in which representation neurons exhibit correlated Gaussian noise, with a covariance matrix that has the same statistic as those of Σ, but in which the form of correlations is not inherited from the network structure through the synaptic matrix $\mathbf{W}$; specifically, we consider a random covariance matrix, $\Sigma_{\text{rand}} = \tilde{\eta}^2\mathbf{I} + \xi^2\mathbf{X}\mathbf{X}^\mathbf{T}$, where $X_{ij} \sim \mathcal{N}(0, 1/L)$. In this case, noise correlations *suppress* the MSE as compared to the independent case (with Σ_{ind}), because the 'cloud' of possible noisy responses is reoriented randomly with respect to the manifold of mean responses. Analytically, the analog of Eq. (10) for the case of covariance matrix given by Σ_{ind} is similar, but skips the lowest-order, deleterious term:

$$\varepsilon^2_{l,\text{rand}} \approx \varepsilon^2_{l,\text{ind}} \left(1 - \frac{N}{L}\frac{\xi^4}{\tilde{\eta}^4}\right). \tag{11}$$

This result, as well as numerical simulations (Fig. 7B), demonstrate that generically coding is improved by random noise correlations, and that this improvement increases with N and with ξ^2. In sum, noise correlations in representation neurons are deleterious if they are inherited the noise in sensory neurons—yet, the effect is quantitatively modest.

3 Discussion

We analyzed the coding properties of neural populations beyond classical models of tuning curves, by considering irregular response profiles as the result of an unstructured connectivity. The model can interpolate between an irregular coding scheme, locally very accurate but prone to catastrophic errors, and a smooth one, more robust to noise. With a straightforward extension of the model, we studied how limitations in the downstream structure affect the optimal arrangement of multi-dimensional tuning curves, distinguishing between the two extreme cases of 'pure' and 'conjunctive' selectivity (Finkelstein et al., 2018; Harel & Meir, 2020). In particular, one instance of this model for three-dimensional stimuli, may explain the large degree of irregularity found in tuning curves of monkey motor cortex (Lalazar et al., 2016). We illustrated the advantage of such irregularities in coding performance, through the comparison of a neural population generated by our model, fitted to experimental data, and one with an homogeneous, smooth description of tuning curves.

Population coding and geometry of neural responses. A large body of literature addressed the theoretical problem of coding low dimensional stimuli with neural populations featured by simple, parametric tuning curves. In the standard case of homogeneous, bell-shaped tuning curves, a benchmark for the study of *peripheral* sensory neurons, the optimal tuning width was studied in function of the population size, stimulus dimensionality (Zhang & Sejnowski, 1999), stimulus geometry (Montemurro & Panzeri, 2006), time available for decoding (Bethge et al., 2002; Yaeli & Meir, 2010). Moreover, several studies analyzed the region of the stimulus space which was best encoded by a single neuron, showing that it varies from the region of maximal slope (the flanks of the tuning curve) or the region of maximal response (the peak), depending upon the population size and the signal-to-noise ratio (Butts & Goldman, 2006; Yarrow & Series, 2015). Few works considered how heterogeneity of the tuning parameters affects the coding properties of the neural population (Wilke & Eurich, 2002; Shamir & Sompolinsky, 2006; Fiscella et al., 2015). Finally, recent papers showed how such an homogeneous population can be warped to optimally encode stimuli with a non uniform prior distribution (Wei & Stocker, 2012; Ganguli & Simoncelli, 2014b; Yerxa et al., 2020). In this paper, we followed this line of works in examining the optimal tuning width as a function of limitations in downstream areas. This has the merit of applying the efficient coding hypothesis in *deeper* neurons, which usually show complex and heterogeneous tuning properties.

With such a complex, multi-peaked shape of tuning curves, the activity of a single representation neuron is not very informative about the stimulus; rather, the neural population, as a whole, is the relevant unit of computation (Saxena & Cunningham, 2019). In this sense, many coding properties can be derived by analyzing the geometry of the neural responses as stimuli parameters are varied (Fig. 1A,C, Fig. 5B), an approach which has been proved to bring fruitful insights in different brain areas (Fusi et al., 2016; Gallego et al., 2017; Stringer et al., 2019; Kobak et al., 2019). In our setting, by tuning σ, we effectively vary the *intrinsic dimensionality* of the coding manifold, defined as the minimum number of coordinates needed to describe it, interpolating between $\sim N$ in case of random uncorrelated responses and ~ 1 in case of very smooth manifolds. In close relation to our results, the work of Stringer et al. (2019) suggests that the manifold evoked by the joint neural activity of neurons in V1 possesses a fractal-like structure, with progressively less coding resources employed to encode finer details of the stimulus. Such an arrangement was suggested to balance the accuracy, given by fine scale irregularities, and the noise-robustness, given by smoother manifolds. We gave new insights on this idea, by quantifying how the balance between the two instances depends on the number of neural resources and the magnitude of the noise. The kind of manifolds we considered belong to a class defined in Lahiri et al. (2016);

