Compressed Coding in Shallow Neural Networks ?
Random Compressed Coding with Neurons

Simone Blanco Malerba[1], Mirko Pieropan[2], Yoram Burak[3,4], and Rava da Silveira[1,5,6]

[1]Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris
[2]Department of Applied Science and Technology (DISAT), Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino,
[3]Racah Institute of Physics, Hebrew University of Jerusalem, Jerusalem
[4]Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem, Jerusalem
[5]Institute of Molecular and Clinical Ophthalmology Basel, Basel
[6]Faculty of Science, University of Basel, Basel

October 24, 2020

Abstract

The brain continuously acquires and transmits information rapidly and efficiently, in spite of the fact that neurons are noisy channels connected in complex patterns. How neural populations achieve this task is far from being fully understood. Historically, neural selectivity to features of external stimuli has been described using simple and parametric 'tuning curves'. Nevertheless, powerful codes in terms of capacity and accuracy emerge when considering more complex tuning profiles, which exploit activity space of neurons more efficiently.

In order to study a neural code with these features, we analyzed analytically and numerically the coding properties of a feedforward neural network with random synaptic weights, where the information present in a first layer is transmitted to a second one of much smaller size. This lack of structure generates heterogeneous and ambiguous single cells responses, which are poorly studied in the literature but ubiquitous in real neural systems. We show that a balance between fine scale *local* errors and large scale *global* ones arises when changing the smoothness of neural responses. Finally, we provide a biologically plausible example of such a system, analyzing data from an arm posture control task in monkeys' motor cortex under the lens of our model. Overall, we show quantitatively how good coding properties emerge in neural systems with disordered connectivity structure and complex single cell responses.

Rework to streamline & make stronger

1 Introduction (Check Neuron limits vs Intro length)

The responses of neurons to features of ~~the~~ external stimuli are historically characterized through tuning curves. These selectivity profiles are often described with simple and smooth parametric functions, like Gaussians or sigmoids, both in theoretical and experimental studies [1, 2, 3, 4, 5]. For example, populations of direction selective cells in retina and visual cortex [? 6] encode the orientation of visual stimuli by means of a set of von Mises-like tuning curves centered on different preferred orientations. Similarly, head-direction cells encode the direction that the animal is facing by means of a population of neurons with different preferred orientations [?], neurons in motor cortex encode the direction of motion firing proportionally to the cosine of the angle with a preferred direction [7], ~~and so on~~. When considering the theoretical problem of 'efficient coding' [8, 9, 10] with noisy neurons, the parameters of these tuning curves, often homogeneous to the whole neural population, are optimized with respect to some loss function. Generally, this loss function is considered to be the error made by an ideal decoder in the stimulus estimate, or quantities related to it like the FI or the Mutual Information [11, 12, 13]. The type of errors in these coding schemes are often *local*, in the sense that the stimulus estimate is close to the true value.

+ constraint

?

Despite all of this, when thinking about the problem of coding in a more general way, this kind of coding schemes are not very efficient, due to the high redundancy in the responses. Using a geometric analogy, firstly introduced by Shannon [14], this kind of codes does not use efficiently the space of all possible *signals*, or pattern of neural responses, to represent the possible *messages*, or stimuli. A good code, in this sense, should map the messages so that they fill the space of signals, making full use of the capacity of the channels. On the other hand, noise creates a region of uncertainty around each point in the signal space, and this ultimately poses a limit on how well the space can be filled without creating disrupting ambiguities in the representation of the messages (what Shannon called the *threshold effect*).

In this sense, are there any more efficient codes in the brain? Grid cells in rats enthorinal cortex [?] (but not only, see [? ?]) fire periodically during the exploration of the environment, and, as a whole population, they represent the animal position in the space by mean of a set of phases within different periods. This coding scheme is highly ambiguous at single cell level, since integer multiples of the spatial period will evoke the same response, and the error can be non-local. Nevertheless, different studies [15, 16, 17] showed how this coding scheme efficiently exploits the space of activity of neurons, outperforming more classical population of neurons in terms of resolution and capacity by means of a reduction of the redundancy of neural representations. In contrast to the previously described coding schemes, which allow the error to decrease as an inverse power of the population size, in grid cells scheme the error decreases exponentially with the number of neurons. Similar geometrical arguments were brought to explain the existence and the advantage of neurons with *mixed selectivity*, that is neurons responding to a non linear combination of task parameters [18, 19, 20]. In all these cases, the information that a single cell (or small subsets) conveys may be highly ambiguous, and the correct information can be retrieved only considering the whole population.

Other than these still 'highly structured' examples, tuning curves inside the same population can show a high degree of heterogeneity and deviations from parametric description. As an extreme case, the authors of [21] found that tuning curves in the motor cortex are more compatible with a random model rather than with a simple parametric form. Interestingly, in the already cited paper, in the context of discrete messages, Shannon showed how a random association between messages and points in the signal space, satisfy the properties of space filling. Could the brain employ this strategy to build efficient neural representations? Can this idea be extended to continuous stimuli? Could we achieve the performance of grid cells with random tuning curves?

We addressed these questions by analyzing the coding properties of a biologically plausible model of a shallow (two-layer) neural network, where the population of neurons in the second layer possesses irregular tuning curves due to randomness in the input from the first layer. Irregularities in neurons' responses may improve the coding performances decorrelating the responses for different stimuli. At the same time, they introduce ambiguity at single cell level. With finite population sizes, there exists a trade-off between the local accuracy given by very irregular codes and the requirement of robustness satisfied with smoother tuning curves. We quantified analytically and numerically this trade-off in the case of one-dimensional stimulus, showing that this coding scheme allows the error to decrease exponentially with the population size. Similar results hold when the stimulus is multidimensional, with qualitative differences depending on how it is encoded in the first layer. Finally, we analyzed data from tuning curves in monkey's motor cortex of a previous experiment. We showed how irregularities, within biologically relevant region of parameters space, boost the coding performance of the population with respect to a classical population with smooth, parametric tuning curves. In the last section we discuss how the introduction of noise in the first layer can have a negative effect, but very small in the regime of interest, due to the structure imposed to the noise covariance matrix.

2 Results

The differences between the coding scheme of grid cells and a 'classical' one, with radially symmetric tuning curves, can be fully appreciated looking at the space of neural activity, Fig. 1a. A coding scheme can be viewed as a map between the stimulus space, in this case a 1-D line, into a neural manifold [1] embedded in the N-dimensional space of neurons' activity. Due to the large overlap

[1] We use the word *manifold* as it has recently become common in neuroscience to describe the topological space described by the neural activity. In our case, we consider the manifold described by the function which associates to each point of the stimulus space x a point in the space of the mean firing rates of the N neurons

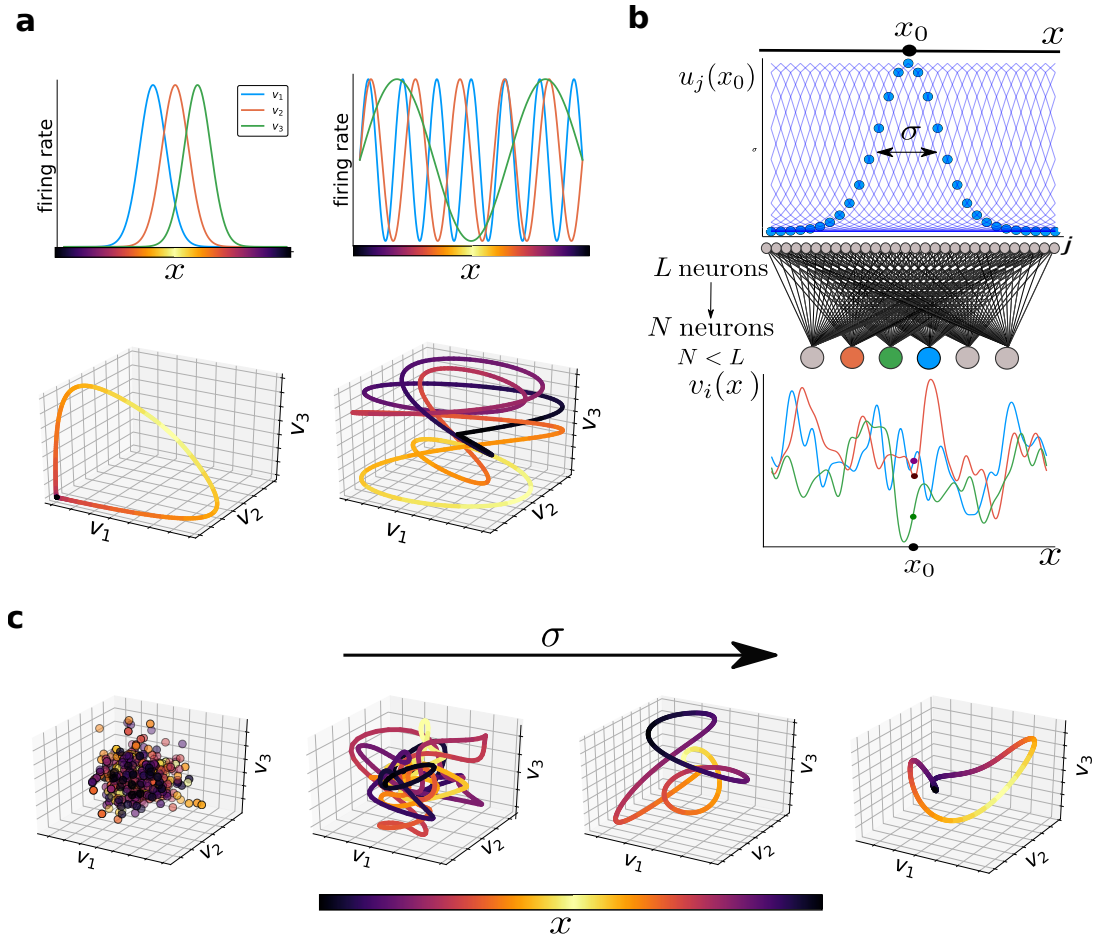

Figure 1: **Different types of encoding schemes and Random Feedforward Network.**(**a**) Differences in encoding a 1D stimulus with radially symmetric tuning functions (left) or periodic ones (grid cells like, right). Top: tuning curves of three sample neurons. Neurons with von Mises tuning curves fire only in a specific region of the stimulus with a substantial overlap of the tuning curves. Instead, grid cells fire periodically across all the stimulus space, conveying information about the phase in a given period (different for the three cells). Bottom: manifold in the activity space of the three neurons; color denotes value of x. Monomodal tuning curves will produce a smooth manifold, similar stimuli evoking similar responses. On the contrary, grid cells activity produces describe a twisted manifold, that does not preserve distance between stimuli (two different stimuli may evoke similar responses, and viceversa). With the addition of neural noise, this could produce large scale errors. At the same time, it 'fills' the activity space much more efficiently and this could allow to encode much more information with the same number of neurons. (**b**) Feedforward neural network. An array of L cells with Gaussian tuning curves (in light blue) encodes a 1-D stimulus into an high dimensional representation. A specific stimulus evokes a Gaussian 'bump' of activity (blue dots), of width σ. This layer projects onto a smaller layer of N cells with an all-to-all random connectivity matrix $\mathbf{W}$, producing irregular tuning curves, three examples are showed. At a single cell level, the responses to the stimulus x_0 (indicated as dots on the tuning curves) are highly ambiguous, since several other stimuli correspond to the same response). Narrower σ is, higher is the ambiguity; this ambiguity can be resolved only considering a whole population. [E' rimasta una sigma molto piccola sull'asse y.] (**c**) Typical shapes of the neural manifold in the activity space, in function of σ. Color denotes stimulus value. For $\sigma \to 0$ (extreme left), responses to different stimuli are uncorrelated, describing a very 'spiky' manifold (left panel). Increasing σ, responses to nearby stimuli start to be correlated and we obtain a smoother curve, still very twisty, similar to the grid cell picture. Increasing further σ, we eliminate the self-intersections and we re-obtain the picture of the Gaussian bump scheme, with a very smooth, ring-like, manifold.

between tuning curves, populations with radially symmetric tuning curves produce a very smooth manifold in the activity space, with similar stimuli represented by nearby points. On the other hand, grid cell–like tuning curves will produce a very twisty manifold, still locally smooth, due to the periodicity of the firing rates. This coding scheme is no more distance preserving, since two similar responses can be evoked by two very distant stimuli (and vice versa, two close stimuli may evoke quite different responses). These self-intersections, with the addition of neural noise that moves the response away from the manifold, are the cause of large scale errors in grid cell scheme. At the same time, one can observe that the activity space is much more filled and the resulting manifold is more 'stretched'. This, at equal population size, can be read as a sign of a higher local accuracy and better error correcting properties. In grid cells, the advantage emerges when the stimulus space is restricted such in a way that there are no more large scale ambiguities, allowing the error to decrease exponentially fast with the population size [16].

2.1 Model: shallow network with random weights

In order to see if these features emerge in a less structured scheme, we studied the coding properties of a population of neurons whose tuning curves heterogeneity arise from randomness in the connectivity structure. Our model consists in a convergent two-layer feedforward neural network, Fig. 1b. The first one is a sensory layer, made of L neurons encoding a one-dimensional stimulus x into a high-dimensional representation by means of Gaussian tuning curves of width σ, each of which is centered on a preferred stimulus c_j of their associated neuron. The preferred stimuli are arranged uniformly to cover the stimulus space (we assumed a uniform prior on the stimulus). As a result, the response vector for a specific stimulus $\mathbf{u}(x_0)$ can be represented as a Gaussian 'bump' of activity centered at x_0.

$$\mathbf{u}(x_0) = \{\frac{1}{Z}\exp\Big(-\frac{(x_0-c_j)^2}{2\sigma^2}\Big) \; with \; c_j = \frac{j}{L}\} \tag{1}$$

This population projects onto a smaller layer of N neurons with an all-to-all connectivity matrix $\mathbf{W}$ with randomly distributed weights. The connectivity structure causes second layer neurons to be sensitive to all the stimulus space, but with an irregular sensitivity profile.

$$v_i(x) = \sum_{j=1}^{L} W_{ij} u_j(x). \tag{2}$$

In Methods we show that under the hypothesis of Gaussian weights, these tuning curves can be modeled as samples from a Gaussian process with squared exponential correlation function. One parameter, σ, governs the smoothness of these tuning curves, that is how far two similar stimuli evoke similar responses. The normalization constant Z is set such that neurons have a constant variance of the responses across the stimulus space. The model produces negative responses; since firing rates are positive by definition, these quantities can be considered deviations from a baseline firing rate. Moreover the model can be enriched passing the random sum through a non linearity, but this complicates the analytical results and does not change qualitatively the behavior.

We kept the model intentionally simple to illustrate concepts and to allow an analytical approach. In particular, we considered the coding properties of the neurons of the second layer when they are affected by isotropic Gaussian noise (see Methods), describing the trial to trial variability as

$$\mathbf{r} = \mathbf{v}(x) + \mathbf{z}, \tag{3}$$

with $\mathbf{z} \sim \mathcal{N}(0, \eta^2\mathbf{I})$. We will discuss the implications of this choice, but introducing plausible levels of input noise does not qualitatively change the results in the considered regimes ($N \ll L$).

The N heterogeneous tuning curves describe the trajectory of the manifold in the activity space as we vary the stimulus. By changing σ, we change the typical shape of this manifold, Fig. 1e. For $\sigma \to 0$, responses to different stimuli are uncorrelated and the manifold is very scattered. The map does not preserve any distance: two stimuli, no matter how close they are, will evoke completely different responses and, vice versa, two close responses can represent two completely different stimuli. Increasing σ, we find that the manifold becomes smoother since similar stimuli

$\mathcal{M} = \{\mathbf{v} \in \mathbb{R}^N | \mathbf{v} = \mathbf{v}(x), x \in S\}$. where S is the space of stimuli and $\mathbf{v}$ is the vector of mean responses to the stimulus (tuning curves). We suppose this map sufficiently smooth so that the topological space locally resembles the Euclidean space.

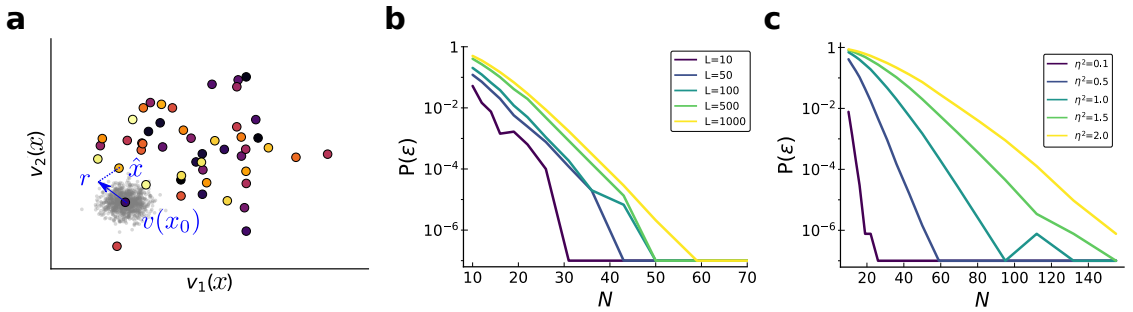

Figure 2: **Error probability for discontinuous random responses.** (**a**) Responses to $L = 50$ stimuli in the activity space of two neurons, color denoting stimulus value. Noise is represented as a cloud of possible responses (in grey) around the true one. We have an error when the noisy response $\mathbf{r}$ happens to be closer to a point representing another stimulus $\hat{x}$ than the true one x_0. Since responses are uncorrelated, that point may represent a very different stimulus. (**b**) Numerical results for the probability of error in function of population size for different numbers of discrete stimuli encoded with uncorrelated random responses ($\eta^2 = 0.5$). The probability scales exponentially with population size, with a prefactor depending on the number of stimuli L. (**c**) Probability of error in function of population size for $L = 500$ discrete stimuli and different noise magnitudes. The exponent of the exponential scaling depends on the noise variance η^2.

evoke similar responses, but can still show a lot of self-intersections. This picture is very similar to grid cells' one, with a locally smooth manifold very twisted in the activity space (Fig. 1a, right bottom panel). When σ is very large, the tuning curves in the second layer become monomodal and we obtain a ring-like manifold, similar to the bottom left panel of Fig. 1a.

What are the coding properties of this population of neurons? How can we solve the trade-off between the local accuracy of uncorrelated responses and the robustness of a smooth code? In the following, we present analytical and numerical results for the specific case where the weights are normally distributed. We used the Root Mean Square Error (RMSE) in the stimulus estimate by an ideal decoder (see Methods) as a measure for the coding performance of the population, a loss function well established in Information Theory [?] and neuroscience [?, ?]. In particular, we used an ideal decoder which minimizes the MSE computing the average of posterior distribution, and has a simple neural network implementation. Other tools of Information Theory, like Fisher Information (FI) and Mutual Information (MI), have been used to measure the coding properties of populations of neurons, often justified by the Cramer Rao bound that sets a limit to the variance of an unbiased estimator. Nevertheless, this is a typical case where these tools give an incomplete, if not misleading, description [12, 13]. Indeed FI is a local measure that will not keep into account large scale errors. Mutual Information is a more global quantity, but still it will fail in keeping into account errors' magnitudes.

2.2 Extremely narrow tuning curves: global errors

While the coding properties of a population of neurons with Gaussian tuning curves (and, more generally, radially symmetric functions) have been largely addressed in the literature, it is instructive to see what happens when $\sigma \to 0$. In this case, the neurons of the first layer respond only to their preferred stimuli, and the response of the second layer neurons for two different stimuli are uncorrelated.

Let's switch for a moment to the case where we have L discrete (ordered) stimuli evoking random uncorrelated responses, with constant variance across different stimuli:

$$v_i(x_j) \sim \mathcal{N}(0,1) \quad with\ x_j = \frac{j}{L} \tag{4}$$

In the activity space this result as an ensemble of scattered points, Fig. 2a, and the noisy responses can be thought as a cloud of points around the true one. A Maximum Likelihood decoder will associate a response to the stimulus corresponding to the closest point. In case this point is the right one, the inference will be correct and the error will be 0. Nevertheless, since responses are uncorrelated, the closest point may represent a completely different stimulus, resulting in a

potentially very large error. Given that the error magnitude is a term of order 1 (the size of the stimulus space), all the coding properties are incorporated in the probability of having such an error. In Methods we computed this probability (averaged over the distribution of the synaptic matrix) as a function of the number of discrete stimuli L, noise variance η^2 and population size N, obtaining the approximated scaling

$$\langle P(\varepsilon)\rangle_W \approx \frac{L}{\sqrt{2\pi N}} \exp\Big(-\frac{N}{2}\log\big(\frac{1+2\eta^2}{2\eta^2}\big)\Big). \tag{5}$$

The probability of error decreases exponentially with the population size, with a prefactor given by the number of discrete stimuli L, Fig. 2b. The slope of this exponential scaling is given by the noise variance, approaching the value of $-\frac{1}{4\eta^2}$ as the variance of the noise becomes larger than the dynamic range of the neurons (defined as the variance of the responses across stimuli), Fig. 2c. A random uncorrelated coding scheme therefore allows an exponential decrease of the error as a function of the number of neurons, but at the price of losing the notion of similarity between stimuli and causing potentially very large errors if the population size is not large enough (or the noise is too large). In a more realistic framework, neurons have a certain degree of smoothness in their selectivity profile, and this is obtained increasing the tuning width σ; this allow us to return to consider the stimulus as a continuous variable.

The general case of broad t.c. : tradeoff between

2.3 ~~Broad tuning curves:~~ local and global errors

In a smooth manifold, a noisy response to the same stimulus may cause two types of qualitatively different errors, Fig. 3a. At a given trial, the response may fall close to a point of the manifold representing a similar stimulus $\hat{x}^I$, causing a small *local* error. This kind of errors are the result of the projection of the noise vector onto the neural manifold, which can be thought as locally linear. Lower values of σ will 'stretch' the manifold, allowing a better resolution. If the manifold is twisted, the noisy response can still happen to be close to a point which represents a completely different stimulus $\hat{x}^{II}$, causing a large scale *global* error. The magnitude of this kind of errors depends on the specific realization of the synaptic matrix, but the probability of having them decreases when we reduce the number of loops. Therefore, increasing the length of the manifold, we increase the local accuracy, but also the number of loops, see Fig. 1c. This trade-off is well explained by the histogram of errors for different widths, Fig. 3b. When σ is very low, the local accuracy is high, and the local errors are of small magnitude. At the same time, the tail of the histogram, representing the probability of large errors, is substantial due to the high number of self-intersections in the manifold. Note that the tail is almost flat, showing the random magnitude of global errors, which are uniformly distributed up to a maximum value. Increasing σ, we lower the tail of the histogram since the manifold is smoother, but the magnitude and the number of local errors increase.

In Methods we formalized this intuition, showing that the average MSE can be approximated as the sum of two contributions

$$\langle \varepsilon^2\rangle_W = \langle \varepsilon_l^2 + \varepsilon_g^2\rangle_W \approx \frac{2\sigma^2\eta^2}{N} + \frac{1}{\sigma\sqrt{2\pi N}}\bar{\varepsilon}_g \exp\Big(-\log\big(\frac{1+2\eta^2}{2\eta^2}\big)\frac{N}{2}\Big), \tag{6}$$

where $\bar{\varepsilon}_g$ is a term of order 1 that depends from the specific stimulus geometry (boundary conditions).

Do not have short & like this.

For limited population sizes and with plausible levels of noise, the two terms are of the same order of magnitude and it is possible to tune σ to achieve a balance, Fig. 3c. Intuitively, it is possible to give an interpretation to the two terms and their scaling. The local error is simply proportional to σ and inversely proportional to N, as the classical error in population of neurons with gaussian tuning curves. Stretching the manifold (decreasing the width) we increase the local accuracy because we allow a better discrimination of nearby stimuli. The second part is a generalization of the error in the case of uncorrelated responses, Eq.(5): the number of uncorrelated manifold segments is inversely proportional to σ. We tested the analytical prediction of Eq.(6) with numerical simulations: we estimated the MSE with a Monte Carlo method, decoding the noisy responses with an ideal decoder (see Methods) and obtaining a good agreement with the theory, Fig. 3d. The range of σ we analyzed was determined by the following considerations. The minimal σ we considered is the spacing between preferred positions, $\sigma_{min} \sim \frac{1}{L} = 0.002$; below this value, we start to have coverage problems (see the following section). The maximum σ was chosen

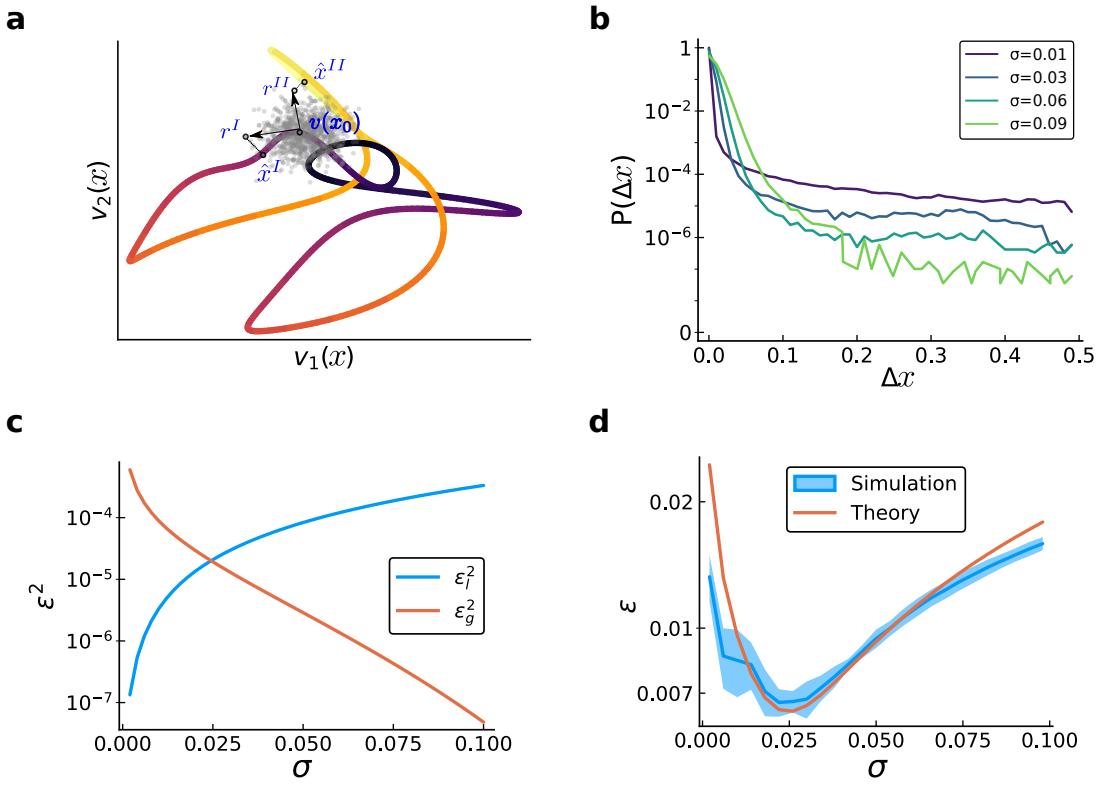

Figure 3: **Balance between local and global errors.** In all simulations, $L = 500$ and $\eta^2 = 0.5$. (**a**) Different types of error in self-intersecting neural manifold (two neurons, color representing stimulus value as in the previous figures). $\mathbf{r}^I$ and $\mathbf{r}^{II}$ are two possible noisy responses to the same stimulus, extracted from the Gaussian cloud surrounding the true response $\mathbf{v}(x_0)$. An ideal decoder will output the stimulus corresponding to the closest point of the manifold. $\mathbf{r}^I$ will cause a local error, falling on a point of the manifold that represents a similar stimulus. Instead, due to self-intersections, $\mathbf{r}^{II}$ happens to be close to a point of the manifold which represents a stimulus quite far from the true one, causing a global error. (**b**) Normalized histogram of errors $\Delta x = |\hat{x} - x|$ made by an ideal decoder, for different values of σ ($N = 25$). We tested the response to 500 stimuli uniformly spaced between $[0, 1]$, 50 trials (noise realization) per stimulus. The histogram is obtained averaging over 8 different realizations of the connectivity matrix. For a better visualization, we considered a stimulus with periodic boundary conditions, such that all global errors magnitudes have the same probability. Contributions of the two types of error change changing σ. For small σ, coding is very precise locally (fast drop of the purple curve for small errors), but we have a great number of global errors (tail of the distribution is quite high). Vice versa, smoother codes yield to poor local accuracy (larger local errors), but an high noise robustness (very few large scale errors). (**c**) Theoretical prediction for the two contributions to the MSE (log scale) in function of σ ($N = 30$). The magnitude of local errors increases with larger widths (blue curve), while the number of global errors decreases (red curve). (**d**) RMSE (log scale) in function of σ: comparison between numerical simulations (blue curve) and theoretical prediction of Eq.(6). Numerical results are obtained averaging over 8 network realizations (shaded region corresponding to 1 s.d.).

such that the Gaussian process had at least one 0-crossing in the stimulus space, corresponding to a single peaked tuning curve: $\sigma_{max} = (2\pi * \sqrt{2})^{-1} \sim 0.1$.

The error curve around the optimal width σ^* is strongly asymmetric. A smaller width will cause a rapid increase of the error due to the weight of global errors, while a broader σ will just lower the local accuracy. This minimum can be thought as the longest manifold of this kind that we can have such that the number of errors coming from the loops are compensated by the local accuracy. How the optimal coding properties of the network scale with population size? We analyzed the error curves in function of σ for different values of N, Fig. 4a. As expected, as soon as the population grows, we can allow a lower σ due to the suppression of the second term in Eq.(6). Analytically, we can obtain the scaling of the optimal width, showing that is exponential in the population size and confirming the results obtained from numerical simulations, Fig. 4c,

$$\sigma^* \approx \Big(\sqrt{\frac{N}{2\pi}}\frac{\bar{\varepsilon}_g}{4\eta^2}\Big)^{1/3} \exp\Big(-\log\big(\frac{1+2\eta^2}{2\eta^2}\big)\frac{N}{6}\Big). \tag{7}$$

Due to the lower bound imposed by the spacing of the preferred positions in the first layer, the optimal σ^* saturates at a finite value. As a consequence, also the optimal error scales exponentially in the number of neurons, Fig. 4d,

$$\varepsilon(\sigma^*) \approx \Big(\frac{\eta\bar{\varepsilon}_g^{2/3}}{(\sqrt{2\pi}N)^{2/3}}\Big) \exp\Big(-\log\big(\frac{1+2\eta^2}{2\eta^2}\big)\frac{N}{3}\Big). \tag{8}$$

In both cases, the slope of the scaling is given by the variance of the noise. Therefore, the optimal width and the error depend from the value of the pair $N - \eta^2$, Fig. 4e,f. Finally, it is worth to see what happen when the width is kept fixed and the number of neurons increases, Fig. 4b. The exponential scaling holds until a critical value of $N^*(\sigma)$. Then, the error simply decreases linearly due to the contribution of local errors. Decreasing σ, we increase the critical N, but at the same time the error at which this transition occurs is lower. In the following section we extend the model to multidimensional stimuli, showing that the way they are encoded in the first layer affects also the balance between local and global errors in the second one.

2.4 Multidimensional Stimuli

Real world stimuli are high dimensional. The previous model can be easily extended to analyze the case of higher dimensional stimuli, and its properties will vary depending on how the stimulus is encoded in the first layer. Sensory neurons encode multidimensional stimuli in several ways, and we describe briefly two extreme cases, keeping in mind that the truth lies in between. On one hand, neurons can be tuned to all dimensions of the stimulus. A first example are place cells, which respond to a specific location in the two dimensional space. In retina and cortex, Direction Selective cells can be tuned to a specific combination of direction and velocity. Such a coding scheme becomes rapidly inefficient as the dimensionality increases, since the number of neurons required to cover the stimulus space grows exponentially. At the same time, it allows more information to be represented by the same neuron. In the other extreme case, neurons are tuned to a single feature only, and insensitive to the others. For example neurons can encode the color of an object disregarding other features like shape, orientation, etc...

We will call the two coding schemes *conjunctive* and *pure*, respectively, following the notation of [22]. In this work, the difference between the two schemes is explored in the context of head direction in bats. The result is that the relative advantage of pure populations (encoding just one angle of head direction) with respect to conjunctive ones (encoding both angles together) changes, depending on the specific constraints on the time available for the decoding and on population size. We enriched this description studying the properties of tuning curves arising from the random projection of these populations; both cases have a corresponding in the literature, and they show a qualitative difference in the balance between local and global errors. The case of 3-D stimulus is of particular interest, since it allows us to make a connection with experimental data; we will therefore present numerical results for this case, referring to Methods for the general case of K-dimensions.

In the pure case, a 3-D stimulus is encoded in the first layer by a population of M neurons, with $L = M/3$ of them monitoring each dimension of the stimulus with 1-D Gaussian tuning curves. In the conjunctive case, the M neurons are sensitive to all dimensions and their tuning curves are multidimensional Gaussian centered on a preferred position in the 3-D space. The population of the first layer randomly projects onto the second layer. We enforced the same constraint as before,

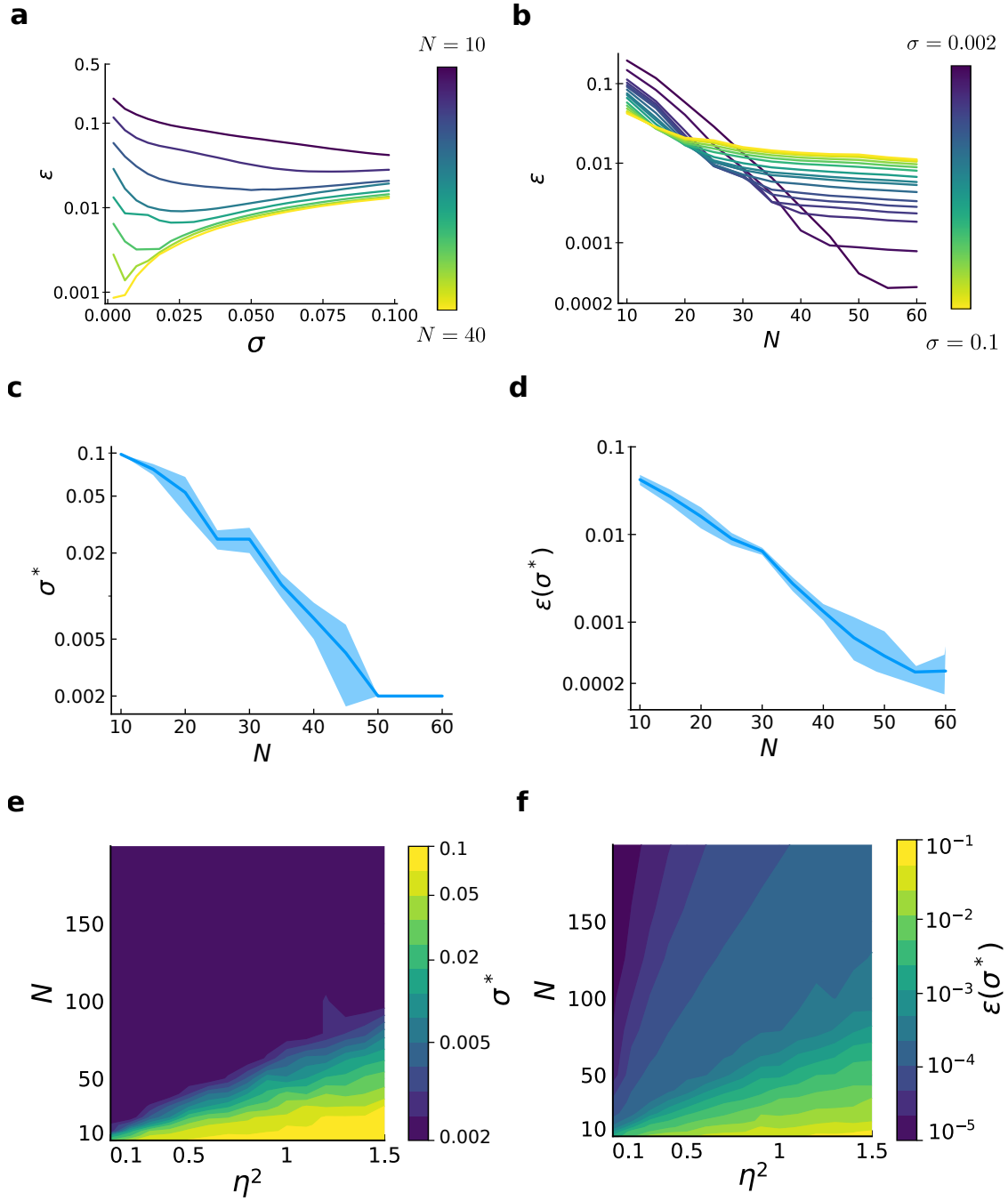

Figure 4: **Numerical results for the scaling of error in a feedforward network with random weights.** In all simulations $L = 500$. (**a**) Error (RMSE, log scale) in function of σ for different population sizes N (increasing from violet to yellow), averaged over 4 network realizations, $\eta^2 = 0.5$. The optimal error is attained at the optimal $\sigma^*(N)$ that decrease increasing N. (**b**) Same data, but the error is showed in function of N, for a fixed σ. The error at first decrease exponentially fast until all global errors are suppressed, then the finite width makes the error decrease linearly. Decreasing σ, we increase the N at which the transition happen, but also the error at this critical value. (**c**) The mean optimal σ^* decreases exponentially fast with the number of neurons, saturating the lower bound imposed by the finite number of neurons of the first layer. Shaded region correspond to 1 s.d. This causes the optimal error in (**d**), $\varepsilon(\sigma^*)$, which is linear in σ, to also decrease exponentially fast with the population size. (**e,f**) Optimal width and error in function of the pair population size - noise variance. Values are in log scale, such that it is possible to appreciate the exponential scaling.

keeping constant the variance of the responses across all stimuli in the second layer. We considered the scalar error in the stimulus estimate $\varepsilon^2 = \varepsilon_x^2 + \varepsilon_y^2 + \varepsilon_z^2$, although others loss functions can be relevant in the context of multidimensional stimuli. Extending the previous theory, analytically

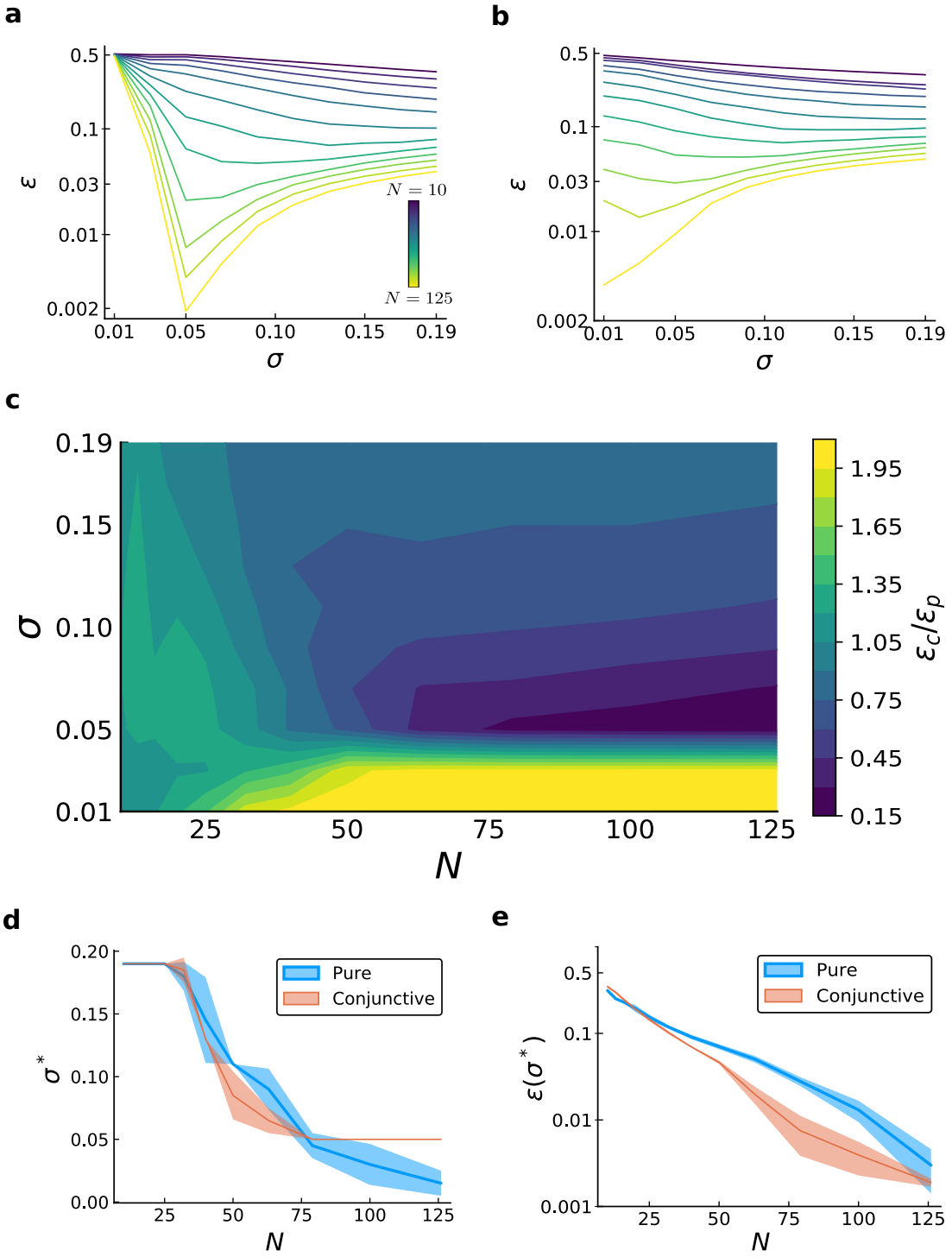

Figure 5: **Numerical results for the case of 3D stimulus, for conjunctive and pure populations in the first layer.** In all simulations $M = 3375$ and $\eta^2 = 1$. *continue to next page*

Figure 5 *(previous page)*: (**a,b**) Error in function of σ for different population sizes N (increasing from violet to blue), when the first layer is made by conjunctive (a) or pure cells (b), averaged over 4 network realizations. An optimal σ, decreasing with N, allows the balance between local and global errors, similarly to the 1D case. In the conjunctive case the rapid increase of the error below $\sigma = 0.05$ is due to the loss of coverage and is independent from N. (**c**) Mean ratio between the error in the two cases $\varepsilon_c/\varepsilon_p$, in function of σ and population size. Yellow (violet) region indicates an outperformance of the pure (conjunctive) population. For a better visualization, the yellow region indicates all the values greater than 2. This region, at low values of σ and basically independently from the population size, is characterized by a better coverage of the pure population. Values greater than 1 are also typical of the low N region, due to the prefactor of the global error being lower in the pure case. As soon as N is sufficiently high and σ allows a good stimulus space coverage, the conjunctive case outperforms the pure case. This effect is stronger in the low σ region, due to the slow scaling of the global errors in the pure case. Gradually, increasing σ, the ratio saturate at the value given by the ratio of the local error only. (**d,e**) Optimal tuning width in function of population size and relative error, for pure (blue) and conjunctive (red) population in the first layer. Shaded region corresponds to 1 s.d. The global error decreases much slower in the pure population, as one can see from both the optimal width and the error being larger and with a smaller slope. At very low population sizes, it is possible to see the difference in the prefactor, that makes a pure code slightly better. At $N \sim 75$ the optimal width for the conjunctive case saturates, due to the loss of coverage. The relative error stops decreasing exponentially and starts decreasing only linearly, while the pure population does not suffer of this problem (it will do at lower widths). Ultimately, since the optimal width will continue to decrease in the pure population, the error will become lower than the conjunctive case.

We obtained the following scaling for global and local errors

$$\begin{aligned}
\langle \varepsilon^2_{l,p} \rangle_W &\approx 3 \langle \varepsilon^2_{l,c} \rangle_W \approx 9 \frac{2\sigma^2 \eta^2}{N} \\
\langle \varepsilon^2_{g,p} \rangle_W &\approx 3 \frac{\bar{\varepsilon}_g}{\sigma \sqrt{2\pi N}} \exp\Big(-\frac{N}{6} \log(\frac{1+2\eta^2}{2\eta^2}) \Big) \\
\langle \varepsilon^2_{g,c} \rangle_W &\approx \frac{\bar{\varepsilon}_g}{\sigma^3 \sqrt{2\pi N}} \exp\Big(-\frac{N}{2} \log(\frac{1+2\eta^2}{2\eta^2}) \Big).
\end{aligned} \tag{9}$$

In both cases the qualitative behavior is the same of the 1D case, with an optimal σ^* that allows the balance between local and global errors and decreases exponentially fast increasing the population size, Fig. 5a,b.

In order to appreciate the difference between the two cases, it is useful to look at the error ratio between them, Fig. 5c. One aspect that is not captured by the analytic computations is the one of loss of coverage. In the two cases the number of neurons is the same, but the pure population cover more efficiently the stimulus space (this difference will grow exponentially with the stimulus dimensionality, making unfeasible a fully conjunctive code for high dimensional stimuli). Under a critical σ, the stimulus is encoded only in the tails of the Gaussian in the conjunctive population, and some stimuli will produce basically no response in the second layer. This makes the error increase, independently from the number of neurons of the second layer; the pure population instead will suffer similar problems at much lower values of tuning width. Therefore, in the very low σ region a pure code is more advantageous, as one can also see by the optimal σ^* that saturates at a certain value in the conjunctive case, while still continues to decrease in the pure one, Fig. 5d. Pure cells are also advantageous in the low N regime, since the prefactor of the exponential scaling is lower in this case (even if this is not a very interesting regime, since the error is very large). As soon as N is sufficiently large, the advantages of the conjunctive code emerge. The local error of the conjunctive population is lower than the one of pure population, as already showed in [22], and this feature is preserved in the second layer. Moreover, the global error scales much more slowly in the pure case, since each dimension is encoded separately and the variance along each dimension is 1/3 of the total one. This will make a conjunctive population much more efficient in the low σ region, where the pure one is heavily affected by global errors. As a result, both the optimal width σ^* and the error $\varepsilon(\sigma^*)$, Fig. 5e, are lower in the conjunctive case. In the conjunctive case the optimal error will stop decreasing exponentially (it will still decrease linearly) after that the

lower bound for the optimal width is reached, and at larger population sizes and lower widths the pure case will ultimately be more advantageous.

2.5 ~~Complex tuning in~~ Compressed coding in the monkey motor cortex

A feedforward network, made up by a layer of conjunctive cells encoding a 3-D stimulus and random projecting onto a smaller layer, have been employed recently in [21] to explain the heterogeneity in the tuning curves of neurons in primary motor cortex. This allows us to apply our theory to a realistic example.

Neurons of primary motor cortex (M1) are tuned to movement parameters. In a static task, like keeping the hand fixed in a certain position, a population of neurons is required to encode the position of the hand in the three-dimensional space, and regulate accordingly muscles' activity. Historically [7], this tuning has been assumed to be a linear function of hand position

$$r_i(\mathbf{x}) = a_{0,i} + a_{1,i}x_1 + a_{2,i}x_2 + a_{3,i}x_3. \tag{10}$$

Changing coordinates, this corresponds to cosine tuning with respect to a "preferred position" (or Positional Gradient)

$$r_i(\mathbf{x}) = a_{0,i} + PP_i \cdot \mathbf{x}, \tag{11}$$

with $PP_i = (a_{1,i}, a_{2,i}, a_{3,i})$. More recently, the authors of [21] analysed a large population of M1 neurons during such a static "hold" task and showed that the linear model was not able to explain the large degree of irregularities in the tuning curves. Indeed, while some neurons showed a good agreement with the linear model, most of them had unpredictable deviations from the parametric shape, some time resulting in a very irregular tuning. To quantify these irregularities, the authors used two summary statistics of the neurons (see Methods). From their analysis, it emerged that the data are more structured than what expected from a completely random distribution of firing rates, but not fully explainable with a simple parametric shape. The authors managed to explain this heterogeneity with a feedforward neural network with random connectivity, very close to our model. The parameters of the models are fitted to match the summary statistics of the data.

Inspired by this analogy, we applied the previous theory to study the coding properties of such a population with irregular tuning curves and its relative advantage with respect to a population of linear neurons, Eq.(10). We generated an irregular population, representing motor neurons with a random network where a first layer of conjunctive cells encodes the target position in the 3D space with Gaussian receptive fields. This population is supposed to represent parietal reach area neurons or premotor neurons, that show similar properties of position encoding in visual space [23]. This information is transmitted to a smaller population, representing the M1 neurons, with random synapses, producing therefore the irregular tuning curves. With respect to our 3D model described in the previous section (see Methods), in the original paper the weights are uniformly distributed and the random sum is passed through a threshold non linearity to enforce positivity of the firing rates.

The linear population is obtained sampling the principal directions on the unit sphere and constraining the dynamical range of the neurons to be the same of the irregular ones. We considered the stimulus space to be the same of the experiment, a cube of side 40 cm. The collected data sample coarsely the tuning curves, since the responses are collected at 27 positions (3 by 3 by 3 grid). Since irregularities are supposed to improve the local accuracy, we generated and tested the neurons' responses at a much finer scale (21 by 21 by 21 grid). To measure the relative advantage of the irregular population we introduced the 'Mean Percent Improvement'

$$\Delta\varepsilon = \frac{\varepsilon_l - \varepsilon_{irr}}{\varepsilon_l}. \tag{12}$$

Firstly, we measured the relative improvement for different population sizes N and tuning width σ, keeping fixed the noise variance, Fig. 6a. When σ is too low, for a given N, the presence of global errors suppresses the local improvements given by the irregularities and a smoother code is more efficient ($\Delta\varepsilon < 0$). Increasing N and keeping fixed σ, the global errors are suppressed exponentially fast and the accuracy given by irregularities boosts the coding performances. This effect is attenuated when σ becomes larger, since the tuning curves approach the linear regime. The region of advantage of the linear case with respect to the irregular one grows with the noise variance (data not shown).

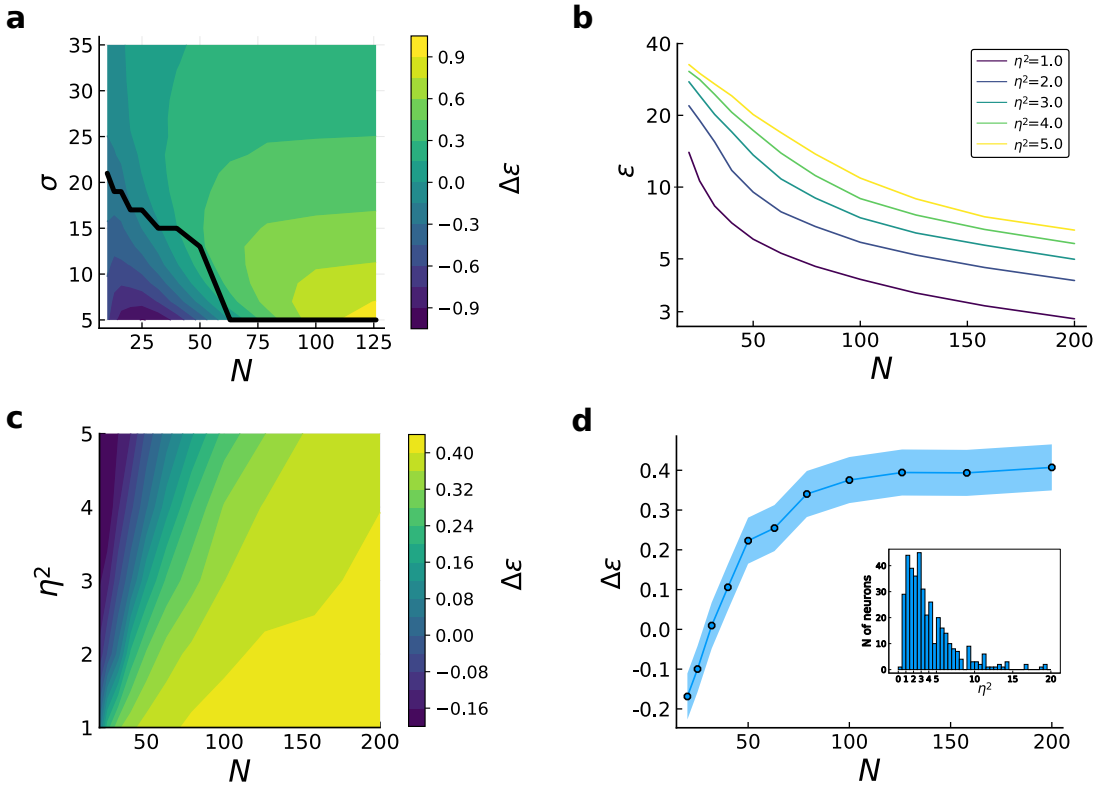

Figure 6: **Linear vs irregular tuning.** (**a**) Mean percent improvement of the irregular population (averaged over 4 different pools of a given size) with respect to the linear one, in function of population size and tuning width. The black line indicates the critical values of $N - \sigma$ after which the irregular population performs better than the linear one. In the region below (violet) the global errors of the irregular scheme make a smoother tuning curve more advantageous. Increasing N, we rapidly fall in the yellow region where the irregularities improve the local accuracy of the code. This advantage is stronger for lower σ, as one can see from the yellow becoming green at larger values of tuning width. (**b**) Error in function of the population size for the data-fitted model σ_f, for different levels of noise variance (averaged over 4 different pools). In this case the stimulus space is a 40x40x40 cube. It is possible to observe that the scaling is exponential for low values of N and then starts to be linear after a critical value, that increases with the noise variance. (**c**) Mean percent improvement of the irregular population, with data fitted parameter σ_f, with respect to the linear one, in function of the pair $N - \eta^2$. At small population sizes the irregular tuning produces global error and smoother tuning curves perform better (violet region, $\Delta\varepsilon < 0$). Increasing N, the global error are suppressed and irregularities improve the local accuracy. The critical N at which the transition happens increases with the noise variance. (**d**) Mean percent improvement of the same population, in function of population size, for a noise model extracted from the data. A noise variance is assigned to each neuron, obtaining a very heterogeneous distribution of noise in the population, as it is showed in the insect. We extracted 8 different pools of size N, each neuron with its noise variance (the same for the two populations), we computed $\Delta\varepsilon$ and we averaged the result to obtain the blue curve. Shaded region represents 1 s.d. For low levels of N, linear tuning produces better results. At $N \sim 40$, the higher local accuracy weights more than the global errors, and the irregular code starts to perform better. The improvement saturates at a finite value of ~ 0.4 when the number of neurons is $N \sim 150$, since global errors are fully suppressed.

Then, we used the real data to analyse the properties of the model in a biologically relevant regime. We compared the distribution of the summary statistics given by our model with the one of the data (preprocessing the data such that the tuning curves have 0 mean and unit variance, consistently with our model, see Methods); in particular, we had just one tunable parameter σ. Our aim was not to fit perfectly the data (see [24] for the problem of fitting implicit generative models), but to obtain a plausible level of irregularities, defined by σ. We obtained that, despite the use of a simpler model, at a non trivial σ_f we were able to capture the distribution of the complexity measure across the population at a level comparable to the original paper. We used this optimal σ_f to analyze the scaling of the error in function of the population size for different levels of noise, Fig. 6b. We chose high levels of noise variance because the data themselves show a great trial to trial variability in the firing rates. With respect to the linear population, two regions are distinguishable in the $N - \eta^2$ plane, Fig. 6c. For small populations and high levels of noise, a smoother code is preferable; after a critical population size, that increases with higher noise variance, the irregular model performs better. Finally, we did the same kind of analysis extracting a noise distribution from the data, Fig. 6d. The noise in the data is not compatible with any simple model (Poisson, standard deviation proportional to the mean, and so on). Therefore we decided to extract a noise variance from the data (Methods) for each neuron ($N \sim 400$) and pre-assign this value to each neuron in simulations. The resulting distribution of noise variance is very heterogeneous, with a mean noise variance of 4.3 and with neurons with very high values, Fig. 6f, insect. For each population size, we extracted $n = 8$ different pools of neurons of that size (neurons of irregular and linear populations having the same noise variance) and we computed $\Delta\varepsilon$ for each pool, averaging then the results. As expected, the relative advantage of the irregularities grows with population size also in this case. It cross 0 when the advantage given by the local accuracy overcomes the effect of global errors. In this regime global errors contribute still substantially to the total error. $\Delta\varepsilon$ continues to grow until when global errors are suppressed, then it saturates since in both populations the error is now only local. Interestingly, it saturates at a population size $N \sim 150$; this is the same order of magnitude of the number of neurons controlling one individual muscle in this specific task, as showed in [21] by using a subset of the neural responses to reproduce EMG signals of individual muscles.

2.6 Compressed coding in presence of input noise

Until now, we considered noise on second layer neurons only; within this setting, the error can be made arbitrarily small increasing N. The number of neurons of the first layer, L, does not affect neither the local nor the global error, as long as it is sufficiently large to allow an overlap of the tuning curves. In other words, without input noise, L only sets a scaling for the optimal σ allowed such that all stimuli are encoded in the bulk of at least one tuning curve (what is called 'tiling property' in [10, 25], and it have been showed to be, numerically, $\sigma \sim \frac{1}{L}$ in [13]).

In a more realistic framework, the input neurons are affected by noise too; this ultimately fix a limit on the information available in the second layer, since shared connections add a non-diagonal part to the noise covariance matrix which result in the so called 'information-limiting' correlations [?]. In [26] the effect of this type of correlations in a two layer feedforward network with layers of equal size is explored. Our main result is that in the regime of compression, $N << L$, this kind of correlations have a small effect. If the first layer neurons are affected by isotropic gaussian noise of variance ξ^2, the responses of the second layer are described by a multivariate gaussian distribution with covariance matrix $\Sigma = \eta^2\mathbf{I} + \xi^2\mathbf{W}\mathbf{W}^\mathbf{T}$. In order to see the effects of correlations on the MSE, we did numerical simulations to compare the case of correlated noise due to input noise to the case where we set to 0 the off-diagonal terms of Σ, $\Sigma_{ind} = (\eta^2 + \xi^2)\mathbf{I}$, which is equivalen to the case of i.i.d. output noise of *effective variance* $\tilde{\eta}^2 = \eta^2 + \xi^2$. We started to analyze the case where $\xi^2 << \eta^2$, and we observed that the effect of correlations start to be slightly relevant only at large values of N, Fig. 7a. We can give an intuitive explanation about the effects of correlations on the two kinds of errors. Correlations shrink the volume of the cloud of possible responses, and this should reduce the probability of global errors. Nevertheless, given the random magnitude of this kind of errors and the fact that they are present only at low values of N, this effect is very difficult to see systematically in numerical simulations. On the other hand, the covariance matrix is related to the coding manifold through the synaptic matrix, and this introduce the so-called 'information limiting correlations' on the local error. Analytically, we performed a perturbative expansion of the inverse of the covariance matrix in order to obtain the linear FI in the correlated case, whose

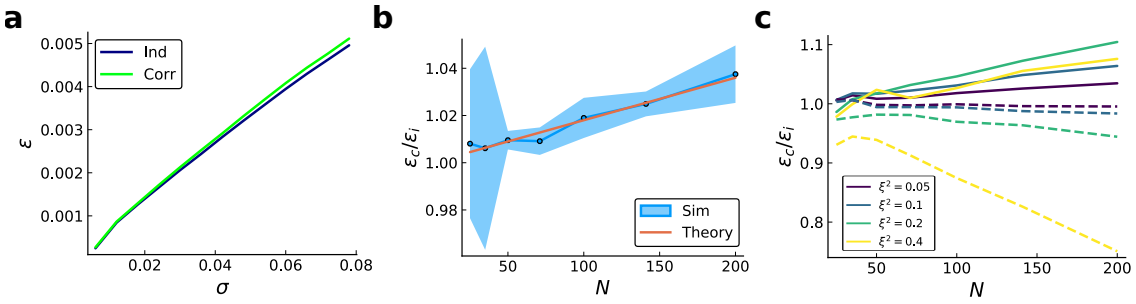

Figure 7: **Effects of correlated noise on compressed coding.** (**a**) RMSE in function of σ in case of correlated noise due to shared connectivity and diagonal noise with effective variance $\tilde{\eta}^2$, averaged over 8 realizations of the synaptic matrix. $N = 70$, $\tilde{\eta}^2 = 0.5$ and the contribution of input noise is small, $\xi^2 = 0.05$. (**b**) Error ratio (MSE) between correlated noise due to input noise and diagonal noise with effective noise variance, in function of N, and theoretical prediction, Eq.(13). The noise variances are the same of the previous figure, $\sigma = 0.045$, average over 8 realizations of synaptic matrix, shaded region indicating 1.s.d. The goodness of the prediction increases with higher values of N, since in this regime the local errors are dominant. (**c**) Error ratio between correlated noise due to input noise and diagonal noise (filled lines), and error ratio between correlated noise with random covariance matrix and diagonal noise (dashed lines). Different colors denote different contributions coming from the off-diagonal terms ξ^2, increasing from violet to yellow, when the effective noise variance is kept fixed $\tilde{\eta}^2 = 0.5$. When correlations come from shared connections, the ratio is positive since we have information-limiting correlations. Their effect are a non-linear function of $\frac{\xi^2}{\tilde{\eta}^2}$, due to the competition between the first order (positive) and second order (negative) corrections. With a random covariance matrix, correlations decrease the error.

inverse is a lower bound to the MSE. We obtained that the local error in case of input noise is given by

$$\varepsilon_l^2 \approx \varepsilon_{l,i}^2 (1 + \frac{N}{L}\frac{\xi^2}{\tilde{\sigma}_\eta^2} - \frac{\xi^4}{\tilde{\eta}^4}\frac{N}{L}), \tag{13}$$

where $\varepsilon_{l,i}^2$ is the error we would obtain considering i.i.d. output noise of effective variance $\tilde{\eta}^2$. We obtained therefore that the first order correction, which are deleterious for the coding properties, are of order $\frac{N}{L}$. We checked this prediction computing the error ratio $\frac{\varepsilon^2}{\varepsilon_i^2}$ for different values of N, averaging over different realizations of the synaptic matrix and tuning width (since, at least for the local error, it should not matter), Fig. 7b, filled lines. We also compared different values of ξ^2, keeping fixed the effective noise variance $\tilde{\eta}^2$ (therefore varying the relative contribution of the input and output noise). We see that increasing ξ^2, the second order term mitigates the negative effect of the first order one.

Finally, it is instructive to see that the information limiting correlations arise precisely because of the structure of the covariance matrix given by shared connections. In order to see it, we examined the case where the covariance matrix has the same statistic of Σ, but is not related to $\mathbf{W}$, for example using a random covariance matrix $\Sigma_{rand} = \tilde{\eta}^2\mathbf{I} + \xi^2\mathbf{X}\mathbf{X}^\mathbf{T}$, where $X_{ij} \sim \mathcal{N}(0, \frac{1}{L})$. In this case, correlations help to decrease the local error with respect to independent case, because the cloud of possible responses becomes randomly oriented with respect to the coding manifold. Analytically, this can be seen using a similar perturbative expansion and observing that the first order correction is missing,

$$\varepsilon_{l,rand}^2 \approx \varepsilon_{l,i}^2 (1 - \frac{\xi^4}{\tilde{\eta}^4}\frac{N}{L}). \tag{14}$$

We performed similar numerical simulations and we showed that the improvement indeed exists, and increase with the number of neurons N and the weight of the off-diagonal terms ξ^2, Fig. 7c, dashed lines.

3 Discussion

We analyzed the coding properties of a population of neurons whose tuning curves are described by a generative model with heterogeneous (random) connectivity. This coding scheme, in the optimal configuration, allows an error scaling that is exponential in the population size, similarly to what found in grid cells [16]; using this analogy, a random uncorrelated code is an Exponentially Strong Population Code (EPC). This is due to the uncorrelated activity at nearby stimuli, which allows the neurons to fill the activity space efficiently. Like in grid cells, this comes at the price of high ambiguity at single cell level, that can lead to large scale errors. This kind of errors are reduced increasing the smoothness of tuning curves. Tuning the width σ of first layer neurons, we interpolate between the high local accuracy of an uncorrelated code (EPC) and the noise robustness of a code with monomodal tuning curves (called Classical Population Code in [16]). The existence in the brain of distributed codes with high (exponential) capacity has been already showed in the context of discrete stimuli [27]. Population of neurons in cortex show a great diversity in the response to face stimuli, and this allows a population of N neurons to encode exponentially many faces. We extended this concept to continuous stimuli. This introduces a notion of magnitude of errors (not present in the context of discrete stimuli, where the task is simply to discriminate between two different stimuli) and we showed how it is possible to balance, in function of population size and noise magnitude, the trade-off between accuracy and robustness requirements.

We considered also the case where the stimulus is multidimensional and we showed how the characteristics of the first layer of neurons affect differently the balance between local and global errors. This adds new possible criteria about the optimal ways of encoding multidimensional variables, well explored in [22]. Interestingly, both models have been already employed in the literature. For the pure case, a model with a set of sensory layers, encoding separately different features of the stimulus, connected randomly to a single population of neurons, have been employed in [28] to explain the flexibility of the working memory. For the conjunctive case, a model with random projections from parietal neurons to motor neurons have been employed to explain the heterogeneity in the tuning curves of M1 neurons. Using the dataset of this experiment, we analyzed the model performance in a biologically relevant region of the parameters space and we showed how such irregularities boost the coding performances with respect to a population of neurons with more classical and regular tuning properties.

Population coding. Overall, we extended the previous work about optimal encoding of low dimensional variables to a population described with a generative model, rather than describing all neurons as an homogeneous population with fixed single cell parameters and common parametric tuning curves (or small deviations in the value of the parameters [29]). Models of this kind gained a lot of attention in the recent years, for their ability to describe the biological richness and diversity present in real data.

With the growth of experimental techniques for large-scale data recordings and the statistical techniques to analyze them, the focus of study shifted from single neurons to populations of neurons [30]. The encoding of external stimuli in neural populations is probably much more than the simple sum of single neurons response properties. One first example is given by neurons with mixed selectivity [19, 20], that is neurons combining non linearly the response to two (or more) features. Again, this kind of responses are highly ambiguous at single cell levels, but they allow to increase the dimensionality of neural representations and, consequently, the computational power of neural circuits. In the opposite direction, some neural populations activities are supposed to lie in a low dimensional manifold to increase the robustness of the neural code. In this case, the joint activity of neural populations is coupled by few 'neural modes' [31, 32]. Recent experimental studies analyzed the dimensionality of population responses in real neural circuits [33, 34] and tried to understand the implications for neural coding. In [34] it is suggested that the neural activity lies in a high dimensional manifold with a power law scaling of the variance explained in each dimension. This suggested a fractal-like structure of the underlying manifold, supposed to be optimal to balance efficiency and robustness.

In this work we analyzed the coding properties of a specific type of neural manifold: a random Gaussian manifold. Varying σ can be thought as exploring different 'intrisic dimensionality' of this manifold. We notice that this type of manifold was already introduced in [35, ?]. They play an important role as a null model to quantify the dimensionality of real neural trajectories in experimental data. Moreover, being randomly oriented with respect to neural axis, they allow downstream circuits to recover their structure even using random subsets of neurons. In this

direction, an important research line is to analyze the coding properties of manifolds extracted from real data and under more realistic noise models, and give an explanation of their characteristics under the lens of efficient coding hypothesis.

Decoder and criteria for optimal encoding. In order to compute the optimal coding properties of the network, we used the error in the stimulus estimate from an ideal decoder. This is a common approach used in the literature, due to the difficulty in treating analytically the MSE, several studies focused on minimizing information theory related quantities, like FI or Mutual Information. As already said, this is a typical case where this kind of approach may give misleading results. Remarkably, other criteria of optimality have been suggested to explain the tuning properties of neurons, like the capacity of downstream circuits to construct a given motor response [36]. In the case of study proposed here, a big assumptions is made regarding the decoder. In particular, the decoder is supposed to know the true responses and the noise variance. The computations of the decoder can be implemented by a biologically plausible neural network, since all the operations have been considered in the literature as canonical computations executed by neural circuits in sensory processing [3, ?, ?]. The structure of the decoder is very similar to the Bayesian Population Vector of [10]. Such a decoder can be thought to be learned through an Hebbian mechanism, for example. Nevertheless, if the manifold is very complicate (i.e. contains a lot of high frequency terms and is very twisted) and we can only have access to the noisy responses, it is not clear how this would impact the learnability of the underlying map. Presumably, limitations on the decoding resources affect criteria for optimal encoding, but until now few studies analyzed this problem and the hypothesis of ideal decoder is ubiquitous.

Compressed Sensing. We considered the problem of transmission of a low dimensional variable x encoded into an high dimensional representation (L neurons), using as less resources as possible (N). This is very similar to the framework of compressed sensing (CS) [?, ?]. A fundamental result in this field is that it is possible compress a high dimensional signal which has a low dimensional structure (i.e. which is sparse in some basis) using a number of random measures (random projecting onto a smaller space) which scales only logarithmically with the dimension of the signal. This corresponds to our result about the exponential scaling of the error with the population size: Eq.(35) can be inverted to compute the number of random projections N such that it is possible to recover a 1D signal embedded in an L-dimensional space with a given error probability. Nevertheless, there are some relevant differences. The typical task of compressed sensing is to reconstruct the high dimensional vector from the random measurements, and the goodness of reconstruction function is measured by the difference between the estimate and the original vector. In our framework, we don't want to reconstruct pattern of activity in the first layer, but we want to obtain a faithful estimate about the low dimensional variable that evoked it. Moreover, we study how the way this low dimensional variable is encoded (what in CS is the sparse basis) affects the minimal number of (noisy) neurons necessary to recover it.

CS gained a lot of attention in the neuroscience field [?]. The brain has to solve the problem of transmission of information across different areas, and often this is complicated by convergent pathways or bottlenecks, and CS offers an attractive idea to study how this compression is possible. This hypothesis has been used to explain the properties of olfactory systems, where a large number of odorant molecules is encoded in the activity of much fewer olfactory neurons [?, 37]. More generally, random connectivity has been proposed as a possible way neural circuits solve the problems of compression and expansion [38, 18] and a possible reason underlying the heterogeneity and diversity of tuning curves within the same population. It is very unlikely that all synapses in the brain are randomly distributed, as experimental [] and theoretical [39] results suggest that the connectivity data reveal a lot about the functions of neural circuits. At the same time, the heterogeneity and irregularities inside the same population of neurons, the high flexibility of models with random connectivity [40], and the difficulty in tuning every single synapse, make plausible the presence of some circuit elements which are random. Studying the coding properties of tuning curves arising from random projecting the activity of more structured sensory layers is therefore a good way to understand which optimization principles may guide the information routing in the brain, which problems neural circuits have to face, and guide the search for evidence of them in real data.

Criteria for faithful coding &
decoder
Rework:
(1) Dividere in 2 part.
(2) Expand part 1.
(3) We should discuss how to present decoder.
activity of
Expand
not clear
Said too naively (cell types, learning)
→ also make less naive.
Excellent discussion of CS in our context!
Overall, I really like the Discussion.

4 Methods

Throughout the discussion, bold letters denote vectors $\mathbf{r} = \{r_1, r_2, ..., r_N\}$, $\|\mathbf{r}\|_2^2 = \sum_i r_i^2$ represents the $L2$ norm, capital bold letters $\mathbf{W}$ represent matrices. Simulations were done using Julia language, [**?**].

4.1 Model description: 1-D stimulus

Random Feedforward Network. We considered a two layer architecture. A 1-D stimulus x is encoded by a sensory layer of L neurons, indexed by j, with Gaussian tuning curves centered on a preferred stimulus c_j. This layer projects onto a layer of N neurons ($N < L$) with normally distributed random weights: [2]

$$\begin{aligned} u_j(x) &= \frac{1}{Z}\exp\Big(-\frac{(x-c_j)^2}{2\sigma^2}\Big) \quad j = 1, ..., L, \\ v_i(x) &= \sum_{j=1}^{L} W_{ij} u_j(x) \qquad i = 1...N, \\ W_{ij} &\sim \mathcal{N}(0, \frac{1}{L}). \end{aligned} \tag{15}$$

Without loss of generality, we can restrict the stimulus space to be $x \in [0, 1]$. We considered an uniform prior on the stimulus, and, consequently, an uniform arrangement of neurons' preferred positions $c_j = \frac{j}{L}$ in the stimulus space.

Constraints. Z is a normalization constant; for different widths, we chose to constrain the variance of responses of second layer neurons across all stimuli, R, also called dynamic range. For each neuron, this quantity depends from the specific realization of the synaptic weights, therefore we imposed the constraint on average. Namely:

$$\begin{aligned} R &= \langle \int_0^1 dx \Big(v_i(x) - \int_0^1 dx v_i(x) \Big)^2 \rangle_W \\ &= \langle \int_0^1 dx \Big(\sum_j W_{ij} u_j(x) - \big(\int_0^1 dx \sum_j W_{ij} u_j(x) \big)^2 \Big)^2 \rangle_W \\ &= \langle \sum_{jj'} W_{ij} W_{ij'} \int_0^1 dx u_j(x) u_{j'}(x) \rangle_i + \langle \sum_{jj'} W_{ij} W_{ij'} \int_0^1 dx u_j(x) \int_0^1 dx u_{j'}(x) \rangle_W . \end{aligned} \tag{16}$$

where $\langle \dots \rangle_W$ denotes the mean over the synaptic weights. We used the approximation $\int_0^1 dx \exp\Big(-\frac{(x-c_j)^2}{2\sigma^2}\Big) \approx \sqrt{2\pi\sigma^2}$, which is valid if c_j is sufficiently far from the borders and σ is small (this will introduce some edge effects, negligible in the regime of σ and L we considered). Using also the fact that weights are statistically independents and have zero mean, $\langle W_{ij} W_{ij'} \rangle = \frac{1}{L}\delta_{jj'}$, we obtain the condition on Z :

$$Z^2 = \frac{\sqrt{\pi\sigma^2} - 2\pi\sigma^2}{R}. \tag{17}$$

In the following, we will often use an approximation for small widths: $Z^2 \approx \frac{\sqrt{\pi\sigma^2}}{R}$.

Gaussian Processes analogy. If we assume that the spacing between preferred positions is

[2] In the following, all the computations are done for a 1D linear stimulus encoded by Gaussian tuning curves. At the same time, we will often assume translational invariance; this will unavoidably introduce edge effects. A more rigorous way would be to consider a circular stimulus and von Mises tuning curves in the first layer:

$$u_j(x) = \frac{1}{Z} exp(\mathcal{K} cos(2\pi(x - c_j))).$$

This complicates the form of the correlation function and it would require a modification of the error function. Anyway, we considered regimes where $\mathcal{K}$ is large and the von Mises function can be approximated locally as a Gaussian with width $\sigma^2 = \frac{1}{\mathcal{K}}$. In the regimes of σ considered in the simulations, the edge effects are small and simulations with circular stimulus did not change qualitatively the results.

small, we can approximate the sum in Eq.(15) with a convolution integral of a random noise process (the synaptic weights) with a smoothing kernel (the tuning curves of first layer neurons)[3]:

$$v_i(x) = \sum_{j=1}^{L} W_{ij} u_j(x) = L \int_0^1 dc_j u(c_j - x) W_i(c_j).$$

This gives rise to a Gaussian process, as described in [?, ?]. Computing the covariance function is straightforward:

$$\begin{aligned}\langle v_i(x) v_i(x') \rangle_i &= \langle \sum_j W_{ij} W_{ij'} u_j(x) u_{j'}(x') \rangle_i = \frac{1}{Z^2} \sum_j \frac{1}{L} \delta_{jj'} \exp\Big(- \frac{(x - c_j)^2}{2\sigma^2} \Big) \exp\Big(- \frac{(x' - c_{j'})^2}{2\sigma^2} \Big) \\ &\approx \frac{1}{Z^2} \int dc_j \exp\Big(- \frac{(x - c_j)^2 + (x' - c_j)^2}{2\sigma^2} \Big).\end{aligned} \tag{18}$$

Assuming translational invariance across all stimulus space, we obtain that the tuning curves are described by a 1D Gaussian process with 0 mean and Gaussian kernel with correlation length $\sqrt{2}\sigma$:

$$\begin{aligned} < v_i(x) > &= 0 \\ K(x, x') = < v_i(x) v_i(x + \Delta x) > &= \frac{\sqrt{\pi \sigma^2}}{Z^2} e^{-\frac{\Delta x^2}{4\sigma^2}}. \end{aligned} \tag{19}$$

This network maps the 1D stimulus space $[0, 1]$ onto a 1D manifold embedded in the N dimensional space of neurons' activity. The coordinates of this manifold are described by N independents samples of the aforementioned Gaussian process. Interestingly, this corresponds to the definition of "Random Gaussian Manifold" proposed in [?, 35].

4.2 Coding - decoding process

Noise Model. We considered an isotropic Gaussian model for the noise on the second layer neurons. At each trial, the vector of responses to a given stimulus x is

$$\mathbf{r} = \mathbf{v}(x) + \mathbf{z}, \tag{20}$$

where $\mathbf{z}$ is a noise vector of independent Gaussian entries with a fixed variance, $\mathbf{z} \sim \mathcal{N}(0, \eta^2 \mathbf{I})$. The likelihood of a response vector given a stimulus (for a fixed realization of the synaptic weights) is given by

$$p(\mathbf{r}|x) = \frac{1}{(2\pi\eta^2)^{N/2}} \exp\Big(- \frac{\|\mathbf{r} - \mathbf{v}(x)\|_2^2}{2\eta^2} \Big). \tag{21}$$

In this case, the variance of the noise does not depend from the response (like in Poisson neurons, for example). What governs the error is the Signal to Noise Ratio ($SNR = \frac{R}{\eta^2}$); we set the variance of the responses $R = 1$ and we varied η^2 to explore different noise regimes.

The noise model can be extended to include correlations in the noise affecting different neurons. Denoting with Σ the noise covariance matrix, the likelihood of the neural response in second layer neurons can be written as a multivariate gaussian distribution,

$$p(\mathbf{r}|x) = \frac{1}{(2\pi)^{N/2} (\det(\Sigma)^{1/2}} \exp\Big(- (\mathbf{r} - \mathbf{v}(x))^T \Sigma^{-1} (\mathbf{r} - \mathbf{v}(x)) \Big). \tag{22}$$

Loss function and decoder. We used the Mean Square Error (MSE) in stimulus estimate as loss function to measure the coding properties of the neural population. An estimator, or decoder $\hat{x}(\mathbf{r})$, is a function that takes in input a noisy response and output an estimate of the stimulus that evoked it. The MSE, for a given network realization, is defined as

$$\varepsilon^2 = \int dx \int d\mathbf{r} p(\mathbf{r}|x) (\hat{x}(\mathbf{r}) - x)^2. \tag{23}$$

[3] This is a delicate integral to treat, as it is not properly defined. One should pass through Ito integration and ito isometry properties to define this object rigourously.

We will often consider the average MSE $\langle \varepsilon^2 \rangle_W$. To give a better idea of the error magnitude, often we showed the Root MSE $\langle \varepsilon \rangle_W = \sqrt{\langle \varepsilon^2 \rangle_W}$. This quantity is generally hard to compute, even knowing a closed form for the estimator. In numerical simulations we computed this integral with standard Monte Carlo method. At each step we extracted a set of L stimuli (one for each preferred position of the first layer neurons) from the uniform distribution and we generated the noisy responses. We then passed the noisy responses through an ideal decoder (see below) and we updated the error estimate. We iterated this process until when the MSE estimate was within a tolerance of 10^{-7} in the last 50 steps.

The estimator which minimizes the MSE is called Minimal Mean Square Error estimator (MMSE), and is given by the average of the posterior distribution. Using Bayes theorem, we obtain that the posterior is simply proportional to the likelihood due to the choice of uniform prior, and the MMSE estimator can be written as

$$\hat{x}_{MMSE} = \int_0^1 dx p(x|\mathbf{r})x = \frac{\int_0^1 dx x p(\mathbf{r}|x)}{\int_0^1 dx p(\mathbf{r}|x)}. \tag{24}$$

This function can be approximated by a simple neural network. Discretizing the stimulus space in M values, $x_m = \frac{m}{M}$, and substituting the expression for the likelihood Eq.(21), we can approximate the integrals as discrete sums

$$\begin{aligned} \hat{x} &= \frac{\sum_m x_m p(\mathbf{r}|x_m)}{\sum_m p(\mathbf{r}|x_m)} = \frac{\sum_m x_m \exp\left(- \frac{\sum_i r_i^2 + v_i^2(x_m) - 2v_i(x_m)r_i}{2\eta^2} \right)}{\sum_m \exp\left(- \frac{\sum_i r_i^2 + v_i^2(x_m) - 2v_i(x_m)r_i}{2\eta^2} \right)} \\ &= \frac{\sum_m x_m \exp\left(\frac{\sum_i 2v_i(x_m)r_i - v_i^2(x_m)}{2\eta^2} \right)}{\sum_m \exp\left(\frac{\sum_i 2v_i(x_m)r_i - v_i^2(x_m)}{2\eta^2} \right)}, \end{aligned} \tag{25}$$

where we removed $\sum_i r_i^2$, common to both numerator and denominator. A layer of M neurons can compute the likelihood function for different stimuli x_m. Calling λ the connectivity matrix between the N neurons of the output layer and the M neurons of the decoder, its entries are given by the true responses to a given stimulus: $\lambda_{mi} = \frac{v_i(x_m)}{\eta^2}$. Then, the sum is passed through an exponential non linearity with the addition of a bias term $b_m = \sum_i \frac{v_i(x_m)^2}{2\eta^2}$ to obtain the output of a single neuron $h_m = \exp\left(\sum_i \lambda_{mi} r_i - b_m \right)$. This layer could implement a winner-take-all dynamics to output the maximum likelihood (ML) estimator

$$\hat{x} = \arg\min_x \|\mathbf{r} - \mathbf{v}(x)\|_2^2 = \arg\max_{x_m} h_m. \tag{26}$$

Alternatively, the output of each neuron can be weighted according to its preferred stimulus (with the addition of a divisive normalization) to obtain the MMSE estimator

$$\hat{x} = \frac{\sum_m x_m h_m}{\sum_m h_m}. \tag{27}$$

In numerical simulations, we adopted for the decoder the same discretization of the stimulus of the first layer, using $M = L$ and spacing uniformly the preferred stimuli x_m. Note that the decoder is ideal, since it is assumed to know the true responses and the variance of the noise. The same decoder can be extended to decode responses of neurons with different noise variance (Fig. 6), with the following modifications $\lambda_{mi} = v_i(x_m)/(2\sigma_{\eta_i}^2)$, $b_m = \sum_i v_i(x_m)^2/(2\sigma_{\eta_i}^2)$.

Similarly, also a non-diagonal noise covariance matrix Σ can be treated, with the difference that the decoding weights and biases are now correlated: $\lambda_m = \mathbf{v}^T(x_m)\Sigma^{-1}$, $b_m = \mathbf{v}^T(x_m)\Sigma^{-1}\mathbf{v}(x_m)$. To estimate the scaling of the error, in the following sections we will often use the ML estimator, since it has an easier geometrical interpretation (minimal distance). In the main text we showed results for the optimal decoder (MMSE), but the performances for the two are very similar.

4.3 Errors' computation

Narrow tuning curves. If $\sigma \to 0$, the first layer neurons respond only to their preferred stimulus. For this extreme case, we suppose that the stimulus can assume only L discrete values $x_j = \frac{j}{L}$. The responses of the second layer neurons are given by $v_i(x_j) = W_{ij}$, with $W_{ij} \sim \mathcal{N}(0,1)$, and are uncorrelated for different stimuli. Let's denote with $p_e(\mathbf{r}|x) = p(\mathbf{r}|x)\Theta(|\hat{x} - x|)$ the (conditioned) probability density function that the noise will produce an error, where we introduced the Heaviside function $\Theta(x) = 1$ only if $x > 0$ (and 0 otherwise). We notice that, taking the average over the synaptic weights, the magnitude of the error is uncorrelated with its probability and no more depends on the specific realization of the noise $\mathbf{r}$. The average MSE can be rewritten as

$$\begin{aligned} \langle \varepsilon^2 \rangle_W &\approx \frac{1}{L} \sum_x \int d\mathbf{r} \langle p_e(\mathbf{r}|x) \rangle_W \langle (\hat{x}(\mathbf{r}) - x)^2 \rangle_W \\ &= \langle P(\varepsilon) \rangle_W \langle \frac{1}{L} \sum_x (\hat{x} - x)^2 \rangle_W, \end{aligned} \tag{28}$$

where $\langle P(\varepsilon) \rangle_W = \langle \int d\mathbf{r} p_e(\mathbf{r}|x) \rangle_W$ is the average probability that, given a stimulus, the noise will cause an error in its estimate; despite the notation, it does not depend from the specific value of x. This formula has an intuitive interpretation: the average MSE is the mean probability of having an error on a stimulus time the mean error magnitude. Let's suppose now to estimate the stimulus through a ML decoder, Eq.(26): we will obtain an error if there exists at least one x' such that $\|\mathbf{r} - \mathbf{v}(x')\|_2^2 < \|\mathbf{r} - \mathbf{v}(x)\|_2^2$. Since, averaging over the synaptic weights, all x' have the same probability to cause such an error, the average size of the squared error will be

$$\left\langle \frac{1}{L} \sum_x (\hat{x} - x)^2 \right\rangle_W = \frac{1}{L^2} \sum_{j=1}^{L} \sum_{j'=1}^{L} \left(\frac{j'}{L} - \frac{j}{L} \right)^2 \approx \frac{1}{6}, \tag{29}$$

where the last approximation holds for large L. The average probability of error can be expressed in terms of the probability of the complementary event

$$\langle P(\varepsilon) \rangle_W = 1 - \left\langle P\Big(\|\mathbf{r} - \mathbf{v}(x')\|_2^2 > \|\mathbf{r} - \mathbf{v}(x)\|_2^2 \Big) \quad \forall x' \neq x \right\rangle_W. \tag{30}$$

Averaging over different realizations of the synaptic matrix, the probability of not having an error on $x'$are i.i.d for different x', and we can write

$$\begin{aligned} \langle P(\varepsilon) \rangle_W &= 1 - \left(1 - \left\langle P\left(\|\mathbf{r} - \mathbf{v}(x')\|_2^2 < \|\mathbf{r} - \mathbf{v}(x)\|_2^2 \right) \right\rangle_W \right)^{L-1} \\ &\approx L \left\langle P\Big(\sum_i \big(v_i(x') - v_i(x) \big)^2 + z_i^2 - 2\big(v_i(x) - v_i(x') \big) z_i < \sum_i z_i^2 \Big) \right\rangle_W, \end{aligned} \tag{31}$$

where we explicitly substituted Eq.(20), we supposed that the average probability of having an error is small (much smaller than $\frac{1}{L}$), and we considered $L - 1 \approx L$. With the specified distribution of synaptic weights, the average difference between the response of the same neuron to two different stimuli $\tilde{v}_i = v_i(x) - v_i(x') = W_{ij} - W_{ij'}$ is normally distributed with variance equal to 2. Finally, averaging also over the noise distribution, we obtain

$$\langle P(\varepsilon) \rangle_W \approx L \int \prod_i d\tilde{v}_i \prod_i dz_i p(\tilde{v}_i) p(z_i) \Theta\left(-\sum_i \tilde{v}_i^2 + 2 \sum_i \tilde{v}_i z_i \right). \tag{32}$$

We have to compute the probability that the quantity $\tilde{d} = \sum_i \tilde{v}_i^2 - 2\tilde{v}_i z_i$ is less than 0, where $\tilde{v}_i \sim \mathcal{N}(0,2)$ and $z_i \sim \mathcal{N}(0, \eta^2)$. We can notice that, fixing $\lambda = \sum_i \tilde{v}_i^2$, the conditioned quantity $\tilde{d}|\{\tilde{v}_i^2\} \sim \mathcal{N}(\lambda, 4\lambda\eta^2)$ is normally distributed. Therefore, using the definition of error function, we can rewrite the error probability as

$$\begin{aligned} \langle P(\varepsilon) \rangle_W &\approx L \int_0^\infty d\lambda p(\lambda) \int_{-\infty}^0 d\tilde{d} p(\tilde{d}|\lambda) \\ &= \frac{L}{2} \int_0^\infty d\lambda p(\lambda) \operatorname{erfc}\left(\sqrt{\frac{\lambda}{8\eta^2}} \right), \end{aligned} \tag{33}$$

where $p(\lambda) = \frac{(\frac{\lambda}{2})^{\frac{N}{2}-1} exp(-\frac{\lambda}{4})}{2^{\frac{N}{2}+1}\Gamma(N/2)}$ is the probability density function of a Chi-squared distribution.

Computing this integral, we obtain

$$\begin{aligned}\langle P(\varepsilon)\rangle_W &\approx L\frac{(\frac{\eta^2}{2})^{\frac{N}{2}}\Gamma(N)}{\Gamma(\frac{N}{2})}{}_2\tilde{F}_1(\frac{N}{2},\frac{1+N}{2},\frac{2+N}{2},-2\eta^2)\\ &= L\frac{(\frac{\eta^2}{2})^{\frac{N}{2}}\Gamma(N)}{\Gamma(\frac{N}{2})\Gamma(\frac{2+N}{2})}\sum_{n=0}^{\infty}\frac{(\frac{N}{2})_n(\frac{N+1}{2})_n}{(\frac{N+2}{2})_n n!}(-2\eta^2)^n,\end{aligned} \tag{34}$$

where ${}_2\tilde{F}_1(a,b,c,x)$ is the regularized 2F1 Hypergeometric function and we substituted its definition. The Pochammer symbol is also defined through Gamma functions $(x)_n = \frac{\Gamma(x+n)}{\Gamma(x)}$. Simplifying and using the identity $\sum_{n=0}^{\infty}\frac{(x)_n}{n!}a^n = (1-a)^{-x}$, we obtain the final expression for the error probability

$$\begin{aligned}\langle P(\varepsilon)\rangle_W &\approx L(\frac{\eta^2}{2})^{\frac{N}{2}}\frac{\Gamma(N)}{\Gamma^2(\frac{N}{2})\frac{N}{2}(1-2\eta^2)^{\frac{N+1}{2}}}\\ &\approx L\frac{1}{\sqrt{2\pi N}}\exp\Big(-\frac{N}{2}\log(\frac{1+2\eta^2}{2\eta^2})\Big),\end{aligned} \tag{35}$$

where in the last step we used the Stirling approximation for the Gamma function.

Broad tuning curves. As soon as $\sigma > 0$, we allow for continuous stimuli and the resulting manifold in the activity space is smooth. In this case, the noise can also produce small scale local errors: we therefore split the error in two contributions, local and global. Since our system has a natural correlation length, we defined as global an error when the difference between the stimulus and its estimate is greater than σ : $|\hat{x}(\mathbf{r}) - x| > \sigma$. This definition is a bit tricky, since for very large σ all the errors will be local. Anyway, we are interested in the case where σ is relatively small, and what matters is that global errors are of the order of the size of the stimulus space. We rewrite the average error as

$$\langle\varepsilon^2\rangle_W = \langle\varepsilon_l^2+\varepsilon_g^2\rangle_W = \langle\int dx d\mathbf{r} p_l(\mathbf{r}|x)\big(\hat{x}(\mathbf{r})-x\big)^2\rangle_W + \langle\int dx d\mathbf{r} p_g(\mathbf{r}|x)\big(\hat{x}(\mathbf{r})-x\big)^2\rangle_W, \tag{36}$$

where with $p_{l/g}(\mathbf{r}|x) = p(\mathbf{r}|x)\Theta\big(\pm(\sigma - |\hat{x}(\mathbf{r})-x|)\big)$ we denoted the probability density function that, given x, the noise will cause a local/global error. It holds the following normalization $\int d\mathbf{r} p_l(\mathbf{r}|x) + p_g(\mathbf{r}|x) = 1$.

Local error. A ML decoder will output the stimulus corresponding to the closest point of the manifold, which in case of local error will correspond to the projection of the noise vector onto the manifold. Expanding linearly the response around x, we obtain

$$\Big\|\mathbf{r}\cdot\hat{\mathbf{v}}'(x)\Big\|_2^2 = \|\mathbf{v}(x+\Delta x)-\mathbf{v}(x)\|_2^2 \approx \|\mathbf{v}'(x)\|_2^2\,\Delta x^2, \tag{37}$$

where $\hat{\mathbf{v}}'(x)$ is the normalized vector in the direction of the derivative of the tuning curves. Clearly, $\Delta x^2 = \big(\hat{x}(\mathbf{r})-x\big)^2 = \frac{\|\mathbf{r}\cdot\hat{\mathbf{v}}'(x)\|_2^2}{\|\mathbf{v}'(x)\|_2^2}$ will be the resulting error. The probability of global error will be exponentially small in N, as we will show, and we can consider the whole Gaussian likelihood function Eq.(21) for $p_l(\mathbf{r}|x)$. Since the noise is isotropic, when integrating over it the average magnitude of the projection onto a fixed unit vector will be simply the variance, and we can write the local error as

$$\langle\varepsilon_l^2\rangle_W = \langle\int dx\frac{\eta^2}{\|\mathbf{v}'(x)\|_2^2}\rangle_W. \tag{38}$$

Computing the derivative of the tuning curves we obtain

$$\begin{aligned}\|\mathbf{v}'(x)\|_2^2 &= \frac{1}{Z^2}\sum_i\sum_{jj'}W_{ij}W_{ij'}\frac{(x-c_j)(x-c_{j'})}{\sigma^4}\exp\Big(-\frac{(x-c_j)^2+(x-c_{j'})^2}{2\sigma^2}\Big)\\ &\approx \frac{\sum_j\exp\Big(-\frac{(x-c_j)^2}{\sigma^2}\Big)}{Z^2\sigma^4} \approx \frac{N\sqrt{\pi\sigma^2}}{2Z^2\sigma^2},\end{aligned} \tag{39}$$

where we took the average over the weights $\langle \sum_{i=1}^{N} W_{ij} W_{ij'} \rangle_W = \frac{N}{L} \delta_{jj'}$ [4] and we substituted the sum with an integral $\sum_j \approx \frac{1}{L} \int dc_j$ (ignoring edge effects when x is not far from the borders). Considering the limit of small σ for Z^2, we finally obtain the local error

$$\langle \varepsilon_l^2 \rangle_W \approx \frac{2\sigma^2 \eta^2}{N}. \tag{40}$$

Note that this quantity corresponds to the inverse of the linear FI, as predicted by the CRAO bound.

Global error. We defined an error as global when the estimate of the stimulus is further than σ from the true value. In this case, we can make the same reasoning of the uncorrelated case, noticing that once we obtain an error of this kind, its average magnitude is uncorrelated with its probability and independent from the noise magnitude $\mathbf{r}$. Therefore we can write, similarly to Eq.(28), the expression for the global error

$$\langle \varepsilon_g^2 \rangle_W = \langle P(\varepsilon) \rangle_W \langle \int dx (\hat{x} - x)^2 \rangle_W. \tag{41}$$

We can assume that in such a case the estimate will be uniformly distributed in the interval $\hat{x} \notin [x - \sigma, x + \sigma]$, and obtain for the average magnitude of global error

$$\bar{\varepsilon}_g = \int dx \int d\hat{x} p(\hat{x}) (\hat{x} - x)^2 \approx \frac{1}{6} + O(\sigma), \tag{42}$$

where we underlined the fact that is a term of order 1 plus corrections of order σ. Finally, we have to compute the probability that, given a stimulus x, the error will be global. This quantity again will not depend from the specific choice of the stimulus. Computing this probability rigorously is hard, due to the correlations between nearby responses. Nevertheless, we know that at a distance of $\sim \sigma$ the responses to two stimuli are uncorrelated. We can therefore imagine to divide the manifold into $\frac{1}{\sigma}$ discrete correlation 'clusters' of responses: we will have a global error when the estimate of the stimulus belong to a cluster other than the true response. We computed the probability of having an error with uncorrelated responses in the previous section, Eq.(35). We simply have to substitute to L the actual number of uncorrelated clusters $\frac{1}{\sigma}$, obtaining for the global error

$$\langle \varepsilon_g^2 \rangle_W \approx \frac{\bar{\varepsilon}_g}{\sigma \sqrt{2\pi N}} \exp\Big(- \frac{N}{2} \log(\frac{1 + 2\eta^2}{2\eta^2}) \Big). \tag{43}$$

Input noise. We considered the case in which the first layer responses are affected by i.i.d Gaussian noise $\tilde{\mathbf{u}}(x) = \mathbf{u}(x) + \mathbf{z^u}$, with $\mathbf{z^u} \sim \mathcal{N}(0, \xi^2 \mathbf{I})$. This results in a multivariate Gaussian distribution for the responses of the second layer, Eq.(22), with covariance matrix $\Sigma = \eta^2 \mathbf{I} + \xi^2 \mathbf{W} \mathbf{W}^T$. The matrix $\mathbf{W}\mathbf{W}^T$ follow the well known Wishart distribution [**?**], with mean $\mathbf{I}$ and fluctuations of the terms of order $\frac{1}{L}$. Therefore the covariance matrix can be rewritten as the sum of the identity plus a perturbation

$$\Sigma = \tilde{\eta}^2 \mathbf{I} + \xi^2 (\mathbf{W}\mathbf{W^T} - \mathbf{I}), \tag{44}$$

introducing an effective noise variance, which is the sum of input and output noise variance $\tilde{\eta}^2 = \eta^2 + \xi^2$. In order to obtain an estimate of the effects of input noise on the local error, we consider the FI as a lower bound to the MSE; the linear FI is computed as

$$J(x) = \mathbf{v}'(x)^T \Sigma^{-1} \mathbf{v}'(x), \tag{45}$$

where, again, $\mathbf{v}'(x)$ denotes the derivative of the tuning curve with respect to the stimulus variable. If the perturbation is small, we can approximate the inverse of the correlation matrix at the second order
$\Sigma^{-1} \approx \frac{1}{\tilde{\eta}^2} \mathbf{I} - \frac{\xi^2}{\tilde{\eta}^4} (\mathbf{W}\mathbf{W}^T - I) + \frac{\xi^4}{\tilde{\eta}^6} (\mathbf{W}\mathbf{W^T} - \mathbf{I})^2$, and write the FI as:

[4] Note that we are approximating the average of the inverse with the inverse of the average, but as soon as N is not too small the two quantities are very similar.

$$
\begin{aligned}
J(x) &= J^{ind}(x) - \delta J(x) \\
&= \frac{\sum_i v_i'^2(x)}{\tilde{\eta}^2} - \frac{\xi^2}{\tilde{\eta}^4}\mathbf{u'}^T(x)(\mathbf{A}^2 - \mathbf{A})\mathbf{u'}(x) + \frac{\xi^4}{\tilde{\eta}^6}\mathbf{u'}^T(x)(\mathbf{A}^3 - 2\mathbf{A}^2 + \mathbf{A})\mathbf{u'}(x),
\end{aligned} \tag{46}
$$

where $\mathbf{A} = \mathbf{W^T W}$ and we used the matrix notation $\mathbf{v}(x) = \mathbf{W}\mathbf{u}(x)$. We recognize in the first term $J^{ind}(x)$ the FI in the case of second layer responses affected by i.i.d. Gaussian noise. All the correction terms to the FI are related to the moments of the matrix $\mathbf{A} = \mathbf{W^T W}$. Since all the entries are Gaussian, it is possible to compute their mean through Isserlis' Wick? theorem. Using the fact that $E[W_{ij}W_{mn}] = \frac{1}{L}\delta_{im}\delta_{jn}$, we obtain:

$$
\begin{aligned}
E[A_{mn}] &= E[\sum_{j=1} NW_{jm}W_{jn}] = \frac{N}{L}\delta_{mn} \\
E[(A^2)_{mn}] &= E[\sum_{i=1}^{L}\sum_{j=1,j'=1}^{N} W_{jm}W_{ji}W_{j'i}W_{j'n}] = (\frac{N}{L} + \frac{N^2}{L^2} + \frac{N}{L^2})\delta_{mn} \\
E[(A^3)_{mn}] &= E[\sum_{i=1,i'=1}^{L}\sum_{j=1,j'=1,j''=1}^{N} W_{jm}W_{ji}W_{j'i}W_{j'i'}W_{j''i'}W_{j''n}] = (\frac{N^3}{L^3} + 3\frac{N^2}{L^3} + 3\frac{N^2}{L^2} + 4\frac{N}{L^3} + 3\frac{N}{L^2} + \frac{N}{L})\delta_{mn}
\end{aligned} \tag{47}
$$

As a result, the mean of the perturbation term (using just the higher powers of $\frac{N}{L}$)

$$
\langle \delta J(x) \rangle_W = \frac{\xi^2}{\tilde{\eta}^4}\frac{N^2}{L^2}\mathbf{u'}(x)^T\mathbf{I}\mathbf{u'}(x) - \frac{\xi^4}{\tilde{\eta}^6}\frac{N}{L}\mathbf{u'}(x)^T\mathbf{I}\mathbf{u'}(x). \tag{48}
$$

Substituting the sum over the indices with an integral, similarly to what we have done in Eq.(39), we obtain the mean FI

$$
\langle J(x) \rangle_W \approx \frac{N\sqrt{\pi\sigma^2}}{2Z^2\sigma^2\tilde{\eta}^2}(1 - \frac{N}{L}\frac{\xi^2}{\tilde{\eta}^2} + \frac{N}{L}\frac{\xi^4}{\tilde{\eta}^4}), \tag{49}
$$

and consequently an approximation to the MSE

$$
\langle \varepsilon^2 \rangle_W \approx \frac{1}{\langle J(x) \rangle_W} \approx \varepsilon_{l,i}^2(1 + \frac{N}{L}\frac{\xi^2}{\tilde{\eta}^2} - \frac{N}{L}\frac{\xi^4}{\tilde{\eta}^4}). \tag{50}
$$

Similar computations can be done assuming a covariance matrix with the same statistic, but not related to the synaptic weights. For example, assuming $\Sigma_{rand} = \eta^2 I + \xi^2 \mathbf{X}\mathbf{X}^T$ with $X_{ij} \sim \mathcal{N}(0, \frac{1}{L})$ similarly to W, but with uncorrelated entries $E[X_{ij}W_{mn}] = 0$. In this case we have no more first order corrections, and the FI increases,

$$
\langle J(x) \rangle_{W,X} \approx \frac{N\sqrt{\pi\sigma^2}}{2Z^2\sigma^2\tilde{\eta}^2}(1 + \frac{N}{L}\frac{\xi^4}{\tilde{\eta}^4}). \tag{51}
$$

4.4 Extension to multidimensional stimuli

A straightforward generalization is to consider a stimulus $\mathbf{x} \in [0,1]^K$ encoded by a first layer of M neurons. We considered the scalar error $\varepsilon^2 = \sum_k \varepsilon_k^2$ as loss function. Similarly to the previous case, the local error along each dimension is computed expanding linearly the tuning curves

$$
\|\mathbf{v}(\mathbf{x} + \Delta x_k) - \mathbf{v}(\mathbf{x})\|_2^2 \approx \left\| \frac{\partial}{\partial x_k}\mathbf{v}(\mathbf{x}) \right\|_2^2 (\Delta x_k)^2. \tag{52}
$$

In an analogous manner, we consider global every error such that $\|\hat{\mathbf{x}} - \mathbf{x}\|_2^2 > \sigma$.

Pure case. In the case the first layer is made up by pure cells, neurons are sensitive to only one stimulus dimension. We assumed their tuning curves to be 1D Gaussian functions $u_{j_k}(\mathbf{x}) = \frac{1}{Z_p}\exp\left(-\frac{(x_k - c_{j_k})^2}{2\sigma^2}\right)$ with preferred positions arranged uniformly along each dimension, $c_{j_k} = \frac{j_k}{L}$ for $j_k = 1, ..., L$ and $L = M/K$. The second layer tuning curves are given by the linear superposition of uncorrelated Gaussian processes along each dimension $v_i^p(\mathbf{x}) = \sum_k \sum_{j_k} W_{ij_k}u_{j_k}(\mathbf{x})$.

Using the same constraint as before, we obtain $Z_p^2 = (\pi\sigma^2)^{1/2} - 2\pi\sigma^2$. In this case each dimension is encoded separately. The tuning curves along one dimension change only by translation $v_i(x_1 + \Delta x_1) = c + v_i(x_1)$, and therefore the local error along each dimension is independent. The squared norm of the derivative along one dimension is reduced by a factor of K (the derivative along each dimension will act only on $1/K$ terms), and consequently the local error along each dimension is

$$\langle \varepsilon_{l,p,k}^2 \rangle_W = \frac{2KZ_p^2\sigma^2\eta^2}{N(\pi\sigma^2)^{1/2}} \approx \frac{2K\sigma^2\eta^2}{N}. \tag{53}$$

Also the probability of having a global error is independent along each dimension. We can approximate the total probability of having a global error as the sum of probabilities along each dimension $P(\varepsilon_g) = \sum_k P(\varepsilon_{g,k})$. Since in this case tuning curves are described by a superposition of uncorrelated Gaussian processes and each dimension contributes equally to the variance, we obtain for the global error in the pure case

$$\langle \varepsilon_{g,p}^2 \rangle_W \approx \frac{K\bar{\varepsilon}_g}{\sigma\sqrt{2\pi N}} \exp\Big(-\frac{N}{2K} \log(\frac{1+2\eta^2}{2\eta^2}) \Big) \tag{54}$$

where the average magnitude of global error, $\bar{\varepsilon}_g$, is again a term of order 1.

Conjunctive case. In the conjunctive case the first layer neurons' responses are given by multidimensional Gaussian functions $u_j(\mathbf{x}) = \frac{1}{Z_c} \exp\Big(-\frac{\|\mathbf{x}-\mathbf{c}_j\|_2^2}{2\sigma^2} \Big)$ with preferred positions arranged on a K dimensional square grid of side $1/L$ with $L = M^{1/3}$. The tuning curves of the second layer neurons $v_i^c(\mathbf{x}) = \sum_j W_{ij} u_j(\mathbf{x})$ are multidimensional Gaussian processes with K-dimensional covariance function $< v(\mathbf{x})v(\mathbf{x}+\Delta\mathbf{x}) >= \frac{1}{Z_c^2} \exp\Big(-\frac{\|\Delta\mathbf{x}\|_2^2}{2\sigma^2} \Big)$. The normalization term is given by $Z_c^2 = (\pi\sigma^2)^{K/2} - (2\pi\sigma^2)^K$ (note that increasing the dimensionality of the stimulus, the edge effects become more relevant). In this case the derivative along one dimension will act on all the terms of the random sum, and the resulting local error is given by

$$\varepsilon_{l,c,k}^2 = \frac{2Z_c^2\sigma^2\eta^2}{N(\pi\sigma)^{K/2}} \approx \frac{2\sigma^2\eta^2}{N} \tag{55}$$

To compute the global error we simply extend the reasoning about uncorrelated clusters. Since stimuli evoke a correlated response within a radius of $\sim \sigma$, the number of uncorrelated clusters scale as $\frac{1}{\sigma^K}$, and the global error is given by

$$\varepsilon_{g,c}^2 \approx \frac{\bar{\varepsilon}_g}{\sigma^K\sqrt{2\pi N}} \exp\Big(-\frac{N}{2} \log(\frac{1+2\eta^2}{2\eta^2}) \Big). \tag{56}$$

4.5 Data analysis and model fitting

The detailed data description is reported in [21]. It consists in $N \sim 500$ neurons' responses (firing rates) recorded during an arm posture "hold" task at 27 different positions (and with 2 hand orientation, up and down) arranged on a virtual cube of size 40x40x40 cm. The response of each neuron for each position is recorded for several trials ($\sim$ 10 trials per position) and the tuning curves are computed averaging over trials. We considered the tuning curves just in function of the position, ignoring the difference in hand orientation. We chose to analyze just "tuned neurons", cells responding with at least 5 spikes/s at more than two positions. We mean-centered and standardized the tuning curves to have variance =1. In order to measure the level of irregularity of one tuning curve in a non parametric form, the authors of [21] decided to introduce a complexity measure. For each neuron, it is defined as the standard deviation of the discrete derivative between the response at one target and its response at the closest target

$$c(D_{min})_i = std\Big(\frac{\|r(x) - r(x+\Delta x)\|}{\sqrt{\|\Delta x\|^2}} s.t. \, \|\Delta x\|_2^2 < D_{min} \Big). \tag{57}$$

In the data, the D_{min} is imposed by the experiment and is equal to 35. This limitation, inherent to the data themselves, prevent us from capturing high frequency components due to aliasing phenomena.

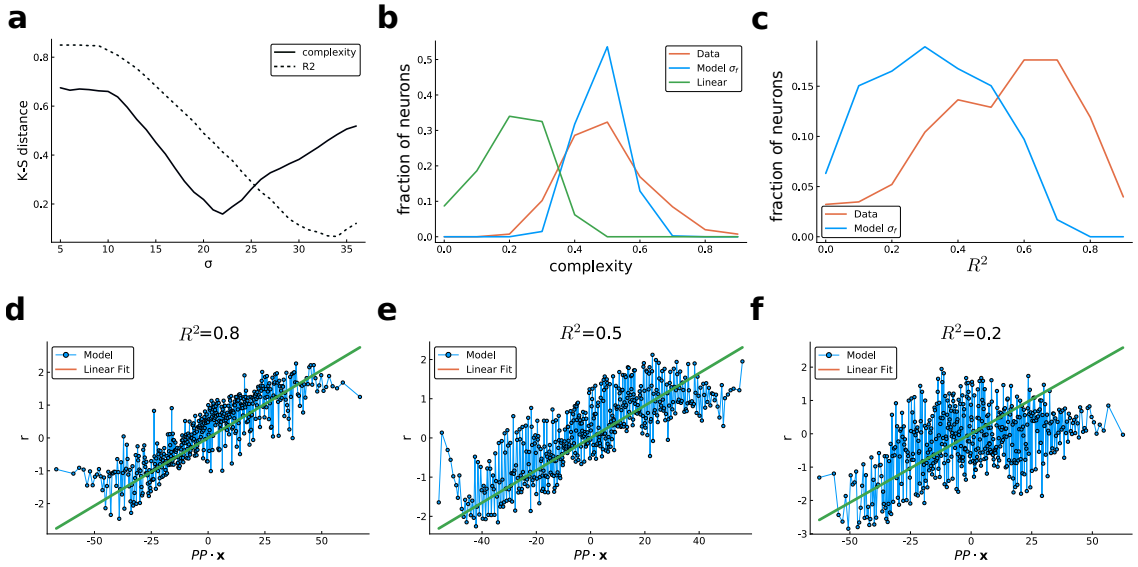

Figure 8: **Model fitting and tuning curves.** (**a**) Kolmogorov-Smirnov distance between the distributions of complexity measure (full line) and R^2 of fitting (dashed) across neurons from the data and the model at different σ. σ_f is chosen to be the value at which the minimum of the distance between complexity distributions is attained, $\sigma_f \sim 22$. (**b**) Normalized histogram of the distribution of complexity measure (arbitrary units) across the neurons of the data (red), the irregular population at σ_f (blue) and a linear population (green). The model is able to capture the bulk of the distribution of the real data much better than a linear model. Nevertheless, the data show a much broader distribution across the population. (**c**) Normalized histogram of the distribution of the R^2 of linear fit across neurons of the data and the irregular population at σ_f (red). Both distributions are broad, but the data show a more consistent linear part. (**d-f**) Three examples of tuning curves of the irregular population at σ_f, showing a broad range of behavior with respect to the linear fit. The tuning curves are plotted in function of the projection of the stimulus (target position) onto a preferred position, obtained by the fit with Eq.(10) (green line). Some neurons are well described by the parametric function (d), some others show consistent deviations (e), while in others the linear behavior is absent (f). This is reflected in the broadness of the distribution of the R^2.

The irregular population was constructed with a 3D model with a first layer of conjunctive cells. To be faithful with the paper and to avoid loss of coverage and boundary effects, we used $M = 100^3$, tiling a 100 by 100 by 100 cube with a grid of side 2. For the connectivity matrix $\mathbf{W}$ we used a sparse random matrix (sparsity = 0.1) with Gaussian entries. The tuning curves were normalized one by one to have variance equal to 1. The only tunable parameter of this model is σ (similarly to the simplest model in the original paper). To find σ_f, we generated the responses of the model to the same 27 stimuli of the real data. We then computed the distribution of the complexity measure (in a.u.) at different σ and we picked σ_f such that the Kolmogorov-Smirnov distance between the distribution of the model and the one of the data is minimal, Fig. 8a. At this optimal σ, the two distributions are very similar, even if real data show a broader distribution of values in both directions; for comparison, a linear model suffers an heavy underestimate of the complexity values across all the populations, Fig. 8b.

The other summary statistic used in the paper is the distribution of R^2 values resulting from the fit with the linear model of Eq.(10),

$$R_i^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_x (r_l(\mathbf{x}) - r(\mathbf{x}))^2}{\sum_x r(\mathbf{x})^2}, \tag{58}$$

where $r(\mathbf{x})$ is the response at stimulus $\mathbf{x}$ and $r_l(\mathbf{x})$ is the response predicted by the linear model. For the sake of completeness, we computed the K-S distance between the model and the data also for this measure, Fig. 8a, red line. The difference in the two distributions simply decreases with σ. The model at σ_f underestimate the linear components of the tuning curves, Fig. 8c. Nevertheless, this is expected since our model has no non linearity, which potentially increases the illusion of linear tuning. It is worth noticing that in the original paper the simpler model (with threshold non linearity) still underestimates the distribution of R^2 values and only the complexity measure was considered in the fitting procedure. The authors obtained a good agreement only considering a more complicate model with more parameters (namely, different thresholds for each neuron and different widths in the first layer).

For numerical simulations in Fig. 6, the tuning curves were computed at a much finer scale than the data (cubic grid of 21 by 21 by 21 points). As expected, the tuning curves show a broad range of behavior with respect to the linear fit, that goes from very linear to very irregular, Fig. 8d-f. The linear population for the comparisons was constructed sampling the preferred positions $((a_1, a_2, a_3)$) uniformly in the unit sphere and using Eq.(10) to generate the responses. Again, the dynamic range (variance of responses) was constrained to be =1.

For Fig. 6d we extracted the noise from the data, assigning to each neuron a noise variance in the following way. For a single neuron, we computed its dynamic range as the variance of the responses across all possible stimuli: $\mathrm{Var}(r) = \langle r^2 \rangle_x - \langle r \rangle_x^2$. Then, we computed the mean variance of the trial to trial variability across all stimuli: $\mathrm{Var}(\eta) = \langle \mathrm{Var}(r(x)) \rangle_x$. Since our tuning curves in the simulations have a dynamic range =1, we assigned to neuron i a variance of the noise equal to $\sigma_{\eta_i}^2 = \frac{\mathrm{Var}(\eta_i)}{\mathrm{Var}(r_i)}$. The decoding error for a population size of N neurons was computed averaging over 8 independent pools of N neurons, each one associated with its noise variance. Also the decoder, Eq.(27), was modified to keep into account each neuron's noise variance. In principle, the noise may be dependent from the mean. To control for this effect, we also preprocessed the data with a variance stabilizing transformation (substituting $r(\mathbf{x})$ with $\sqrt{r(\mathbf{x})}$, [**?**]). The distribution of the noise variance across neurons does not vary substantially. The data are publicly available at https://osf.io/u57df/.

5 Acknowledgements

References

[1] K. Zhang and T. J. Sejnowski, "Neuronal tuning: To sharpen or broaden?," *Neural Computation*, vol. 11, no. 1, pp. 75–84, 1999.

[2] K. Kang, R. M. Shapley, and H. Sompolinsky, "Information Tuning of Populations of Neurons in Primary Visual Cortex," *Journal of Neuroscience*, vol. 24, no. 15, pp. 3726–3735, 2004.

[3] S. Deneve, P. E. Latham, and A. Pouget, "Reading population codes: A neural implementation of ideal observers," *Nature Neuroscience*, vol. 2, no. 8, pp. 740–745, 1999.

[4] M. A. Montemurro and S. Panzeri, "Optimal tuning widths in population coding of periodic variables," *Neural Computation*, vol. 18, no. 7, pp. 1555–1576, 2006.

[5] S. Yaeli and R. Meir, "Error-based analysis of optimal tuning functions explains phenomena observed in sensory neurons," *Frontiers in Computational Neuroscience*, vol. 4, no. October, pp. 1–16, 2010.

[6] M. Fiscella, F. Franke, K. Farrow, J. Müller, B. Roska, R. A. da Silveira, and A. Hierlemann, "Visual coding with a population of direction-selective neurons," *Journal of Neurophysiology*, vol. 114, no. 4, pp. 2485–2499, 2015.

[7] W. Wang, S. S. Chan, D. A. Heldman, and D. W. Moran, "Motor cortical representation of position and velocity during reaching," *Journal of Neurophysiology*, vol. 97, no. 6, pp. 4258–4270, 2007.

[8] H. B. Barlow, "Possible Principles Underlying the Transformations of Sensory Messages," *Sensory Communication*, pp. 216–234, 2013.

[9] M. Chalk, O. Marre, and G. Tkačik, "Toward a unified theory of efficient, predictive, and sparse coding," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 115, no. 1, pp. 186–191, 2018.

[10] D. Ganguli and E. P. Simoncelli, "Efficient sensory encoding and Bayesian inference with heterogeneous neural populations," *Neural Computation*, 2014.

[11] N. Brunel and J. P. Nadal, "Mutual Information, Fisher Information, and Population Coding," *Neural Computation*, 1998.

[12] M. Bethge, D. Rotermund, and K. Pawelzik, "Optimal short-term population coding: When Fisher information fails," *Neural Computation*, vol. 14, no. 10, pp. 2317–2351, 2002.

[13] P. Berens, A. S. Ecker, S. Gerwinn, A. S. Tolias, and M. Bethge, "Reassessing optimal neural population codes with neurometric functions," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 11, pp. 4423–4428, 2011.

[14] C. E. Shannon, "Communication in the Presence of Noise," *Proceedings of the IRE*, vol. 37, no. 1, pp. 10–21, 1949.

[15] I. R. Fiete, Y. Burak, and T. Brookings, "What grid cells convey about rat location," *Journal of Neuroscience*, vol. 28, no. 27, pp. 6858–6871, 2008.

[16] S. Sreenivasan and I. Fiete, "Grid cells generate an analog error-correcting code for singularly precise neural computation," *Nature Neuroscience*, vol. 14, no. 10, pp. 1330–1337, 2011.

[17] A. Mathis, A. V. Herz, and M. B. Stemmler, "Resolution of nested neuronal representations can be exponential in the number of neurons," *Physical Review Letters*, vol. 109, no. 1, pp. 1–5, 2012.

[18] O. Barak, M. Rigotti, and S. Fusi, "The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off," *Journal of Neuroscience*, vol. 33, no. 9, pp. 3844–3856, 2013.

[19] S. Fusi, E. K. Miller, and M. Rigotti, "Why neurons mix: High dimensionality for higher cognition," *Current Opinion in Neurobiology*, vol. 37, no. April, pp. 66–74, 2016.

[20] M. Rigotti, O. Barak, M. R. Warden, X. J. Wang, N. D. Daw, E. K. Miller, and S. Fusi, "The importance of mixed selectivity in complex cognitive tasks," *Nature*, vol. 497, no. 7451, pp. 585–590, 2013.

[21] H. Lalazar, L. F. Abbott, and E. Vaadia, "Tuning Curves for Arm Posture Control in Motor Cortex Are Consistent with Random Connectivity," *PLoS Computational Biology*, vol. 12, no. 5, pp. 1–27, 2016.

[22] A. Finkelstein, N. Ulanovsky, M. Tsodyks, and J. Aljadeff, "Optimal dynamic coding by mixed-dimensionality neurons in the head-direction system of bats," *Nature Communications*, vol. 9, no. 1, 2018.

[23] R. A. Andersen, G. K. Essick, and R. M. Siegel, "Encoding of Spatial Location by Posterior Parietal Neurons Author(s): Richard A. Andersen, Greg K. Essick and Ralph M. Siegel Source:," *Science*, vol. 230, no. 4724, pp. 456–458, 1985.

[24] T. Arakaki, G. Barello, and Y. Ahmadian, *Inferring neural circuit structure from datasets of heterogeneous tuning curves*, vol. 15. 2019.

[25] X.-X. Wei, J. Prentice, and V. Balasubramanian, "The Sense of Place: Grid Cells in the Brain and the Transcendental Number e," pp. 1–17, 2013.

[26] V. Pernice and R. A. da Silveira, "Interpretation of correlated neural variability from models of feed-forward and recurrent circuits," *PLoS Computational Biology*, vol. 14, no. 2, pp. 1–26, 2018.

[27] L. F. Abbott, E. T. Rolls, and M. J. Tovee, "Representational capacity of face coding in monkeys," *Cerebral Cortex*, vol. 6, no. 3, pp. 498–505, 1996.

[28] F. Bouchacourt and T. J. Buschman, "A Flexible Model of Working Memory," *Neuron*, vol. 103, no. 1, pp. 147–160, 2019.

[29] M. Shamir and H. Sompolinsky, "Implications of neuronal diversity on population coding," *Neural Computation*, vol. 18, no. 8, pp. 1951–1986, 2006.

[30] S. Saxena and J. P. Cunningham, "Towards the neural population doctrine," *Current Opinion in Neurobiology*, vol. 55, pp. 103–111, 2019.

[31] J. A. Gallego, M. G. Perich, L. E. Miller, and S. A. Solla, "Neural Manifolds for the Control of Movement," *Neuron*, vol. 94, no. 5, pp. 978–984, 2017.

[32] J. P. Cunningham and B. M. Yu, "Dimensionality reduction for large-scale neural recordings," *Nature Neuroscience*, vol. 17, no. 11, pp. 1500–1509, 2014.

[33] D. Kobak, J. L. Pardo-Vazquez, M. Valente, C. K. Machens, and A. Renart, "State-dependent geometry of population activity in rat auditory cortex," *eLife*, vol. 8, pp. 1–27, 2019.

[34] C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris, "High-dimensional geometry of population responses in visual cortex," *Nature*, vol. 571, no. 7765, pp. 361–365, 2019.

[35] P. Gao, E. Trautmann, B. Yu, G. Santhanam, S. Ryu, K. Shenoy, and S. Ganguli, "A theory of multineuronal dimensionality, dynamics and measurement," p. 214262, 2017.

[36] E. Salinas, "How behavioral constraints may determine optimal sensory representations," *PLoS Biology*, vol. 4, no. 12, pp. 2383–2392, 2006.

[37] S. Qin, Q. Li, C. Tang, and Y. Tu, "Optimal compressed sensing strategies for an array of nonlinear olfactory receptor neurons with and without spontaneous activity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 116, no. 41, pp. 20286–20295, 2019.

[38] B. Babadi and H. Sompolinsky, "Sparseness and Expansion in Sensory Representations," *Neuron*, vol. 83, no. 5, pp. 1213–1226, 2014.

[39] M. Farrell, S. Recanatesi, R. C. Reid, S. Mihalas, and E. Shea-Brown, "Autoencoder networks extract latent variables and encode these variables in their connectomes," *bioRxiv*, p. 2020.03.04.977702, 2020.

[40] D. Sussillo and L. F. Abbott, "Generating Coherent Patterns of Activity from Chaotic Neural Networks," *Neuron*, vol. 63, no. 4, pp. 544–557, 2009.