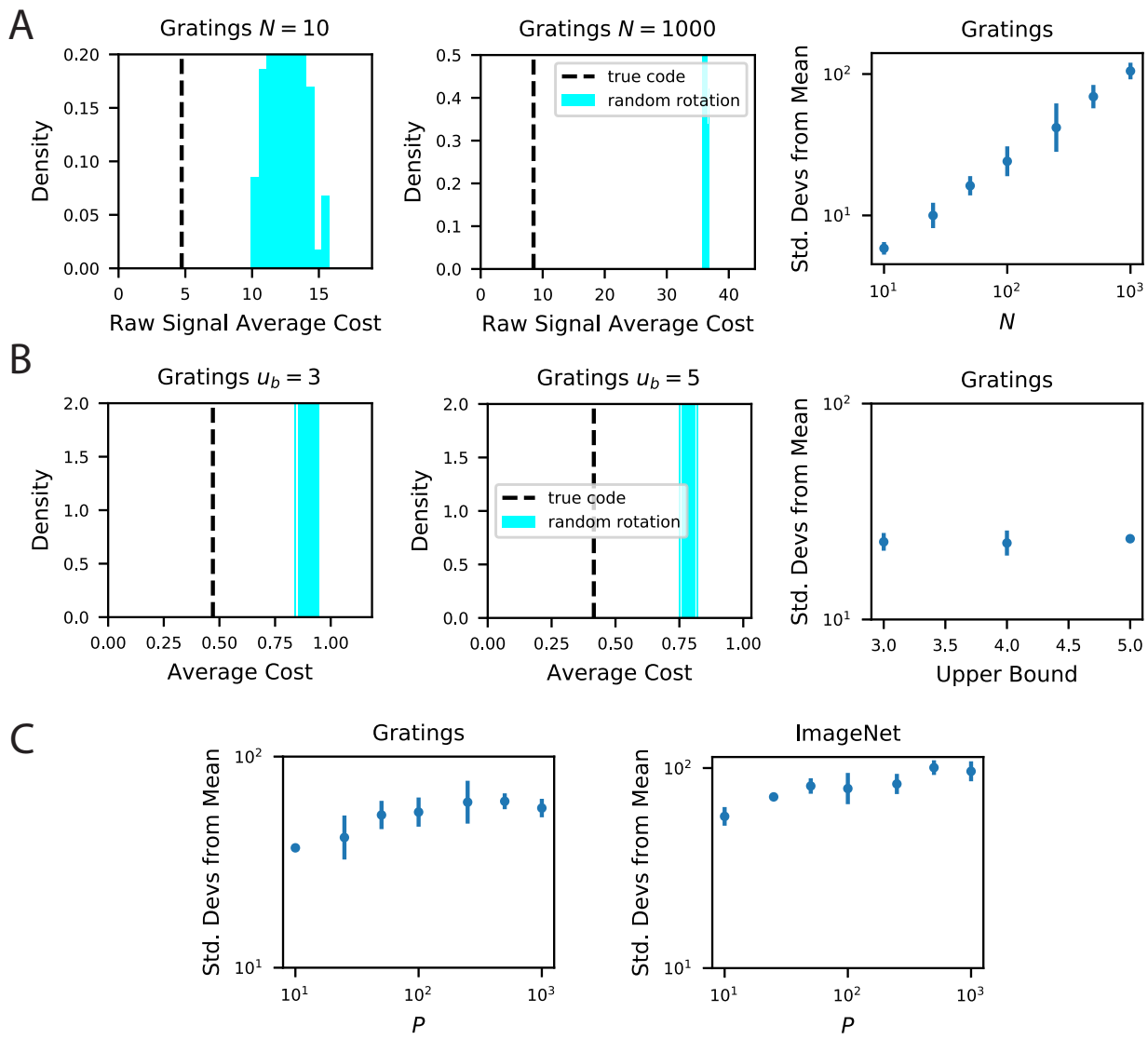

Figure App.1: Our metabolic efficiency finding is robust to different pre-processing techniques and upper bounds on neural firing. **A** We show the same result as the main text except we use raw (non z-scored) estimate of responses for each stimulus. **B** Our result is robust to imposition of firing rate upper bounds u_b on each neuron. This result uses the z-scored responses to be consistent with the rest of the paper. The biological code achieves a maximum z-score values in the range $[3.2, 4.7]$, which motivated the range of our tested upper bound values $\{3, 4, 5\}$. **C** Our finding is robust to the number of sampled stimuli P as we show in an experiment where rotations in $N = 500$ dimensional subspace.

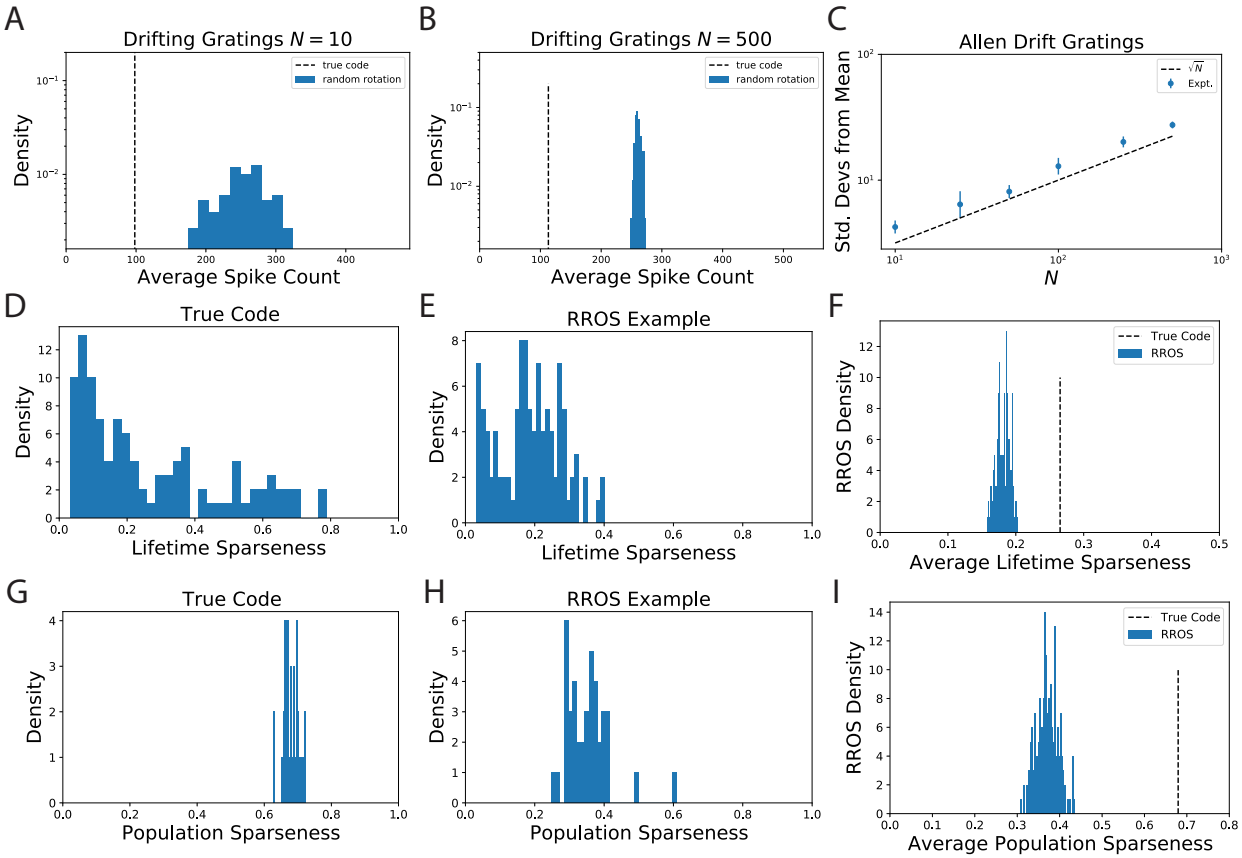

Figure App.2: The observation that randomly oriented codes with the same kernel require higher spike counts than the original code is reproduced from electrophysiological recordings of Mouse visual cortex (VISp and VISal) from the AIBO. **A** Distribution of average spike count for a random selection of $N = 10$ neurons. **B** The same for $N = 500$ randomly selected neurons. **C** Since the distribution of average spike count for the randomly rotated codes concentrates with N the number of standard deviations the true code is from the mean increases with N. We show that this scaling is approximately like $\sqrt{N}$ which suggests that the variability in average spike count for randomly rotated codes goes scales with neuron count like $\frac{1}{\sqrt{N}}$. This is intuitive since the average spike count over the rotated code is an empirical average over N random variables. **D** The lifetime sparseness of neurons in the true code are spread out over a large range. **E** The lifetime sparseness distribution for an example RROS code does not have the same range. **F** The average (over neurons) lifetime sparseness of RROS codes over random rotations is significantly lower than the average lifetime sparseness of the true code. **G-I** The true visual cortex code has much higher population sparseness over each grating stimulus as well

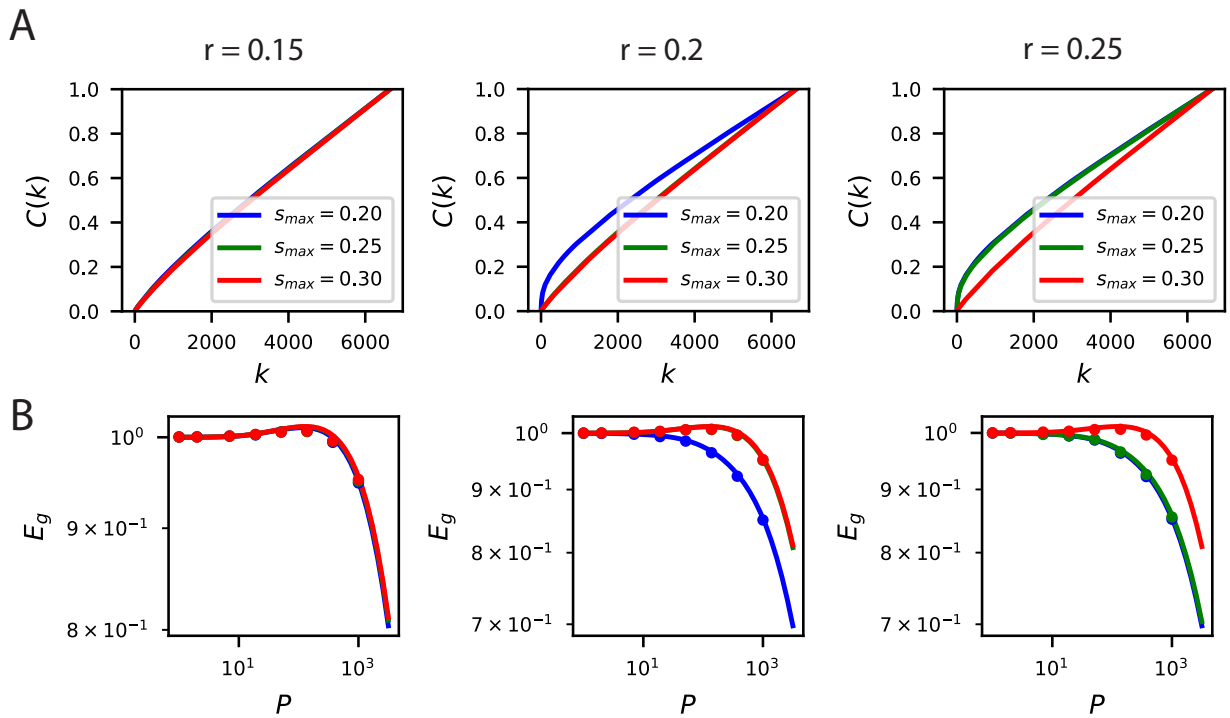

Figure App.3: The inductive bias to reconstruct low spatial frequency components of natural scenes from the population responses holds over several band-pass filters of the form $|\mathbf{k}| \in \left[\max(\sqrt{s_{max}^2 - r^2}, 0), s_{max}\right]$. **A** The cumulative power curves for reconstruction of the filtered images with different values of c and s_{max}. Lower values of s_{max} and higher values of c preserve mostly low frequency content in the images and are easier to reconstruct from the neural responses. **B** The learning curves respect the ordering of the cumulative power metric $C(k)$.

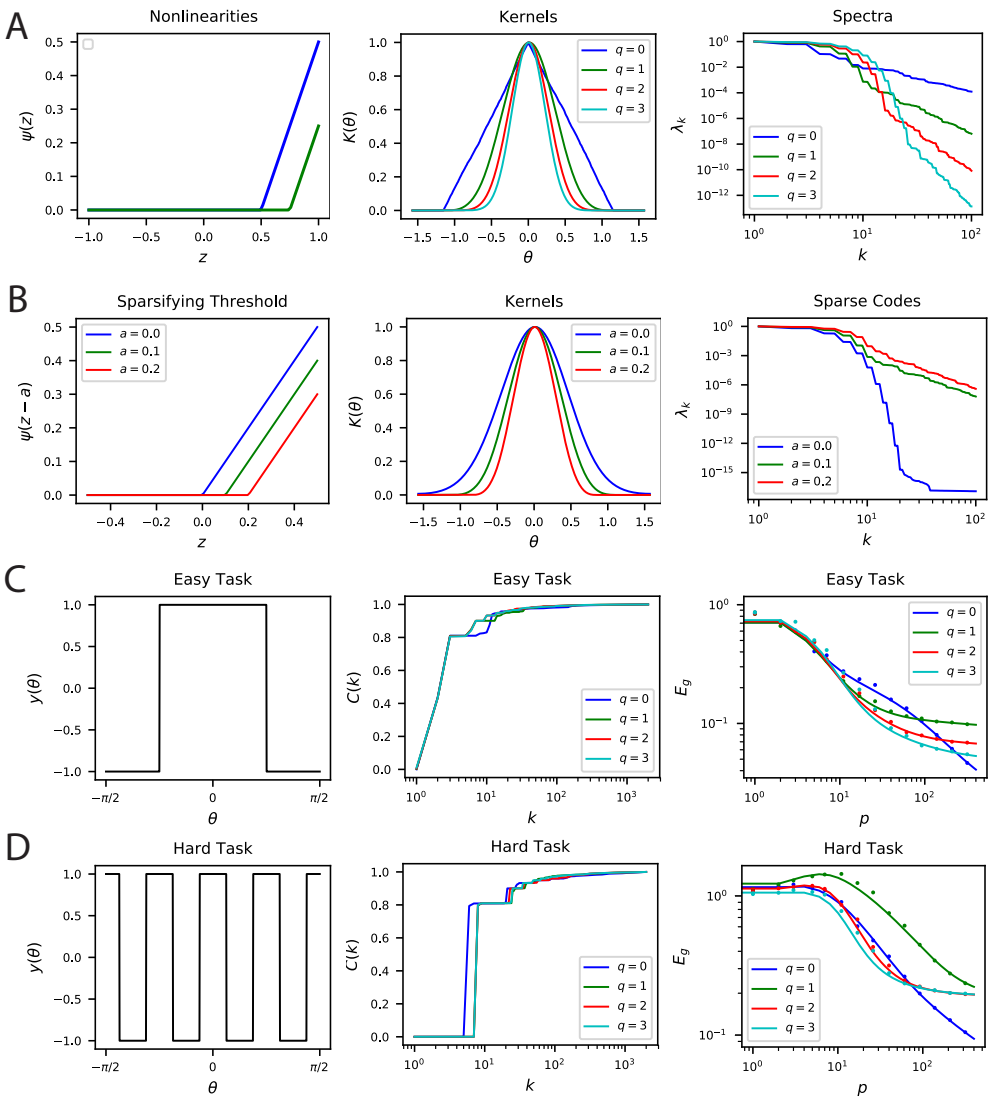

Figure App.4: Nonlinear Rectification and proportion of simple and complex cells influences the inductive bias of the population code. **A** The choice of nonlinearity has influence on the kernel and its spectrum. If the nonlinearity is $g(z) = \max(0, z^q)$, then $\lambda_k \sim k^{-2n-2}$. **B** The sparsity can be increased by shifting the nonlinearity $g(z) \to g(z-a)$. Sparser codes have higher dimensionality. Note that $a = 0$ is a special case where the neurons behave in the linear regime for all inputs θ since the currents $\mathbf{w} \cdot \mathbf{h}$ are positive. Thus, for $a = 0$, the spectrum decays like a Bessel Function $\lambda_k = I_k(\beta)$. **C-D** Easy and hard orientation discrimination tasks with varying nonlinear polynomial order q. At low sample sizes, large q performs better, whereas at large P, the step function nonlinearity $q = 0$ achieves the best performance.

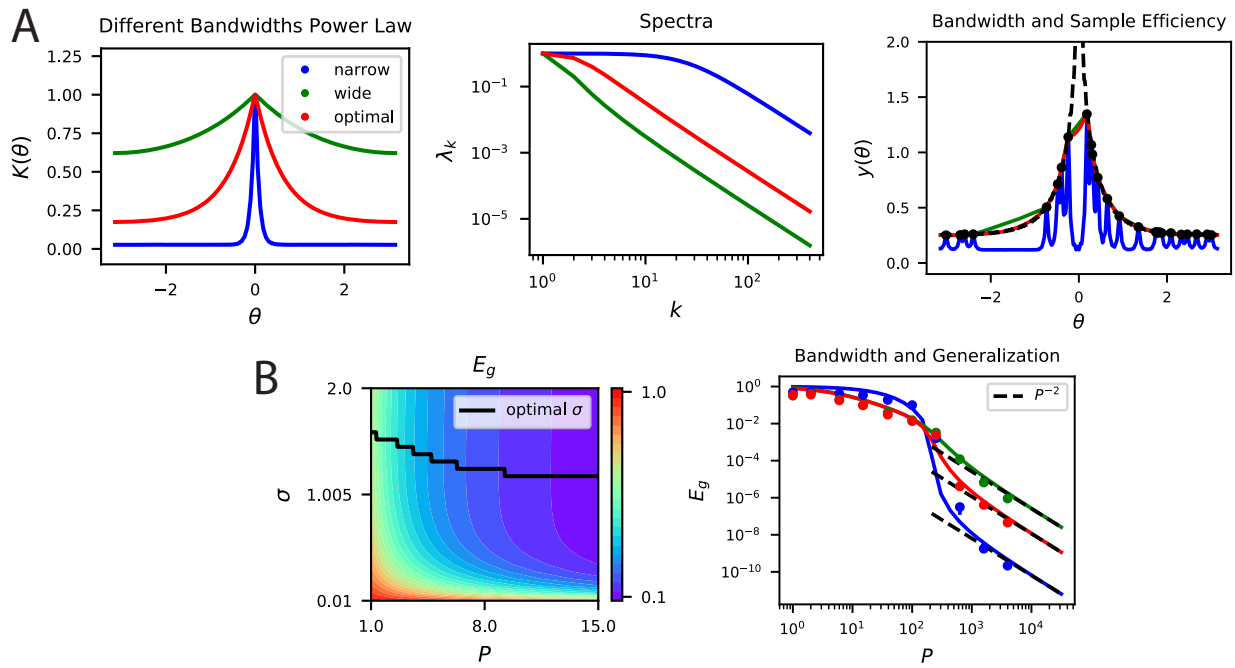

Figure App.5: **A, B** Kernel regression experiments are performed with Laplace kernels of varying bandwidth on a non-differentiable target function. The top eigenvalues are modified by changing the bandwidth, but the asymptotic power law scaling is preserved. Generalization at low P is shown in the contour plot while the large P scaling is provided in the generalization. In A-right and B-right, color code is the same as in the main text.

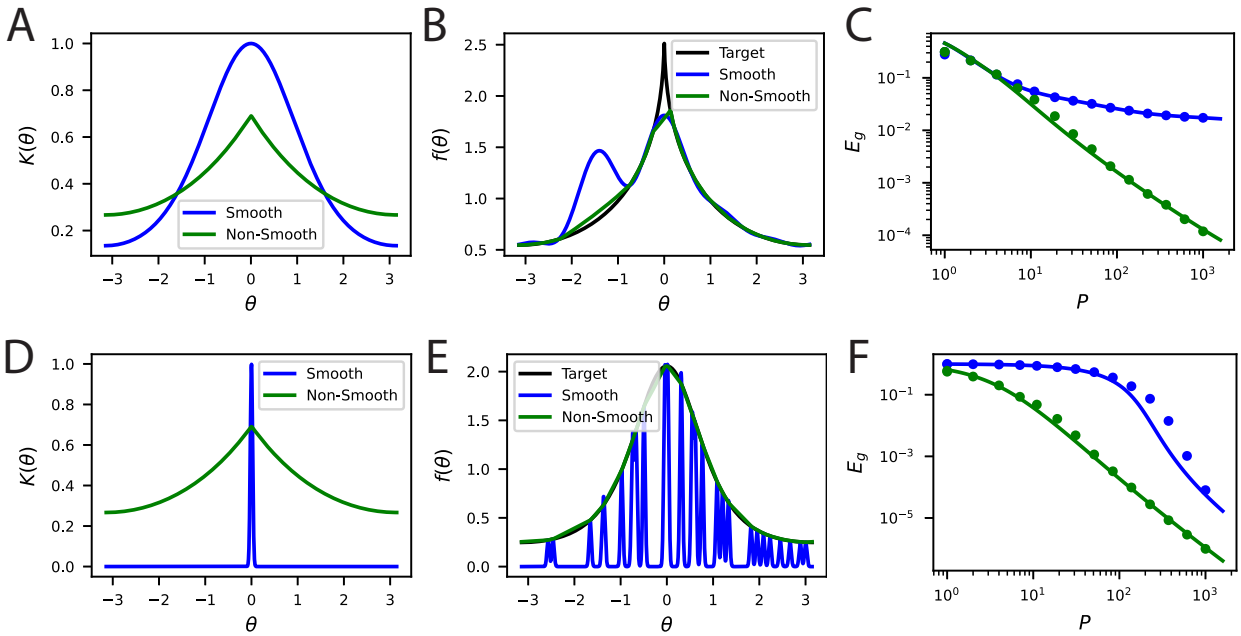

Figure App.6: Non-differentiable kernels can generalize better than infinitely differentiable kernels in a variety of contexts. **A** An infinitely differentiable von-Mises kernel (blue) will be compared to a non-differentiable Laplace kernel (green). **B** For a non-differentiable target function $y(\boldsymbol{\theta})$ with a cusp, the non-smooth kernel can provide a better fit to the target function at $P = 30$ samples. **C** The learning curves for this task. Solid lines are theory, circles are simulations. **D** The lengthscale of the kernel can be more important than local smoothness. We will now compare a narrow von-Mises kernel (still infinitely differentiable) with a non-differentiable wide kernel. **E** The non-smooth kernel generalizes on a smooth task almost perfectly whereas the narrow smooth kernel only locally interpolates. **F** The learning curves for this smooth task.

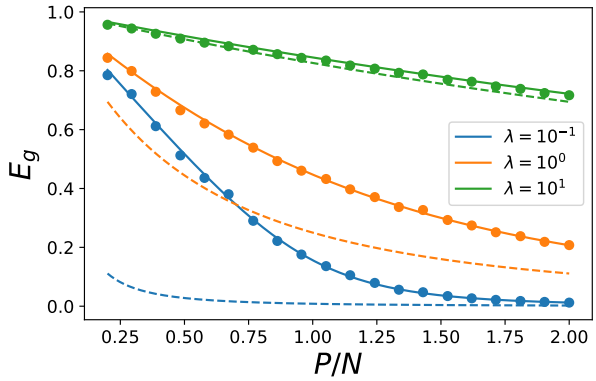

Figure App.7: Our theoretical prediction for generalization error E_g (solid lines) on a linear kernel $K(\mathbf{x}, \mathbf{x}') = \mathbf{x} \cdot \mathbf{x}'$ regression problem with uncorrelated Gaussian design $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$ and linear teacher $y = \beta \cdot \mathbf{x}$ in $N = 500$ dimensions. The generalization error as a function of sample size P is plotted predicted by the equivalent kernel/Wiener filter (dashed). Each color represents a different regularization level λ. The experimental generalization is shown as dots of the corresponding color and shows the numerical error obtained from solving the kernel regression problem with regularization level λ. We see that the two theories coincide at large λ but are drastically different at low levels of explicit regularization λ.

References

[1] Alex Townsend and Lloyd N. Trefethen. Continuous analogues of matrix factorizations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2173):20140585, 2015.

[2] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[3] Harold W Kuhn and Albert W Tucker. Nonlinear programming. In *Traces and emergence of nonlinear programming*, pages 247–258. Springer, 2014.

[4] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[5] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation, 2020.

[6] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[7] Bernard Widrow and Marcian E. Hoff. Adaptive switching circuits. In *1960 IRE WESCON Convention Record, Part 4*, pages 96–104, New York, 1960. IRE.

[8] John Hertz, Anders Krough, and Richard Palmer. *Introduction To The Theory Of Neural Computation*, volume 44. 01 1991.

[9] Bernhard Schölkopf, Ralf Herbrich, and Alex J. Smola. A generalized representer theorem. In *Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, COLT '01/EuroCOLT '01, page 416–426, Berlin, Heidelberg, 2001. Springer-Verlag.

[10] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, in press, 2021.

[11] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1024–1034. PMLR, 13–18 Jul 2020.

[12] Peter Sollich. Learning curves for gaussian processes. In *Neurips*, 1998.

[13] Peter Sollich. Gaussian process regression with mismatched models. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14, pages 519–526. MIT Press, 2002.

[14] Thomas Kailath. Correlation detection of signals perturbed by a random channel. *IRE Transactions on Information Theory*, 6(3):361–366, 1960.

[15] Simon Haykin. Adaptive filter theory. prentice-hall, 1991.

[16] Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

[17] E. Adelson and J. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America. A, Optics and image science*, 2 2:284–99, 1985.

[18] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[19] Oleg Rumyantsev, Jérôme Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark Schnitzer. Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580:1–6, 04 2020.

[20] Jonathan Dong, Ruben Ohana, Mushegh Rafayelyan, and Florent Krzakala. Reservoir computing meets recurrent kernels and structured transforms, 2020.

[21] Greg Yang. Tensor programs i: Wide feedforward or recurrent neural networks of any architecture are gaussian processes, 2019.

[22] Greg Yang. Tensor programs ii: Neural tangent kernel for any architecture, 2020.