Gao et al. (2017) as *random Gaussian manifolds*. These are of theoretical interest, because they saturate the upper bound on the intrinsic dimensionality, given the smoothness imposed by biological constraints. (Which can be measured experimentally, as the autocorrelation length of responses variability in function of stimulus parameters). Thanks to this property, they can be used as null model to quantify the dimensionality of neural trajectories in experimental data. Our results can be used as a benchmark to be compared with recorded neural populations; we gave an example of this approach by re-analyzing the data from Lalazar et al. (2016).

Combinatorial codes and randomness. At the optimal network configuration, the error decreases exponentially fast with the number of neurons, similarly to observations on the coding of position by grid cells (Fiete et al., 2008; Sreenivasan & Fiete, 2011; Mathis et al., 2012; Wei et al., 2015). Using the terminology of these works, the random coding scheme of neurons in the second layer is an another example of *exponentially strong* population code. This result is tightly related to ideas that were explored already by Claude Shannon (Shannon, 1949). He proposed a geometrical representation of an abstract communication system, as a map which associates points in the space of *messages* (corresponding to stimuli in our case) to points in the space of *signals* (neural activity). The decoding process, which a *receiver* is supposed to do, corresponds to the inverse mapping from signals to messages. By using the one-dimensional message space case as example, he noticed that, in order for such a map to be as efficient as possible, the corresponding one-dimensional curve should wander back and forth through the high dimensional space of signals, to be as long as possible. In this way, the region of uncertainty created by the noise will be small relatively to the length of the line, leading to a higher dynamic range. Nevertheless, this 'signal space filling' map has to be such that the noise does not create large scale ambiguities in the represented message (what he called *threshold effect*, corresponding to global errors in our case). Astonishingly, he showed that this map, which achieves the maximal transmission capacity, need not to be carefully designed, and that optimality can be achieved through random associations between messages and corresponding signals.

The existence in the brain of distributed codes with high (exponential) capacity, and without any evident structure, has been showed in the context of discrete stimuli (Abbott et al., 1996). For example, neural populations in the cortex exhibit great diversity in their responses to face stimuli, and this allows a population of N neurons to encode exponentially many faces. Here, we extended these ideas to continuous stimuli. The treatment of continuous stimuli introduces a notion of magnitude of errors, not present in the context of discrete ones, where the task is simply to discriminate between two different stimuli. This gives rise to the trade-off between local and global errors, constraining the smoothness of the random code.

Compression and expansion in neural systems. Random *divergent* connectivity have been used as a benchmark model to study the *expansion* of low-dimensional, *dense* neural patterns, into high-dimensional, *sparse* representations (Barak et al., 2013; Babadi & Sompolinsky, 2014; Lindsay et al., 2017; Maoz et al., 2020). This process features neurons with the so-called *mixed selectivity* (Rigotti et al., 2013; Fusi et al., 2016). The resulting representations of input patterns facilitates the readout and the associative learning in downstream areas (Litwin-Kumar et al., 2017), and it has been suggested to play a role in the flexibility of working memory (Bouchacourt & Buschman, 2019). Such divergent pathways and mixed selectivity have been observed in many sensory and cortical regions, e.g., prefrontal cortex, cerebellum, insects' mushroom body and hippocampus (Bernardi et al., 2020; Cayco-Gajic & Silver, 2019)

On the other side, in as many cases neural systems exhibit *convergent* pathways, or bottlenecks, where the information encoded in a large population is compressed into a lower number of neurons (Ganguli & Sompolinsky, 2012). In signal processing, techniques for acquiring high-dimensional, sparse signals with a small number of measurements goes under the name of Compressed Sensing (CS) (Donoho, 2006). One of the key results in this field is that, given a high (L)-dimensional signal, which is K-sparse in some basis (meaning that it is possible to express it as a vector with only K components different from 0), it is possible to reconstruct it using a minimal number of noisy 'measurements' (linear projections) which scales only logarithmically with the dimensionality, $N > O\left(K \log\left(L/K\right)\right)$. Notably, the acquisition matrix need not to be carefully designed, as random matrices achieve optimality (Candes & Tao, 2006; Baraniuk et al., 2008; Baraniuk & Wakin, 2009). In neuroscience, the framework has been successfully applied to model the coding properties of neurons in the olfactory pathway: these sensory neurons do not show any evident structure of selectivity to odors, which can be considered as sparse combinations of molecules (Stettler & Axel, 2009; Zhang & Sharpee, 2016; Schaffer et al., 2018; Qin et al., 2019).

The analogy with our setting is clear, as we considered a low-dimensional stimulus x, encoded in the high-dimensional activity of L neurons, then projected onto N neurons. Indeed, by inverting the Eq.(5) to compute the minimal number of random projections, N, such that it is possible to decode the L stimuli with a given error probability, we obtain that this number grows only logarithmically with the number of stimuli. The difference is that the CS task is to reconstruct the high-dimensional vector, while we are not interested in the reconstruction of the pattern of activity of the first layer, but rather in obtaining an estimate of the low

dimensional variable which evoked it. Although in many cases the connectivity of a neural circuit is deeply linked to the underlying functions (Kim et al., 2014; Litwin-Kumar & Turaga, 2019; Farrell et al., 2020), the brain is a complex system and it is plausible that exhibits some 'unstructured' components. These observations, together with the computational properties of random matrices, lead to consider random convergent synapses as a benchmark for the study of compressed neural representations.

Efficient coding criteria and decoders. In order to compute the optimal coding properties of the network, we used the error in the stimulus estimate as obtained from an ideal decoder. The use of this loss function is justified in information theory (Cover & Thomas, 2005) and neuroscience (Salinas & Abbott, 1994; Dayan & Abbott, 2001). Due to the difficulty in treating analytically the MSE, several studies used the Fisher information as a proxy. According to the Cramer-Rao inequality, this quantity sets a lower bound to the variance of an unbiased estimator. Furthermore, it is related to other interesting measures, such as the mutual information (Brunel & Nadal, 1998; Wei & Stocker, 2016; Huang & Zhang, 2019), which was the loss function originally assumed in Barlow's seminal work (Barlow, 1961). Nevertheless, Fisher information is a local quantity and therefore fails in keeping track of global errors; our results show a relevant example where its use leads to wrong conclusions about the optimal coding parameters (Bethge et al., 2002; Yaeli & Meir, 2010; Berens et al., 2011).

In our work, a strong assumption is made about the decoder. The ideal decoder is implementable as a two-layer neural network (see Methods): a first layer computes a discrete approximation of the posterior distribution over stimuli and the second one computes an average, returning therefore the Minimal MSE estimator. A similar network decoder has been used by Ganguli & Simoncelli (2014a); for the similarity with the classical population vector (Georgopoulos et al., 1986), it has been called *Bayesian* population vector. (Since, instead of weighting the preferred stimuli through the relative neural responses, it uses the correct posterior probability). All the required operations, linear filtering, non linearity and normalization, have been assumed as canonical computations in neural circuits (Deneve et al., 1999; Kouh & Poggio, 2008; Carandini & Heeger, 2012). Nevertheless, the ideal parameters of the decoder (i.e., the synaptic weights) depend on the knowledge of the mean neural responses (tuning curves) and noise variance. It is not clear if these ideal parameters can be learnt with simple, biologically plausible, rules, and how global errors and small scale irregularities affect the learning process. In close relation to these considerations, Bordelon et al. (2020) showed how deep networks trained with gradient descent fit firstly low components of the target function, suggesting how irregularities are learnt much slowly, and require an higher number of examples. The impact of limitations in the decoding architecture on the optimal encoding parameters is an interesting question, which we leaves for future research.

4 Methods

Throughout the paper, bold letters denote vectors $\mathbf{r} = \{r_1, r_2, ..., r_N\}$, $\|\mathbf{r}\|_2^2 = \sum_i r_i^2$ represents the L_2 norm, capital bold letters $\mathbf{W}$ denote matrices. Numerical simulations and data analysis were done using a custom code written in Julia (Bezanson et al., 2017).

Model description: one-dimensional stimulus

Random Feedforward Network. We considered a two layer architecture. A one-dimensional 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: