Jointly efficient encoding and decoding in neural populations

Simone Blanco Malerba[1,†], Aurora Micheli[1], Michael Woodford[2], and Rava Azeredo da Silveira[1,3,4]

1 Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS,3 Sorbonne Université, Université de Paris, Paris
2 Department of Economics, Columbia University, New York
3 Institute of Molecular and Clinical Ophthalmology Basel, Basel
4 Faculty of Science, University of Basel, Basel
†Present address: Center for Molecular Neurobiology, University Medical Center Hamburg-Eppendorf, Hamburg

* correspondingauthor@institute.edu

Abstract

The efficient coding approach proposes that neural systems represent as much sensory information as biological constraints allow. It aims at formalizing encoding as a constrained optimal process. A different approach, that aims at formalizing decoding, proposes that neural systems instantiate a generative model of the sensory world. Here, we put forth a normative framework that characterizes neural systems as jointly optimizing encoding and decoding. It takes the form of a variational autoencoder: sensory stimuli are encoded in the noisy activity of neurons to be interpreted by a flexible decoder; encoding must allow for an accurate stimulus reconstruction from neural activity. Jointly, neural activity is required to represent the statistics of latent features which are mapped by the decoder into distributions over sensory stimuli; decoding correspondingly optimizes the accuracy of the generative model. This framework results in a family of encoding-decoding models, which result in equally accurate generative models, indexed by a measure of the stimulus-induced deviation of neural activity from the prior distribution over neural activity. Each member of this family predicts a specific relation between properties of the sensory neurons—such as the arrangement of the tuning curve means (preferred stimuli) and widths (degrees of selectivity) in the population—as a function of the statistics of the sensory world. Our approach thus generalizes the efficient coding approach. Notably, here, the form of the constraint on the optimization derives from the requirement of an accurate generative model, while it is arbitrary in efficient coding models. Finally, we characterize the family of models we obtain through other measures of performance, such as the error in stimulus reconstruction. We find that a range of models admit comparable performance; in particular, a population of sensory neurons with broad tuning curves as observed experimentally yields both low reconstruction stimulus error and an accurate generative model.

Introduction

Normative models in neuroscience describe stimulus representation and information transmission in the brain in terms of optimality principles. Among these, the efficient

coding principle [1] posits that neural responses are set so as to maximize the information about external stimuli, subject to biological resource constraints. Despite this minimal assumption, this hypothesis has been successful in predicting neural responses to natural stimuli in various sensory areas [2–4]. The approach consists in specifying an *encoding* model as a stochastic map between stimuli and neural responses. The parameters of this model are then chosen so as to optimize a function that quantifies the coding performance, e.g., the mutual information between stimuli and neural responses. This optimization is carried out under some metabolic cost proportional, e.g., to the energy needed to emit a spike [5]. The decoding process is assumed to be ideal and is carried out in a Bayesian framework. Prior knowledge about the environment is combined with the evidence from neural activity to form a posterior belief about the stimulus [6, 7].

The idea that the brain is capable of manipulating probabilities and uncertainty dates back to Helmoltz's view of perception as an inference process, in which the brain learns an internal statistical model of sensory inputs [8]. Mathematically, such an internal model can be formalized as a generative model in which stimuli are generated by sampling from a distribution conditioned by one of a set of 'latent,' elementary features [9, 10]. These features can be chosen so as to allow for a semantic interpretation, such as oriented edges or textures in generative models of natural images [11, 12], but this does not have to be the case [13]. It is then assumed that the role of sensory areas is to perform statistical inference by computing the posterior distribution over the latent features conditioned on the sensory observation, thereby 'inverting' the internal model. This posterior distribution is assumed to be represented in the neural activity, and different representation schemes have been proposed [14–16]. As opposed to the efficient coding approach, which prescribes a stochastic mapping from stimulus to neural activity, the generative model approach prescribes a stochastic mapping from neural activity to stimulus. This mapping implies a posterior distribution on neural activity, which can be read off from neural data.

Here, we consider an extended efficient coding approach: while, typically, only the sensory encoding process is optimized, we consider jointly the encoding and decoding processes. In addition to a class of encoding transformations from stimuli to neural responses in a sensory area, we assume a class of generative models implemented downstream. These define maps from neural activity patterns, corresponding to latent variables, to distributions over stimuli. Optimality is achieved when the generative distribution matches the true distribution of stimuli in the environment. If one assumes that the encoder and the decoder are jointly optimized in this framework, the system has the structure of a variational autoencoder (VAE) [17].

Similarly to the classical efficient coding framework, here the encoder is set so as to maximize a variational approximation to the mutual information between stimuli and neural responses under a constraint on the neural resources. However, an important aspect of this formulation is that the constraint, rather than being imposed by hand, is a direct consequence of the assumption of an optimal internal model. This constraint is obtained as the statistical distance between the stimulus-evoked distribution of neural activity and the prior distribution of neural activity assumed by the generative model. The latter, in turn, can be interpreted as the statistics of spontaneous neural activity [18]; the statistical constraint can thus be viewed as the metabolic cost of stimulus-induced deviations from spontaneous neural activity.

We apply our theoretical framework to the study of a population coding model with neurons with classical, bell-shaped tuning curves. By capitalizing on recent advances in the VAE literature, we solve the optimization problem as a function of the constraint on neural resources: we obtain a family of solutions which yield equally satisfying generative models [19]. However, these solutions make different predictions about the

corresponding neural representations, which correspond to different arrangements of tuning curves, statistics of prior over neural activity, and coding performances. Related approaches have been explored in the literature, and predictions about the optimal allocation of coding resources, i.e., the tuning curves, as a function of the stimulus distribution have been derived [6, 20]. We examine how, in our framework, the optimal allocation of coding resources as a function of the statistics of stimuli varies as a function of the constraint. Despite the differences in the objective function, our results are consistent with previous predictions in a weakly-constrained regime, while more complex behaviors arise in a strongly-constrained regime. Our results illustrate how the interactions between the encoder and the internal model shape neural representations of sensory stimuli.

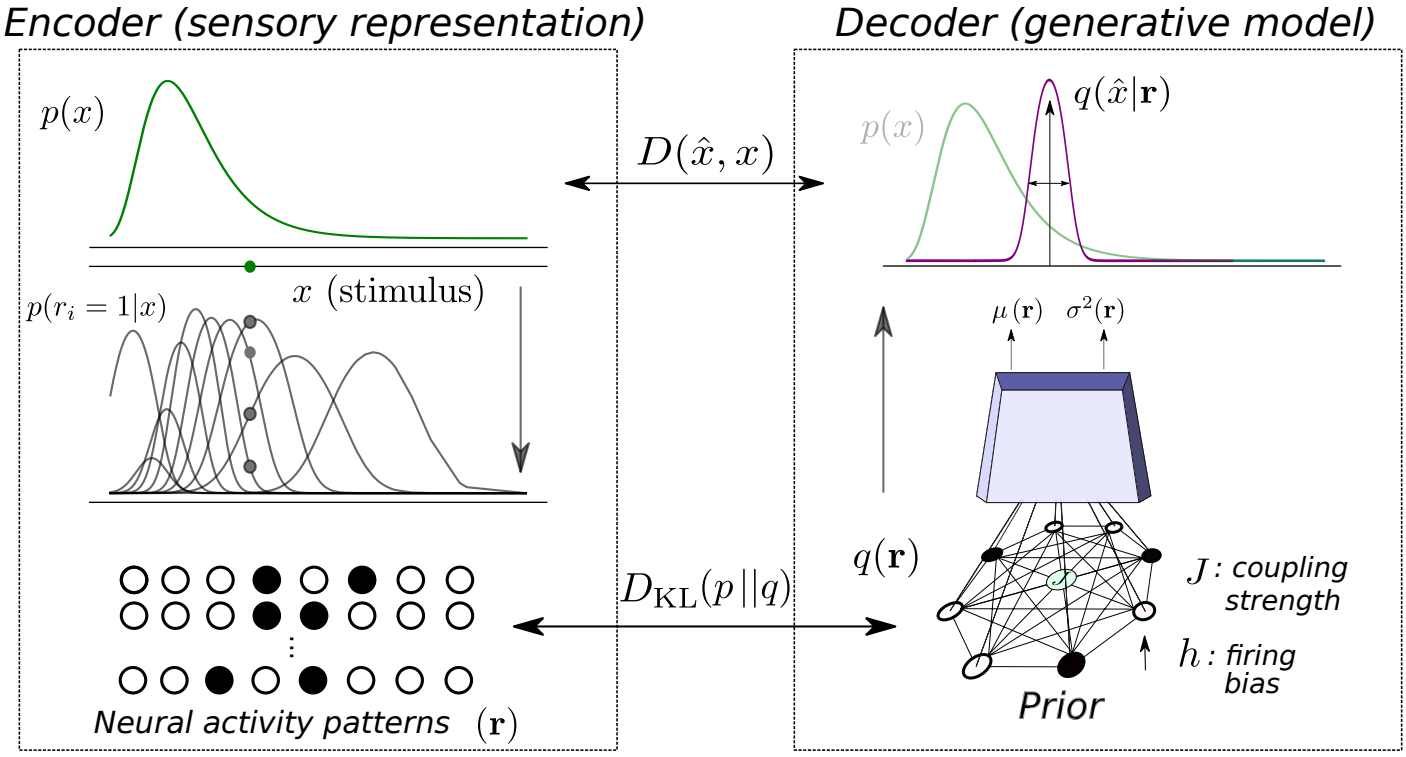

Fig 1. Model architecture. Left: encoder, or sensory representation. Neurons emit spikes according to bell-shaped tuning curves in response to a stimulus, x, drawn from the distribution $p(x)$ (green curve). The population response consists in a neural activity pattern, $\mathbf{r}$. Right: decoder, or generative model. The generative model maps neural activity patterns, sampled from the prior distribution (a Boltzmann machine), $q_\psi(\mathbf{r})$, to parameters, μ and σ, of a Gaussian distribution over stimuli, $q_\psi(x|\mathbf{r})$. When an activity pattern is observed, $q_\psi(x|\mathbf{r})$ is used to obtain an estimate of the stimulus, as well as an associated uncertainty (purple curve). The distortion term drives the system to maximize the likelihood of the observed stimulus given the generative distribution, while the rate term penalizes deviations of the conditional encoding distribution away from the distribution of prior distribution of neural activity of the network.

Materials and methods

In what follows, we denote vectors in bold font and scalars in regular font. We denote by $\langle f(z)\rangle_{p(z)}$ the expectation of a function f of a random variable z distributed

according to $p(z)$,

$$\langle f(z)\rangle_{p(z)} = \int dz p(z) f(z). \tag{1}$$

Encoder (sensory representation)

We consider a population of N neurons responding to a continuous scalar stimulus, x, distributed according to a prior distribution, $p(x)$ (Fig. 1, left). In order to avoid confusion with the prior distribution over neural activity patterns, $q(\mathbf{r})$, defined below, we will refer to $p(x)$ as the data, or stimulus, distribution. We consider neural activity in the limit of short time intervals, such that each neuron either emits one spike or is silent. The set of possible activity patterns is then the set of binary vectors, $\mathbf{r} = (r_1, r_2, ..., r_N)$ where $r_i \in \{0, 1\}$; in what follows, the sum $\sum_{\mathbf{r}} \cdot$ denotes the sum over these 2^N binary patterns. The encoding distribution is the conditional probability distribution over neural activity patterns given the stimulus, $p_\theta(\mathbf{r}|x)$, where θ denotes the set of parameters. We assume neurons to spike independently, such that $p_\theta(\mathbf{r}|x) = \prod_{i=1}^N p_\theta(r_i|x)$.

We consider the limit of small time bins of the Poisson model for spiking neurons [21, 22], by taking into account only the first two terms of the Poisson distribution. With proper normalization, the probability of spiking of a neurons is obtained as

$$p_\theta(r_i = 1|x) = \frac{f_i(x)}{1 + f_i(x)}, \tag{2}$$

where $f_i(x)$ is the neuron's tuning curve. We parametrize tuning curves as Gaussian functions, a shape widely observed in early sensory areas, as

$$f_i(x) = A_i \exp\left(-\frac{(x - c_i)^2}{2w_i^2}\right), \tag{3}$$

with c_i the preferred stimulus of neuron i, w_i the tuning width, and A_i the amplitude. Thus, the probability of spiking of a neuron can be written as $p_\theta(r_i = 1|x) = \mathcal{S}(\eta_i(x))$, with $\eta_i(x) = \frac{(x-c_i)^2}{2w_i^2} - \log A_i$ and $\mathcal{S}(y) = 1/(1 + \exp(-y))$, the logistic function. In the canonical form of the exponential family, the resulting multivariate Bernoulli distribution can be written as

$$p_\theta(\mathbf{r}|x) = \exp\left(\boldsymbol{\eta}(x)^T \mathbf{r} - \sum_{i=1}^N \log\left(1 + e^{\eta_i(x)}\right)\right), \tag{4}$$

with $\boldsymbol{\eta}(x) = (\eta_1(x), ..., \eta_N(x))$ the vector of natural parameters and $\theta = \{A_i, c_i, w_i\}_{i=1}^N$ the set of parameters of the encoder.

Decoder (generative model)

We define an internal model of the environment as a generative model, by specifying a parametric joint probability of neural activity patterns and sensory stimuli, $q_\psi(\mathbf{r}, x)$, where ψ denotes the set of parameters (Fig. 1, right). The neural activity patterns are treated as latent variables, sampled from a prior distribution, $q_\psi(\mathbf{r})$, and mapped to a distribution over stimuli, $q_\psi(x|\mathbf{r})$. As the prior distribution does not depend on the stimulus, we can interpret $q_\psi(\mathbf{r})$ as describing the statistics of the spontaneous neural activity. We model this distribution as the maximum-entropy distribution constrained by the first- and second-order statistics of neural activity, a model which has been proposed as a model of the distribution of activity in neural systems, e.g., in retina and

in cortex [23]. In the case of binary patterns, this maximum-entropy distribution takes the form of an Ising model, or Boltzmann machine,

$$q_\psi(\mathbf{r}) = \exp\left(\mathbf{h^T r} + \mathbf{r^T} J \mathbf{r} - \log Z\right), \tag{5}$$

where $\mathbf{h}$ is the vector of biases, J is the matrix of couplings (with our choice of parametrization, the diagonal elements of J vanish), and $Z = \sum_{\mathbf{r}} \exp(\mathbf{h^T r} + \mathbf{r^T} J \mathbf{r})$ is a normalization constant (also called partition function).

On the basis of experimental findings, it has been suggested that the brain encodes both a stimulus estimate and the associated uncertainty [24, 25]. Thus, we model the generative distribution as a Gaussian, whose mean (stimulus estimate) and variance (uncertainty) are generic functions of neural activity patterns,

$$q_\psi(x|\mathbf{r}) = \mathcal{N}\left(\mu_\phi(\mathbf{r}), \sigma_\phi(\mathbf{r})\right); \tag{6}$$

we parameterize these functions as two-layer neural networks, and we denote by ϕ the set of weights and biases. The parameters of the generative distribution and of the prior, $\psi = \{\phi, \mathbf{h}, J\}$, constitute the set of parameters of the generative model. In this framework, while the encoding distribution and the prior of the generative model are defined on the space neural activity patterns, the generative distribution is defined on the space of stimuli, which can be related to behavioral outputs (stimulus estimate). The neural network, thus, is not intended to be interpreted as a biological neural circuit, but just as a flexible model of the map between neural activity and behavioral output.

Training objective

The internal model is deemed optimal when the output probability distribution, $q_\psi(x) = \sum_{\mathbf{r}} q_\psi(x|\mathbf{r}) q_\psi(\mathbf{r})$, matches the true distribution of stimuli, $p(x)$. We achieve this by setting the parameters of the generative model so as to minimize the Kullback-Leibler (D_{KL}) divergence between the data and the generative distributions,

$$\min_\psi \left\{ D_{\mathrm{KL}}\left(p(x)||q_\psi(x)\right) = H(p) - \langle \log q_\psi(x) \rangle_{p(x)} \right\}, \tag{7}$$

where $H(p)$, the stimulus entropy, does not depend on the parameters. (In what follows, we will denote shortly this divergence by $D_{\mathrm{KL}}(p||q)$). In order to learn the optimal parameters on the basis of a set of data points, we assume a two-stage encoding-decoding process. The encoder maps a stimulus sample, x, to a neural activity pattern, $\mathbf{r}$, according to $p_\theta(\mathbf{r}|x)$. The activity pattern corresponds to a realization of the latent variable in the generative model, and is mapped back ('decoded') to a distribution over stimuli according to $q_\psi(x|\mathbf{r})$. By including the encoder, we can rewrite the second term on the right-hand-side of Eq. (7) as the sum of three terms,

$$\begin{aligned} \langle \log q_\psi(x) \rangle_{p(x)} = \Big\langle & D_{\mathrm{KL}}\left(p_\theta(\mathbf{r}|x)||q_\psi(\mathbf{r}|x)\right) \\ & + \sum_{\mathbf{r}} p_\theta(\mathbf{r}|x) \log q_\psi(x|\mathbf{r}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{r}|x)||q_\psi(\mathbf{r})\right) \Big\rangle_{p(x)}. \end{aligned} \tag{8}$$

The first term in the sum involves the posterior distribution over neural activity patterns, $q_\psi(\mathbf{r}|x) = q_\psi(x|\mathbf{r}) q_\psi(\mathbf{r}) / q_\psi(x)$; calculating $q_\psi(x)$ requires summing over all patterns of activity, $\mathbf{r}$, which is computationally prohibitive. Instead, we use the fact that the D_{KL} divergence is positive, and vanishes only when the two distributions are identical, to convert Eq. (8) into an inequality,

$$\langle \log q_\psi(x) \rangle_{p(x)} \geq \left\langle \sum_{\mathbf{r}} p_\theta(\mathbf{r}|x) \log q_\psi(x|\mathbf{r}) - D_{\mathrm{KL}}\left(p_\theta(\mathbf{r}|x)||q_\psi(\mathbf{r})\right) \right\rangle_{p(x)}. \tag{9}$$

Since the generative distribution, $\log q_\psi(x)$, is often referred to as the 'evidence' for a data point, x, the quantity on the right hand side of Eq. (9) goes by the name of 'evidence lower bound' (ELBO). We note that the maximum value of the ELBO corresponds to minus the stimulus entropy, yielding a vanishing $D_{\rm KL}$ divergence in Eq. (7).

We can then address a variational approximation to the problem in Eq. (7) by maximizing the ELBO. Equivalently, we can optimize the encoder and decoder parameters so as to the minimize the negative ELBO, written as the sum of two terms,

$$\min_{\{\psi,\theta\}} \{-\text{ELBO} = D + R\}\,; \tag{10}$$

borrowing the nomenclature from rate-distortion theory, we call *distortion* the quantity

$$D = \left\langle -\sum_{\mathbf{r}} p_\theta(\mathbf{r}|x) \log q_\psi(x|\mathbf{r}) \right\rangle_{p(x)}, \tag{11}$$

equal to the opposite of the first term on the right-hand-side of Eq. (9), which measures the average log-probability of a stimulus, x, after the encoding-decoding process, and *rate* the quantity

$$R = \langle D_{\rm KL}\left(p_\theta(\mathbf{r}|x)||q_\psi(\mathbf{r})\right)\rangle_{p(x)} = \left\langle \sum_{\mathbf{r}} p_\theta(\mathbf{r}|x) \log\left(\frac{p_\theta(\mathbf{r}|x)}{q_\psi(\mathbf{r})}\right)\right\rangle_{p(x)}, \tag{12}$$

equal to the opposite of the second term, which measures the statistical distance between the encoding distribution and the prior assumed by the generative model. This framework goes by the name of variational autoencoder (VAE) [17]. As one typically does not have access to the true data distribution, but only to a set of samples, the average over $p(x)$ is approximated by an empirical average over a set of P samples, $\langle f(x)\rangle_{p(x)} \approx \sum_{i=1}^{P} f(x_i)/P$.

We note that, due to the fact that the variance of the generative distribution depends on the neural responses, the distortion differs from the more usual mean squared error (MSE) loss function of classical autoencoders, also commonly employed to measure the performance of neural codes. Indeed, here the distortion function is written as

$$D = \left\langle \sum_{\mathbf{r}} p(\mathbf{r}|x) \left(\frac{\left(\mu_\phi\left(\mathbf{r}\right) - x\right)^2}{2\sigma_\phi^2(\mathbf{r})} + \frac{1}{2}\log\left(2\pi\sigma_\phi^2(\mathbf{r})\right)\right)\right\rangle_{p(x)}, \tag{13}$$

while the MSE is obtained as

$$\varepsilon^2 = \left\langle \sum_{\mathbf{r}} p(\mathbf{r}|x) \left(\mu_\phi\left(\mathbf{r}\right) - x\right)^2 \right\rangle_{p(x)}, \tag{14}$$

where we have used the fact that the optimal estimator is given by the mean of the posterior.

In a Bayesian framework, if $q(x|\mathbf{r})$ is an accurate approximation of the posterior distribution of stimulus given neural responses, its mean approximates the minimum MSE estimate. In the Results section, we also consider the MSE obtained when the stimulus estimate, $\hat{x}$, is sampled from the posterior distribution, $\hat{x} \sim q(\hat{x}|\mathbf{r})$, as

$$\begin{aligned} \varepsilon^2_{\rm sampling} &= \left\langle \sum_{\mathbf{r}} p(\mathbf{r}|x) \int d\hat{x} q(\hat{x}|\mathbf{r}) \left(\hat{x} - x\right)^2 \right\rangle_{p(x)} \\ &= \left\langle \sum_{\mathbf{r}} p(\mathbf{r}|x) \left[\left(\mu_\phi(\mathbf{r}) - x\right)^2 + \sigma_\phi^2(\mathbf{r})\right] \right\rangle_{p(x)}. \end{aligned} \tag{15}$$

Constrained optimization and connection with efficient coding

It is a known issue in the VAE literature that, when the generative distribution is flexible given the data distribution (meaning that $q_\psi(x|\mathbf{r})$ has enough degrees of freedom to approximate complex distributions), the ELBO optimization problem exhibits multiple solutions (Fig. 3). Optimization algorithms based on stochastic gradient descent are biased towards solutions with low rate and high distortion, a phenomenon which goes by the name of posterior collapse [19,26]. In the extreme case, the model relies entirely on the power of the decoder and ignores the latent variables altogether: all realizations of the latent variables are mapped to the data distribution, $q_\psi(x|\mathbf{r}) \approx p(x)$, and, consequently, all stimuli are mapped to the same representation, $p_\theta(\mathbf{r}|x) \approx q_\psi(\mathbf{r})$.

We overcome this issue by addressing a related constrained optimization problem. We minimize the distortion subject to a maximum, or 'target,' value of the rate, $\bar{R}$:

$$\begin{aligned} \min_{\{\theta,\psi\}} \quad & D \\ \text{subject to} \quad & R \leq \bar{R}. \end{aligned} \tag{16}$$

The set of parameters $\{\theta, \psi\}$ that satisfy the constraint $R \leq \bar{R}$ is called feasible set. By writing the associated Lagrangian function with multiplier $\beta \geq 0$, we have that

$$\max_{\beta\geq 0} \left\{ L(\theta,\psi,\beta) = D + \beta(R-\bar{R}) \right\} = \begin{cases} D & \text{if } \{\theta,\psi\} \text{ is feasible} \\ \infty & \text{otherwise} \end{cases}. \tag{17}$$

Solutions of Eq. (16) can thus be found as solutions to the 'minimax' problem,

$$\min_{\{\theta,\psi\}} \max_{\beta\geq 0} \left\{ L(\theta,\psi,\beta) = D + \beta(R-\bar{R}) \right\}. \tag{18}$$

The Lagrangian has a form similar to that of the negative ELBO, with an additional β factor multiplying the rate; this framework was presented as an extension of the classical VAE, with the aim of obtaining disentangled latent representations, in Refs. [19,27].

Before addressing the optimization problem, we note that the two terms contributing to the ELBO are related to the mutual information of stimuli and neural responses,

$$I_p(\mathbf{r}, x) = \left\langle \log \frac{p_\theta(\mathbf{r}, x)}{p(x) p_\theta(\mathbf{r})} \right\rangle_{p_\theta(\mathbf{r},x)}, \tag{19}$$

through the bounds

$$H(p) - D \leq I_p(\mathbf{r}, x) \leq R, \tag{20}$$

where $H(p)$ is the entropy of the stimulus distribution [19]. The two inequalities arise because in the variational approximation the posterior over stimuli, $q_\psi(x|\mathbf{r})$, replaces $p_\theta(x|\mathbf{r})$, and the prior over activity patterns, $q_\psi(\mathbf{r})$, replaces $p_\theta(\mathbf{r})$, respectively. Since we are considering continuous stimuli, H is a differential entropy, and is thus defined up to a constant, and D can take negative values. Below, we will illustrate the properties of the generative model also through the D_{KL} divergence, Eq. (7), which is non-negative.

Equation (20) has two important consequences. First, it allows us to interpret the problem in Eq. (16) as an efficient coding problem, where the objective is to maximize a lower bound to the mutual information, $H - D$, subject to a bound on the neural resources, $\bar{R}$. Contrary to the classical efficient coding literature, in which a metabolic constraint is imposed by hand, here it results from the original formulation of the problem as optimization of the ELBO, and it is affected by the assumptions made on the generative model (more specifically, on the prior distribution).

Second, it prescribes a bound on the solutions of Eq. (16). If the variational distributions, $q_\psi(\mathbf{r})$ and $q_\psi(x|\mathbf{r})$, are flexible enough to approximate $p_\theta(\mathbf{r})$ and $p_\theta(x|\mathbf{r})$,

we can achieve both inequalities, and we have $D = H - R$. Along this line in the rate-distortion plane, the negative ELBO achieves its minimum value, equal to the stimulus entropy.

We address the minimax problem of Eq. (18) numerically through a two-timescale alternated stochastic gradient descent-ascent, Alg. 1. We denote by $\{\theta^*, \psi^*, \beta^*\}$ the optimal parameters. If the Lagrangian function is convex in the parameters $\{\theta, \psi\}$, then the algorithm converges to a saddle point [28], i.e., we have

$$L(\theta^*, \psi^*, \beta) \leq L(\theta^*, \psi^*, \beta^*) \leq L(\theta, \psi, \beta^*), \tag{21}$$

for all feasible parameters and $\beta \geq 0$. According to the saddle point theorem (see, e.g., [29]), Eq. (21) implies that $\{\theta^*, \psi^*\}$ is a solution of the problem defined in Eq. (16). The convergence properties in the general case with L possibly non convex in the parameters, but concave in β, are the object of ongoing research; Ref. [30] shows that a gradient descent-ascent algorithm converges to a stationary point of the function $g(\cdot) = \max_{\beta \geq 0} L(\cdot, \beta)$.

Solutions of Eq. (18) obey $\beta^*(R(\theta^*, \psi^*) - \bar{R}) = 0$, i.e., if $\beta^* > 0$, the constraint on the rate is satisfied as an equality, $R = \bar{R}$ (this mechanisms is also known as the complementarity slackness in the Karush–Kuhn–Tucker conditions [31]). Moreover, if the solution is a differentiable point and a saddle point (or, more generally, a stationary point) of the Lagrangian, we have that $\frac{dD}{d\bar{R}}\big|_{\theta^*,\psi^*} = -\beta^*$. (This can be shown by noting that

$$\left.\frac{dL}{d\bar{R}}\right|_{\theta^*,\psi^*} = \left.\frac{dD}{d\bar{R}}\right|_{\theta^*,\psi^*}, \tag{22}$$

and that

$$\begin{aligned}\left.\frac{dL}{d\bar{R}}\right|_{\theta^*,\psi^*,\beta^*} &= \left.\frac{\partial\theta}{\partial\bar{R}}\,\frac{\partial L}{\partial\theta}\right|_{\theta^*,\psi^*,\beta^*} + \left.\frac{\partial\psi}{\partial\bar{R}}\,\frac{\partial L}{\partial\psi}\right|_{\theta^*,\psi^*,\beta^*} + \left.\frac{\partial\beta}{\partial\bar{R}}\,\frac{\partial L}{\partial\beta}\right|_{\theta^*,\psi^*,\beta^*} + \left.\frac{\partial L}{\partial\bar{R}}\right|_{\theta^*,\psi^*,\beta^*} \\ &= -\beta^*,\end{aligned} \tag{23}$$

since the partial derivatives evaluated at the stationary points vanish.)

If the stationarity condition is satisfied and we find $\beta^* = 1$ as a result of our optimization scheme, then it is possible to show, under some assumptions, that the parameters $\{\theta^*, \psi^*\}$ maximize the ELBO. This is obviously true if the solution belongs to the line $D = H - R$, where the ELBO achieves its upper bound. In general, $\beta^* = 1$ implies that the ELBO is optimized if the distortion-rate curve, $D(\bar{R})$ (i.e., the curve defined by the solutions of Eq. (16) as a function of $\bar{R}$), is convex. This observation can be proved with a simple geometric argument. We denote by $\bar{R}_1$ the point at which we have $dD/d\bar{R} = -1$; at this point, the tangent line to the distortion-rate curve is defined by $D = -\text{ELBO}_1 - \bar{R}$, with $-\text{ELBO}_1 = D(\bar{R}_1) + \bar{R}_1$, as the constraint is satisfied as an equality, $R(\bar{R}_1) = \bar{R}_1$. The convexity of the distortion-rate function implies that it lies above this tangent line. Indeed, the convexity property implies that

$$D(\lambda\bar{R}_1 + (1-\lambda)\bar{R}_2) \leq \lambda D(\bar{R}_1) + (1-\lambda)D(\bar{R}_2), \tag{24}$$

with $0 < \lambda < 1$. We now assume, without loss of generality, $\bar{R}_1 < \bar{R}_2$. By subtracting $D(\bar{R}_1)$ and dividing both sides in Eq. (24) by $(1-\lambda)(\bar{R}_2 - \bar{R}_1) < 0$, we obtain

$$\frac{D\left(\lambda\bar{R}_1 + (1-\lambda)\bar{R}_2\right) - D(\bar{R}_1)}{\left(\lambda\bar{R}_1 + (1-\lambda)\bar{R}_2\right) - \bar{R}_1} \geq \frac{(1-\lambda)\left(D(\bar{R}_2) - D(\bar{R}_1)\right)}{(1-\lambda)(\bar{R}_2 - \bar{R}_1)}; \tag{25}$$

we now take the limit $\lambda \to 1$, which yields

$$\left.\frac{dD}{d\bar{R}}\right|_{\bar{R}_1} = -\beta^*(\bar{R}_1) \geq \frac{D(\bar{R}_2) - D(\bar{R}_1)}{\bar{R}_2 - \bar{R}_1}. \tag{26}$$

Finally, by rearranging the terms, we obtain

$$D(\bar{R}_1) + \bar{R}_1 = -\text{ELBO}_1 \leq D(\bar{R}_2) + \bar{R}_2, \tag{27}$$

where we have used $\beta^*(\bar{R}_1) = 1$. We now define the negative ELBO at $\bar{R}_2$, $-\text{ELBO}_2 = D(\bar{R}_2) + R(\bar{R}_2)$. Equation (27) directly implies that $\text{ELBO}_1 \geq \text{ELBO}_2$ when the constraint is satisfied as an equality, $R(\bar{R}_2) = \bar{R}_2$. Instead, when $R(\bar{R}_2) < \bar{R}_2$, we can consider the problem defined in Eq. (16) with $\bar{R}_3 = R(\bar{R}_2)$. In this case, we have $D(\bar{R}_2) = D(\bar{R}_3)$ and, since $D(\bar{R}_2)$ is achieved when $R = \bar{R}_3$, the constraint is satisfied as an equality, $R(\bar{R}_3) = \bar{R}_3$; thus, Eq. (27) implies $\text{ELBO}_1 \geq \text{ELBO}_2 = \text{ELBO}_3$. This proves that $\bar{R}_1$ maximizes the ELBO.

Algorithm 1 Two-timescale optimization algorithm.

Inputs: target rate $\bar{R}$, dataset $\mathcal{D}$
Initialize: $\beta = 1$, encoder/decoder parameters= $\{\theta_i, \psi_i\}$
while convergence **do**
    Define β-ELBO: $L_\beta = D + \beta R$
    for `batch in` $\mathcal{D}$ **do**
        Update parameters: $(\theta, \psi) \leftarrow \text{Adam}\left(\nabla_\theta L_\beta(\texttt{batch}), \nabla_\psi L_\beta(\texttt{batch})\right)$
    end for
    $\beta \rightarrow \max\{\beta + \eta_\beta(R - \bar{R}), 0\}$
end while
return

Numerical optimization and related computations

Numerical simulations are carried out using PyTorch. We solve the optimization problem in Eq. (18) through stochastic gradient descent on the loss on a dataset with $P = 5000$ samples from $p(x)$, divided in minibatches of size 128, with the Adam optimizer [32] with learning rate equal to 10^{-4} and otherwise standard hyperparameters. The learning rate for β, η_β, is set to 0.1. The training is iterated over multiple passes over the data (epochs) with a maximum of 5000 epochs and it is stopped when the training loss running average remains unchanged (with a tolerance of 10^{-5}) for 100 consecutive epochs. The parameters are initialized as follows. The preferred positions, c_i, are initialized as the centroids obtained by applying a k-means clustering algorithm (with $k = N$) to the set of stimuli in the dataset. Tuning widths are initialized by setting $w_i = |c_i - c_j|$, with c_j the closest preferred position to c_i, and the amplitude is set equal to 1, corresponding to a maximum probability of spiking of 0.5. Random noise of small variance is then applied to the initial value of the parameters. We illustrate results obtained by averaging over different random initializations. An example of the evolution of D, R, and β during training is illustrated in Fig. 2.

We illustrate results for N small enough so that it be possible to compute explicitly the sums over activity patterns appearing in the loss function. This also allows us to explore regimes in which the information is compressed in the activity of a finite population of neurons. In Sec. Supporting information, we discuss the numerical issues encountered when the population size is large, and we mention proposed solutions.

Results

We optimize jointly an encoder, a population of neurons with simple tuning curves which stochastically maps stimuli to neural activity patterns, and a decoder, a neural

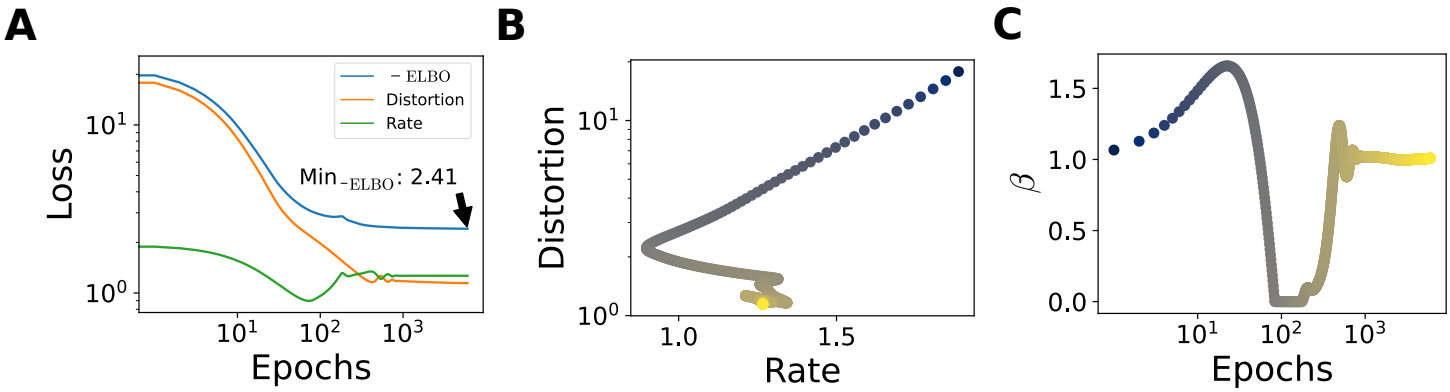

Fig 2. Example of training. $p(x) = \text{Lognormal}(1, 1)$**,** $N = 12$**,** $\bar{R} = 1.32$ (**A**) Evolution of negative ELBO, and the two terms, D and R, with training epochs. Plot in log-log scale. (**B**) Joint evolution of R and D in the rate-distortion plane, colored according to the epoch (increasing from blue to yellow, colors in logarithmic scale). (**C**) Evolution of β during training.

network which maps activity patterns, interpreted as latent variables, to distributions over stimuli. The system is set so as to minimize a bound to the Kullback-Leibler (D_{KL}) divergence between the generative distribution and the true distribution of stimuli (Fig. 1). By formulating the training objective as a constrained optimization problem, we characterize the space of optimal solutions as a function of the value of the constraint; we then discuss the properties of the encoder and of the decoder in the family of solutions.

Degeneracy of optimal solutions

We begin by illustrating two alternative solutions of the ELBO optimization problem, Eq. (10), characterized by different contributions of the two terms, D and R. We first consider the simple, but instructive, case of a Gaussian distribution over stimuli, $p(x) = \mathcal{N}\left(\mu_p, \sigma_p^2\right)$. In order to minimize the rate, a possible solution is to set the parameters of the encoder so as to map all stimuli to the same distribution over neural activity patterns, which takes a similar form as the prior distribution, $p_\theta(\mathbf{r}|x) \approx q_\psi(\mathbf{r})$. This is achieved through neurons with low selectivity, i.e., with broad and overlapping tuning curves (Fig. 3A, top). Despite the non-informative neural representation, a perfect generative model is obtained (in this special, Gaussian case) by mapping all activity patterns to the parameters of the data distribution, $\mu_\psi(\mathbf{r}) = \mu_p$ and $\sigma_\psi^2(\mathbf{r}) = \sigma_p^2$ for all $\mathbf{r}$; in this way, the generative distribution becomes independent from the neural activity, $q_\psi(x|\mathbf{r}) \approx p(x)$ (Fig. 3B, top). The rate term is then negligible and the distortion is equal to the stimulus entropy, thereby satisfying the leftward inequality in Eq. (20). The sum of the two terms is equal to the stimulus entropy, the lower bound of the negative ELBO; the neural representation, however, retains no information about the stimulus.

At the opposite extreme, it is possible to minimize the distortion by learning an injective encoding map that associates distinct stimuli to distinct activity patterns. The decoder can then map each activity pattern to a narrow Gaussian distribution over stimuli. In our framework, this is achieved through narrow and non-overlapping tuning curves that tile the stimulus space (Fig. 3A, bottom). For a given encoding distribution,

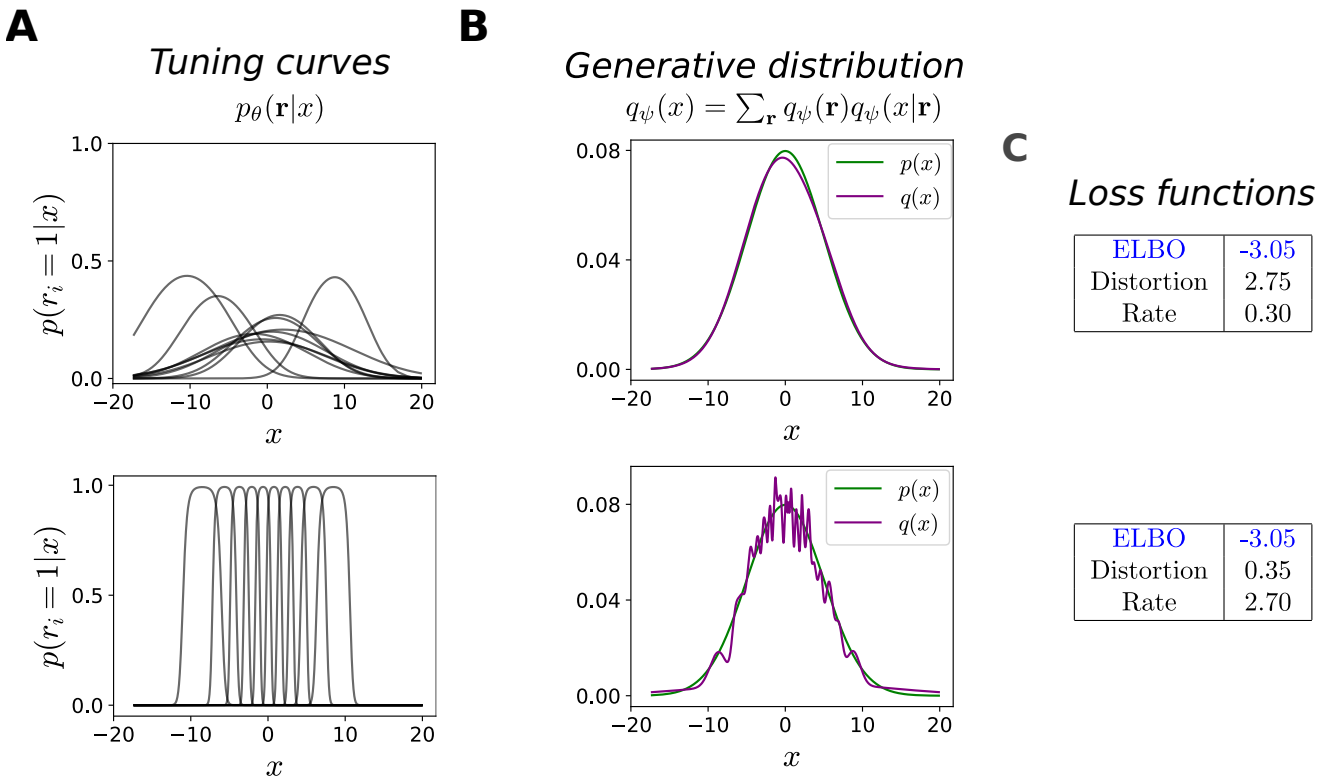

Fig 3. Qualitatively different optimal configurations. In all simulations, $N = 10$ and $p(x) = \mathcal{N}(0, 5)$. Top row: high-distortion, low-rate solution. Bottom row: low-distortion, high-rate solution. (**A**) Bell-shaped tuning curves of the encoder (probability of neuron i to emit a spike, as a function of x). (**B**) Comparison between the stimulus distribution, $p(x)$ (green curve), and the generative distribution, $q(x) = \sum_{\mathbf{r}} q(x|\mathbf{r})q(\mathbf{r})$ (purple curve). (**C**) Numerical values of the ELBO, and the distortion and rate terms.

the optimal prior distribution which minimizes the rate, Eq. (12), is equal to the marginal encoding distribution,

$$q_{\psi^*}(\mathbf{r}) = \langle p_\theta(\mathbf{r}|x) \rangle_{p(x)} \,. \tag{28}$$

(See Ref. [33] for an application of this optimal prior in the context of VAEs.) If the encoding distribution is different for each stimulus, the rate term does not vanish, but, numerically, we find that the parameters of the prior can still be set so as to approximate Eq. (28), achieving the rightward inequality in Eq. (20). As a consequence, the negative ELBO again achieves its lower bound, and it is possible to obtain a generative model that approximates closely the stimulus distribution, though less smoothly (Fig. 3B, bottom).

Thus, although these two solutions yield comparable values of the ELBO (Fig. 1C) and equally accurate generative models, the corresponding neural representations are utterly different. This case is special and contrived, because the conditional generative distribution has the same functional form as the stimulus distribution, and thus a perfect generative model is obtained even when it ignores the latent variables. However, the reasoning extends to more complex cases, and the choice of the forms of the decoding distribution and the prior determines the ability of the system to optimize the ELBO in different ways [19]. In order to achieve an optimal distortion at low rates, the generative distribution must be complex enough to approximate the data distribution

even when the latent variables carry no information about the stimulus. Conversely, prior distributions which can fit marginal encoding distributions in which each data point is mapped precisely to a realization of the latent variables, achieve the optimal values of the rate at low values of the distortion. Indeed, as we show next, we observe the existence of multiple solutions of the ELBO optimization problem also for more complex stimulus distributions.

Analysis of the family of optimal solutions

We explore systematically the space of solutions which optimize the ELBO by minimizing the distortion subject to a constraint on the maximum ('target') value of the rate, $\bar{R}$, a formulation which yields a generalized objective function (Eq. (18)) with a factor β that weighs the rate term (see Methods). The value of $\bar{R}$ is an upper bound to the mutual information between stimulus and neural response; it thereby imposes a degree of 'compression' of the information in the encoding process. We illustrate results for the simple, yet non-trivial, choice of a log-normal stimulus distribution, which exhibits a similar degeneracy as the simple described above (Fig. S1). Similar observations are valid for other distributions as well: in Fig. S2 we illustrate the case of a more complex, multimodal distribution.

Each solution is associated with a point $(\bar{R}, D)$ in the rate-distortion plane. By varying the value of $\bar{R}$, we trace the curve of the optimal distortion as a function of the target rate (Fig. 4A). We focus on the range of values of $\bar{R}$ resulting in $\beta^* = 1$, for which $R = \bar{R}$ and the corresponding solutions also yield an optimal value of the ELBO (shaded grey area). These solutions fall on the line $D = H(p) - R$, with $H(p)$ the stimulus entropy, such that both inequalities in Eq. (20) are achieved; as a result, the mutual information is equal to $\bar{R}$ (Fig. 4A, inset). Deviations from this line appear for extreme values of the target rate. On the one hand, as the stimulus and the generative distributions do not belong to the same parametric family, it is not possible to achieve an optimal distortion with $R = 0$. On the other hand, for sufficiently large $\bar{R}$, the distortion stops decreasing and saturates; this occurs when the tuning curves are as narrow as possible while still tiling the stimulus space (Fig. 3B, bottom). (The distortion can be further decreased by increasing the number of available activity patterns, which depends on the population size (Fig. 5A)).

The quality of the generative model is quantified by the D_{KL} divergence between the generative distribution, $q_\psi(x)$, and the stimulus distribution, $p(x)$; it is negligible for all values of $\bar{R}$ in the region of interest (Fig. 4B) (We recall that the ELBO, up to a constant, is a lower bound to this quantity, and the gap is the D_{KL} divergence between the true and the approximate posterior distribution over neural activity, Eq. (8)). The U-shape is due to the jaggedness of the generative model at high values of $\bar{R}$, which is attenuated as the population sizes increases (Fig. 5B)).

Different values of $\bar{R}$ also result in different encoders, corresponding to different arrangements of the tuning curves (Fig. 4C). For small values of $\bar{R}$, tuning curves are broad and the spacing between preferred positions is small, causing large overlaps: different stimuli are mapped to similar distributions over neural activity patterns. Moreover, they are characterized by low amplitudes and, thus, higher stochasticity; indeed, stochastic neurons yield compressed representations [34]. Increasing $\bar{R}$ causes noise to be suppressed through an increase in the amplitude, and narrower and more distributed tuning curves.

The solutions also differ in the structure of the prior over neural activity, $q_\psi(\mathbf{r})$ (Fig. 4D, insets). In the regime in which the decoder ignores the latent variables, i.e., $q_\psi(x|\mathbf{r}) \approx q_\psi(x)$, the prior, $q_\psi(\mathbf{r})$, is unstructured, and the couplings, J, are weak. By contrast, when $\bar{R}$ is large, the structure of the stimulus distribution affects the coupling matrix in the prior, inducing coupling strengths that depend on the distances between

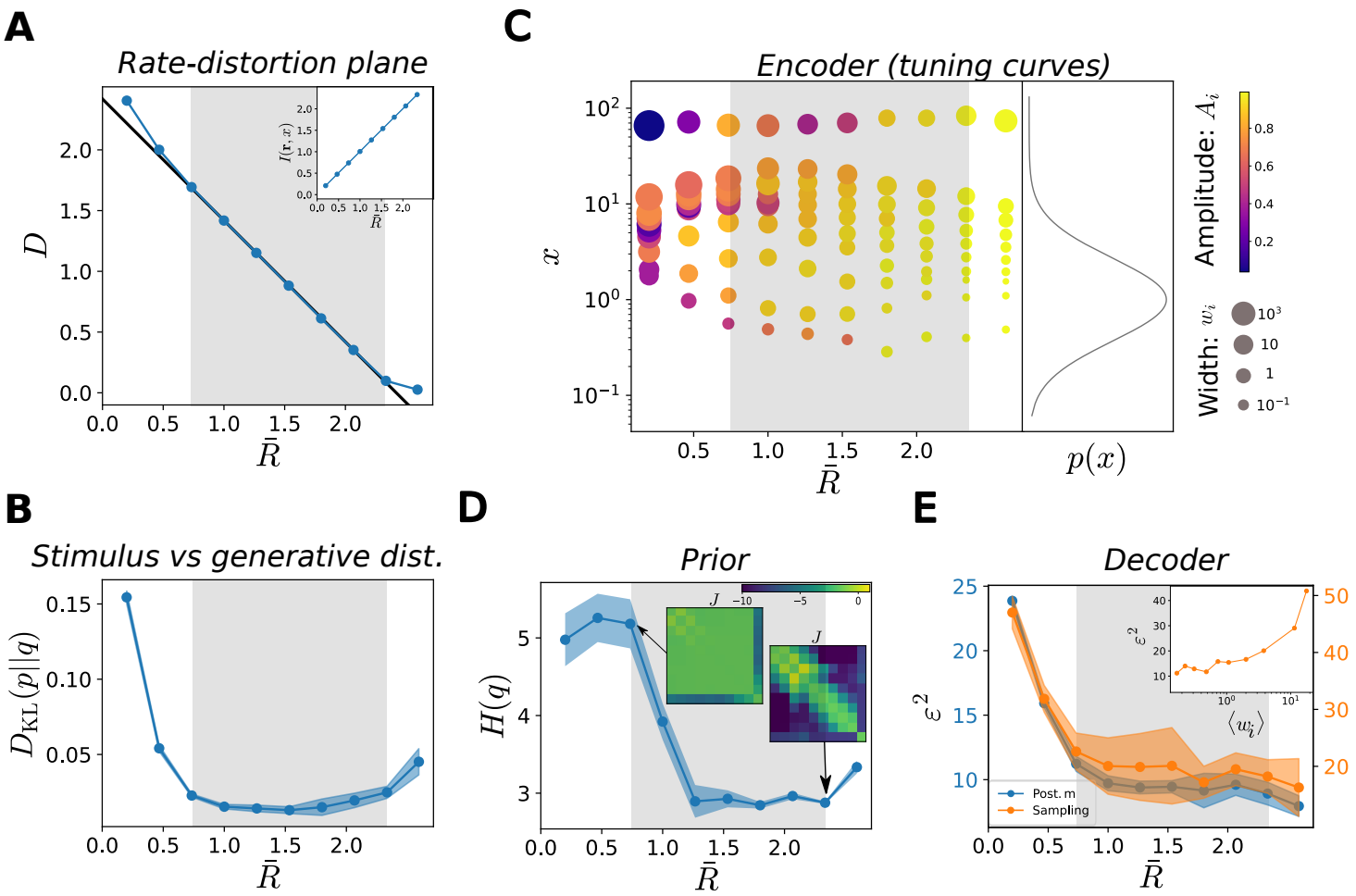

Fig 4. Characterization of the optimal solutions as functions of the target rate. In all simulations, $N = 12$, $p(x) = \text{Lognormal}(1, 1)$, and results are averaged over 16 initializations of the parameters. (**A**) Solutions of the ELBO optimization problem as a function of target rate, $D(\bar{R})$ (blue curve), and theoretical optimum, $D = H(p) - \bar{R}$ (black curve), in the rate-distortion plane. Values of $\bar{R}$ where the solutions coincide with the theoretical optimum (grey region). Solutions depart from the optimal line when the rate is very low (poor generative model) or very high (saturated distortion). Inset: mutual information between stimuli and neural responses as a function of $\bar{R}$. (**B**) Kullback-Leibler divergence between the stimulus and the generative distributions, as a function of $\bar{R}$. (**C**) Optimal tuning curves for different values of $\bar{R}$. Each dot represents a neuron: the position on the y-axis corresponds to its preferred stimulus, the size of the dot is proportional to the tuning width, and the color refers to the amplitude (see legend). The curve on the right illustrates the data distribution, $p(x)$. (**D**) Entropy of the prior distribution over neural activity, $q_\psi(\mathbf{r})$, as a function of $\bar{R}$. Insets show two configurations of the coupling matrices, with rows ordered according to the neurons' preferred stimuli, and coupling strengths colored according to the legend. (**E**) MSE in the stimulus estimate, obtained as the mean of the posterior (blue curve, scale on the left y-axis), or from samples (orange curve, scale on the right y-axis), as a function of $\bar{R}$. Inset: MSE (sampling) as a function of the average tuning width.

the neurons' preferred positions. As the coupling strengths increase, the entropy of the prior distribution decreases (Fig. 4D). We note, however, that in more complex distributions for which, even when the rate is low, a structure is imposed to the prior through the biases, $\mathbf{h}$, the entropy can exhibit a non-monotonic behavior (Fig. S2E).

Finally, we characterize the decoding properties, in terms of a quantity commonly used in perceptual experiments and theoretical analyses: the mean squared error (MSE)

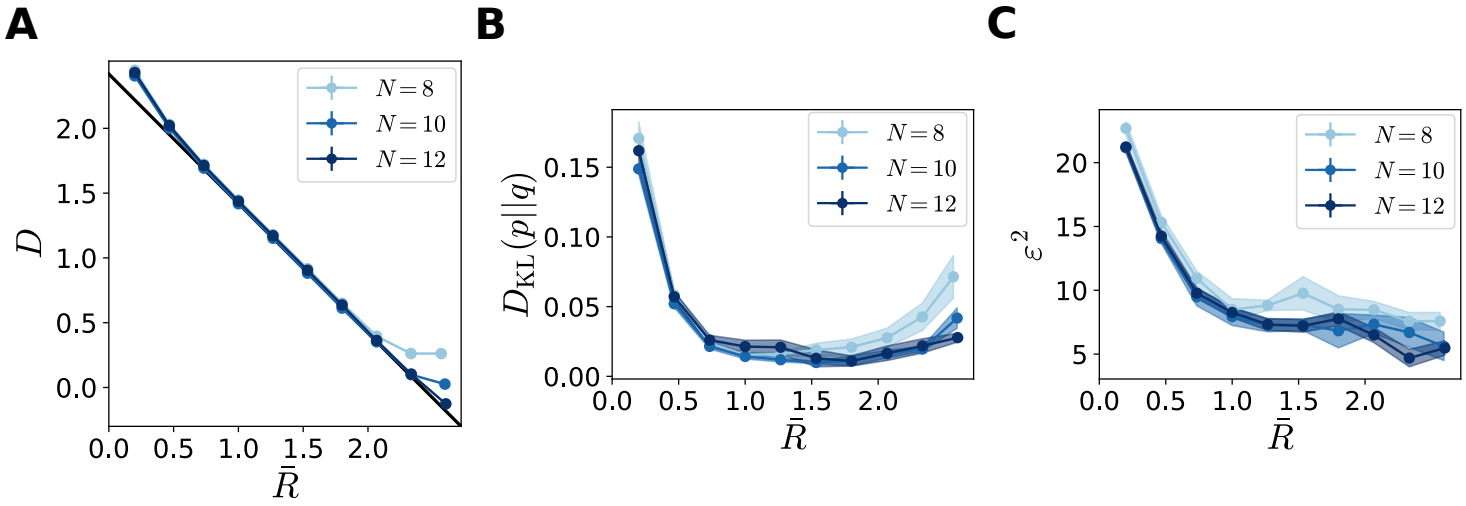

Fig 5. Dependence of the results on the population size. Same simulations as in Fig. 4, with different values of the population size. (**A**) Optimal solutions (blue curves), $D(\bar{R})$, for different population sizes, N, and theoretical bound (black curve), $D = H(p) - \bar{R}$, in the rate-distortion plane. (**B**) Kullback-Leibler divergence between the stimulus and the generative distributions, as a function of $\bar{R}$. (**C**) MSE in the stimulus estimate, obtained as the mean of the posterior, as a function of $\bar{R}$.

in the stimulus estimate. The estimate which minimizes the MSE is the mean of the decoding distribution, $q(x|\mathbf{r})$, $\hat{x} = \mu_\phi(\mathbf{r})$. We also consider the MSE when the stimulus estimate is sampled from the posterior distribution, $\hat{x} \sim q(x|\mathbf{r})$. In Methods we compute the two corresponding functional forms, (see Eqs. (14)-(15)), which differ by a term equal to the posterior variance.

As expected from the higher mutual information between stimuli and neural responses, the decoding performance of the system increases as a function of $\bar{R}$, with a similar qualitative behavior of the error in the two cases (Fig. 4E). But it is worth examining the behavior quantitatively. In both cases, the MSE does not decrease linearly with $\bar{R}$, but rather it exhibits a rapid decrease followed by a slower one; the quantitative value at hgih rates depends on the population size (Fig. 5C). In particular, the system achieve comparable decoding performances for a broad range of values of the tuning width (Fig. 4E, inset). Jointly, the results of Figs. 4B and E suggest that intermediate representations, yielding a smooth approximation of the stimulus distribution, yet achieving a low coding error, are preferred to representations with extremely narrow tuning curves.

Optimal allocation of neural resources and coding performance

The classical efficient coding hypothesis prescribes an allocation of neural resources as a function of the stimulus distribution: more frequently stimuli are represented with higher precision. This has been proposed as an explanation of a number of measurements of perceptual accuracy and behavioral bias [6,35,36]. We investigate, in our model, the relations between stimulus distribution, the use of neural resources (tuning curves), and coding performance, and how each these with $\bar{R}$. We emphasize that the functional form of the stimulus distribution affects these relations, through its interplay with the functional form not only of the encoder (as in the classical efficient coding framework), but also of the generative distribution. In order to make statements

about the typical behavior of the system, we average our results over different random initializations of the parameters; single solutions might deviate from the average behavior due to the small number of neurons and the high dimensionality of the parameters space. We illustrate results for the non-trivial log-normal distributions over stimuli; in Fig. S3 we report results obtained in the Gaussian case. Our conclusions can be compared with results from previous studies. In particular, we invoke the analytical results derived in Ref. [6] for a similar population coding model; in Sec. S2 Appendix. Optimally heterogeneous allocation of neural resources, we provide an alternative derivation of these results and we comment on the main differences with our model. Here, we note that our results are obtained by considering a regime of strong compression of the information (small population sizes), while previous studies focused on the asymptotic regime with $N \to \infty$.

As illustrated in Fig. 4B, the target rate affects the neural density, i.e., the number of neurons with preferred stimuli within a given stimulus window. In previous work, maximizing the mutual information required that the neural density be proportional to the stimulus density, $d(x) \propto p(x)$ [6, 20, 37]. In our case, the range of possible behaviors is richer, especially when the stimulus distribution is non-trivial (i.e., it does not have the same functional form as that of the generative distribution). At low rates, the location of maximum density might be different from the mode of the stimulus distribution, depending on the interplay between the generative and the stimulus distributions (Figs. 6A, S3A). The neural density becomes more similar to the stimulus distribution for large values of $\bar{R}$: a power law functional form, $d(x) = A_d p(x)^{\gamma_d}$, yields a good agreement with our numerical results, with an exponent, γ_d, close to 1/2 (Fig. 6A).

In Ref. [6, 38], analytical results were obtained by constraining the neural density and the tuning width relative to each other. This is equivalent to fixing the overlap between tuning curves, by imposing $w(x) \propto d^{-1}(x) \propto p(x)^{-1}$ (see Sec. S2 Appendix. Optimally heterogeneous allocation of neural resources). In our case, the tuning width and neural density vary independently of each other, and the distribution of widths exhibits an intricate behavior at small values of $\bar{R}$ (Figs. 4B,S3B). However, at large values of $\bar{R}$, the tuning width decreases for large values of the stimulus distribution, and its behavior is well described by a power law, $w_i = A_w / p(c_i)^{\gamma_w}$. As a result, as $\bar{R}$ increases, the inverse correlation between the neural density and the tuning width becomes sharper (Figs. 4C,S3C).

A consequence of the heterogeneous allocation of neural resources is a non-uniform coding performance across stimuli. Figure 7A shows that the MSE exhibits an inverse relation as a function of the stimulus distribution, with more frequent stimuli encoded more precisely. This is broadly consistent with previous studies [6, 39], which maximized the mutual information to obtain the expression

$$\varepsilon^2(x) \propto \frac{1}{p^2(x)}. \tag{29}$$

More precisely, this expression was derived using the Fisher information, whose inverse is a lower bound to the variance of any unbiased estimator and which can be related to the mutual information in some limits. Here, for all values of $\bar{R}$, the error is well described by a power law, $\varepsilon^2(x) = A_e / p(x)^{\gamma_e}$, although the exponent changes as a function of $\bar{R}$ (Figs. 7 A,S3D). Finally, we illustrate how the configuration of the tuning curves affects the coding performance, by plotting the MSE as a function of the neural density and tuning width. We observe a correlation between high coding performance and regions of high neural density as well as with narrow tuning widths (Figs. 7B,C, S3E,F).

To summarize, given our choice of the loss function, which constrains the encoding stage as a function of the decoding stage, we obtain a range of possible optimal neural

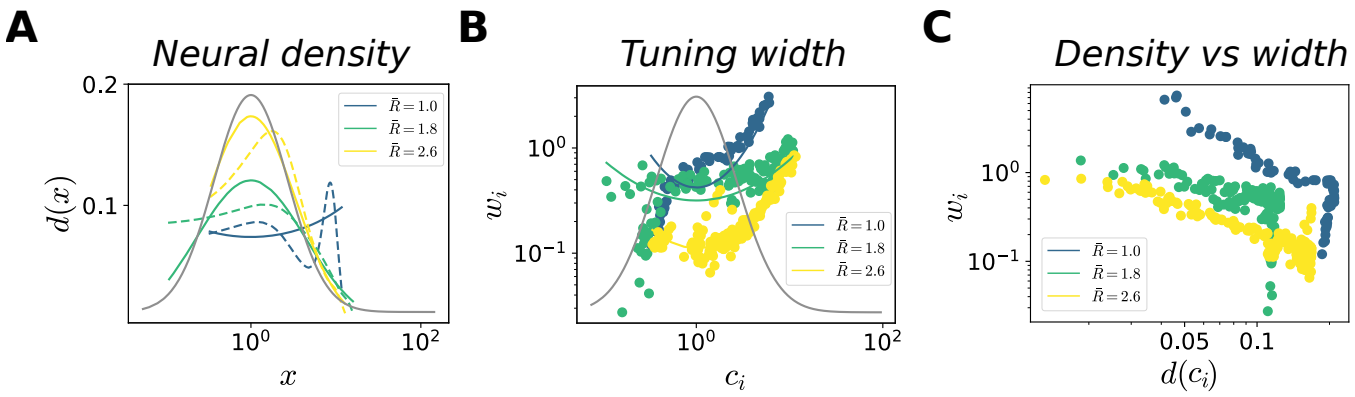

Fig 6. Optimal allocation of neural resources. In all simulations, $N = 12$ and results are averaged over 16 initializations of the parameters, $p(x) = \text{Lognormal}(1, 1)$. Results are illustrated for regions of the stimulus space where the coding performance is sufficiently high, defined as the region where the MSE is lower than the variance of the stimulus distribution. Below, we mention exponents of the power law fit when the variance explained is larger then a threshold, $R^2 \geq 0.7$. (**A**) Neural density as a function x (dashed curves) and power-law fits (solid curves, $R^2 = (0.21, 0.83, 0.95)$, $\gamma_d = (-, 0.43, 0.62)$), for three values of $\bar{R}$ (low, intermediate, and high); the grey curve illustrates the stimulus distribution. The density is computed by applying kernel density estimation to the set of the preferred positions of the neurons. (**B**) Tuning width, w_i, as a function of preferred stimuli, c_i (dots), and power-law fits (solid curves, $R^2 = (0.41, 0.18, 0.92)$, $\gamma_w = (-, -, 0.71)$) for three values of $\bar{R}$; the grey curve illustrates the stimulus distribution. (**C**) Tuning width, w_i, as a function of the neural density, $d(c_i)$, for three values of $\bar{R}$; Pearson correlation coefficient $\rho = (-0.78, -0.78, -0.90)$.

representations. In weakly constrained systems (large values of $\bar{R}$), we qualitatively recover previously derived relationships between tuning curves, stimulus distribution, and coding performance. (The difference in the numerical values of the exponents in the power laws can be explained by the differences between the two models; see Sec. S2 Appendix. Optimally heterogeneous allocation of neural resources. We note that, in Ref. [6] the numerical value of the exponents also change as a function of the form of the loss function.) In systems with severe information compression (small values of $\bar{R}$), the optimal resource allocation exhibits a more intricate behavior, that depends on the interaction between the stimulus distribution and the properties of the generative model.

Case study: neural encoding of acoustic frequencies

Finally, we validate our theory on existing data by considering the empirical distribution of acoustic frequencies in the environment. This distribution was obtained in Ref. [38] by fitting the power spectrum of recordings data, $S(f)$, with a power-law,

$$S(f) = \frac{A}{f_0^p + f^p}, \tag{30}$$

with $A = 2.4 \times 10^6$, $f_0 = 1.52 \times 10^3$, $p = 2.61$ (Fig. 8A, inset).

Despite the heavy tail in the stimulus distribution, we observe a broad range of values of $\bar{R}$ characterized by comparable values of the ELBO (Fig. 8A), and, thus, characterized by comparable generative model performances ($D_{\text{KL}}(p||q) \simeq 0.6$). Also in this case the solutions are characterized by an encoder with increasingly narrow tuning curves, and preferred stimuli more distributed according to the stimulus probability, for increasing $\bar{R}$ (Fig. 8B).

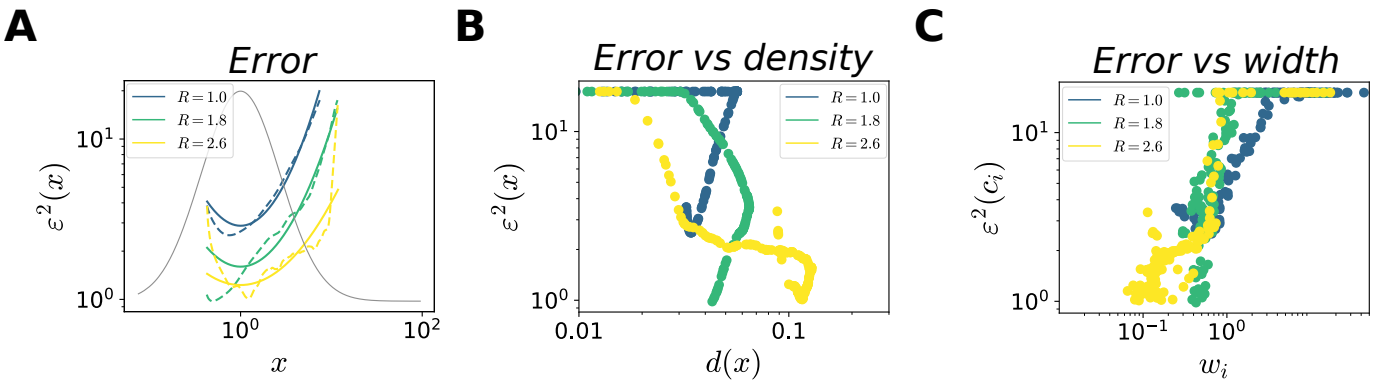

Fig 7. Optimal allocation of coding performance. Same numerical simulations as in Fig. 6. (**A**) MSE (estimate obtained through sampling) as a function of x (dashed curves), and power-law fits (solid curves, $R^2 = (0.92, 0.94, 0.81)$, $\gamma_e = (0.85, 0.70, 0.55)$), for three values of $\bar{R}$. (**B**),(**C**) MSE as a function of the neural density (B) and tuning width (C), for three values of $\bar{R}$; Pearson correlation coefficient $\rho_{density} = (0.05, -0.91, -0.87)$, $\rho_{width} = (0.61, 0.53, 0.72)$

Finally, we test the prediction of our model regarding the dependence of the error on the stimulus value by comparing it to experimental data. We borrow experimental measurements of the so-called frequency-difference limens, the minimum detectable changes in the frequency of a sinusoidal sound wave, from Ref. [40]. Instead of invoking a decision rule, we employ the MSE of the stimulus estimate as a proxy for perceptual resolution (ideally, the two quantities are related through the Fisher information [41]). Since the small number of neurons imposes a fundamental bound to the coding performance, we scale the MSE by a constant factor, a, which can be thought of as a population size gain, to allow for a comparison. The functional form of the MSE captures well the behavior of the frequency-difference limens for a broad range of values of $\bar{R}$ (Fig. 8C). These results show that, despite a large variability in the parameters of the encoder, as is commonly observed in biological systems, robust predictions in the perceptual domain can be derived and are consistent with experimental data.

Discussion

We studied neural representations that emerge in a framework in which populations of neurons encode information about a continuous stimulus with simple tuning curves, but with the additional assumption that the task of the decoder is to maintain a generative model of the stimulus distribution. The consequence of this specific task imposed on the *decoder* is that the *encoder* is set so as to maximize a bound to the mutual information between stimulus and neural activity, as postulated by the efficient coding hypothesis, subject to a constraint on the relative entropy between evoked and the prior distributions over neural activity. Under this constraint, different optimal solutions are obtained, corresponding to equally accurate generative models but (qualitatively) different neural representations of the stimulus (Fig. 4). These representations differ in the degree of compression of information in the neural responses, reflected in encoding (neural) properties (Figs. 4 and 6), in the generative model prior over neural activity (Fig. 4D), and in the coding performance (Figs. 4 and 7).

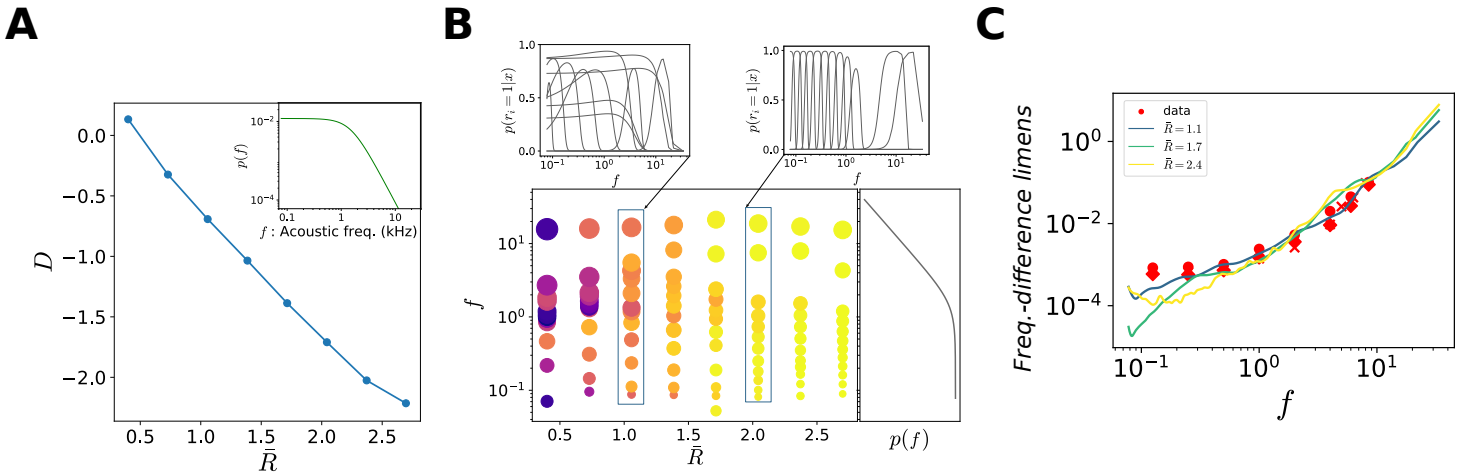

Fig 8. Generative model for distribution of acoustic frequencies. In all simulations, $N = 12$ and results are averaged over 16 initializations of the parameters. (**A**) Solutions of the ELBO optimization problem as a function of target rate, $D(\bar{R})$ (blue curve), and theoretical optimum, $D = H(p) - R$ (black curve), in the rate-distortion plane. Inset: environmental distribution of acoustic frequencies, $p(f)$. (**B**) Optimal tuning curves for different values of $\bar{R}$. Each dot represents a neuron: the position on the y-axis corresponds to its preferred stimulus, the size of the dot is proportional to the tuning width, and the color refers to the amplitude (see legend in Fig. 4). The curve on the right illustrates the data distribution, $p(f)$. Insets show two examples. (**C**) Frequency discrimination as a function of acoustic frequency. Red markers are data points from three different subjects, data from Ref. [40]. Solid curves are the MSE (stimulus estimate obtained through sampling) for three values of $\bar{R}$, scaled by a factor of $a = (227, 121, 111)$.

What causes the degeneracy of the optimal solutions?

Degeneracy in the space of solutions results from the flexibility of the generative model. Indeed, the marginal distribution, $q_\psi(x) = \sum_{\mathbf{r}} q_\psi(x|\mathbf{r}) q_\psi(\mathbf{r})$, is a Gaussian mixture, which is a universal approximator of densities (i.e., a well-chosen Gaussian mixture can be used to approximate any smooth density function [42, 43]). With a population of N neurons we have, in principle, a mixture of 2^N different Gaussians. The prior distribution, defined by $\sim N^2$ parameters, as well as the functional form of the tuning curves, constrain the configurations of latent variables which are exploited. However, despite the undersampling in the space of neural activity patterns, the decoder still retains sufficient flexibility to minimize the ELBO in multiple ways, and the more so the simpler the stimulus distribution (e.g., compare a unimodal distirbution, Fig. 4, to a multimodal one Fig. S1). Here, we focused on relatively simple, one-dimensional stimulus distributions. As the statistics of many natural features are dominated by low-frequency components (e.g., spatial frequencies in natural images), and if powerful decoders are to represent deep brain areas [12, 44], we expect degeneracy in the space of solutions even in the case of multi-dimensional stimuli. The corresponding encoding schemes are interesting objects of future studies.

Internal models and perception as inference

Our choice on the form of the decoder stems from the broad assumption that organisms interact with their environment with the use of internal models. These allow them to perform inference and make predictions. But what form do internal models take and what is their neural substrate? In previous studies [15, 16, 18, 45], internal models were defined by conditioning the probability of a stimulus, x, on the realization of a latent

variable, z, through their joint distribution, $q(z, x) = q(x|z)q(z)$. How the latent variable was related to a specific neural representation was not prescribed. Sensory areas were then assumed to compute a posterior distribution over the latent variable, $q(z|x)$. Only after this step is the neural activity invoked as a way to represent this posterior distribution, either approximately from samples [16, 45], or through the use of a specified parametric form and with the assumption that the values of the parameters are encoded in neural activity [13, 14].

Instead, we define the generative model directly as a joint distribution of two random variables, $q(\mathbf{r}, x)$; $\mathbf{r}$ is the neural activity, while x is defined on the space of stimuli. The neural activity itself plays the role of a latent representation of the stimulus, but it is not set, a priori, to some interpretable feature, such as the presence or the intensity of a Gabor filter in models involving natural images (as in Refs. [16, 45]). In order to constrain sensory areas, we assume the generative model to be implemented in downstream areas and we model its output with a flexible function, a neural network, which outputs a point estimate and an uncertainty about the value of the stimulus [24, 25]. This output corresponds to a perceptual representation of the stimulus in the brain, and can be related to behavioral measurements (as in Fig. 8C)).

Mathematically, the encoding distribution, $p_\theta(\mathbf{r}|x)$, is obtained as a variational approximation of the posterior distribution of the generative model, $q_\psi(\mathbf{r}|x)$, as in previous work. This distribution, however, is defined on the space of neural activity patterns, and not on a set of abstract features. This choice has the drawback of the absence of a simple semantic interpretation of the latent features, but presents the advantage of a natural connection with an encoder based on properties of a neural system, e.g., a set of tuning curves and a model of neural noise. In the case of flexible generative models, different statistics of the latent variables turn out to be optimal. In this sense, the choice of the encoder, as well as the prior of the generative model, is useful to impose a structure on the characteristics of the neural representations.

Optimal tuning width

Our choice of encoding model allows us to compare our results with those of earlier studies that considered the optimal arrangement of neurons with bell-shaped tuning curves in the presence of non-uniform stimulus distributions [6, 20]. While for higher values of the target rate we recover the previously derived allocation of neural resources as a function of the stimulus distribution, the behavior for lower values of the target rate is more intricate, and depends on the specifics of the stimulus distribution. Thus, in our case, the constraint on neural resources has a stronger impact on their optimal allocation than, for example, in Ref. [6], where the bound on the mean activity in the population acts merely as a scaling factor, and the behavior of the tuning curves is more constrained. In particular, in Ref. [6] the tuning width was fixed a priori to be inversely proportional to the neural density, to enforce a fixed amount of overlap between tuning curves: it was not optimized. This choice was made to avoid a common issue in this type of calculations: in the case of a one-dimensional stimulus and in the asymptotic limit of infinitely many neurons, the maximization of the mutual information yields the pathological solution of infinitely narrow tuning curves [46, 47]. Metabolic constraints on the neural activity do not solve the issue, as narrow tuning curves can exhibit a moderate activity (as long as their amplitude is not too large).

In our framework, instead, the optimal tuning width and the amount of overlap between tuning curves are both optimized and vary as a function of $\bar{R}$. Moreover, a regime with intermediate values of the constraint, in which tuning curves are broad, exhibits both a smooth generative model (low D_{KL} divergence) *and* a low MSE (Fig. 4B,E). Broad tuning curves are beneficial to obtain smooth generative models, while still allowing high for coding performance.

Interpretation of the resource constraint

The constraint in Eq. (16) consists in the divergence between the evoked neural activity and its prior distribution according to the generative model. This formulation is different from usual metabolic constraints which are designed to account for the energetic cost of neural activity [5], and one may ask whether such a constraint, statistical in nature, also comes with a biological interpretation.

To answer this question, we invoke the results of Ref. [18], in which the prior distribution over latent variables of the internal model is related to the spontaneous neural activity. The authors start from the observation that in a well-calibrated internal model the prior equals the mean posterior (Eq. (28)) [10]. By comparing the average evoked activity to the spontaneous activity according to the D_{KL} divergence, the authors show that the two quantities become closer during development, and that this phenomenology is specific to naturalistic stimuli. This finding is then proposed as evidence of an internal model in primary visual cortex optimized for natural images, acquired gradually during development. In this picture, the prior distribution of the generative model is identified with the spontaneous neural activity; we note, however, that there is no a priori reason to expect this relationship.

In our case, the prior distribution is parametrized by the biases and couplings of an Ising model. As we have shown, there are multiple ways to achieve a statistically optimal internal model and to minimize the D_{KL} divergence between the two sides of Eq. (28), which differ in the value of the rate. At low rates, Eq. (28) is approximated by relying on the optimization of the encoder parameters which are set so as to make $p_\theta(\mathbf{r}|x)$ similar to the prior for all stimuli; this then results in an unstructured coupling matrix in the prior distribution (Fig. 4E, top). Conversely, at high rates, the encoder has a well defined structure which achieves a low distortion, and Eq. (28) is approximated by optimizing the parameters of the prior and embedding the structure of the average posterior distribution in the connectivity matrix (Fig. 4E, bottom). The value of the target rate can therefore be thought of as a cost of imposing structure in prior (spontaneous activity), through circuit properties. Thus, our model suggest an alternative normative principle to govern neural couplings as compared to information maximization, as proposed in Ref. [48].

VAEs in neuroscience: related studies

VAEs are among the state of the art approaches to unsupervised learning, and in recent years they have been applied in different contexts in neuroscience to model neural responses. Several studies have considered neuroscience-inspired VAEs, in which the generative model was based on a decomposition of natural images into sparse combinations of linear features [49]. It was then paired with a powerful encoder, which models the sensory encoding process, and specific assumptions on the prior distribution of the latent variables, to obtain representations similar to the ones observed in the early visual pathway (in V1 and V2) [12, 50, 51]. In these models, the simplicity of the generative distribution prevented posterior collapse. We note that, in our case, we reverse this approach, by assuming a specific a simple and biologically motivated form of the encoder (a set of tuning curves), while we allow for a flexible decoder.

In the context of higher visual areas instead, more complex generative models were needed to capture neural representations [44]; to overcome the issue of posterior collapse, the authors used a loss function akin to the one in Eq. (18), but the value of β was chosen by hand. In doing so, they obtained an empirical advantage in the semantic interpretability of the latent variables, at the cost of abandoning the requirement that the loss function be a bound to the log-likelihood. This, so-called, β-VAE approach was also employed in Ref. [52] to study optimal tuning curves in a population coding model

of spiking neurons similar to ours. In this study, however, the population as a whole was constrained to emit one spike only, limiting the number of available activity patterns to N (the number of neurons). Moreover, the encoder and the decoder are not optimized independently; this choice prevented the emergence of multiple alternative neural representations in the $\beta = 1$ case. By varying β, the authors obtained neural representations which differed in the shape of the optimal tuning curves, but, since for $\beta \neq 1$ the ELBO was not optimized, the resulting generative model was not accurate.

Acknowledgments

We thank Trang-Anh Nghiem and Luc Stebens for comments on a earlier version of the manuscript.

Supporting information

S1 Appendix. Numerical approaches in the case of large neural populations

To extend our model to larger populations, there are two numerical issues to consider. The first one concerns the distortion term and the gradient with respect to the parameters of the encoder. In order to obtain a low-variance estimate of the gradient, an approach is to use the so-called reparametrization trick together with a continuous relaxation of the discrete random variable, $\mathbf{r}$, (or Gumbel-softmax trick [53, 54]), and calculate the gradient as

$$\begin{aligned} \nabla_\theta D(x) &= \nabla_\theta \left\langle \log q_\psi(x|\mathbf{r}) \right\rangle_{p_\theta(\mathbf{r}|x)} \\ &\approx \left\langle \nabla_\theta \log q_\psi(x|f_\theta(\xi,x)) \right\rangle_{p(\xi)} , \end{aligned} \tag{31}$$

with $p(\xi) = \mathcal{U}(0,1)$. Here,

$$f_\theta(\xi,x)) = \mathcal{S}\left(\frac{\boldsymbol{\eta}_\theta(x) + \mathcal{S}^{-1}(\xi)}{\tau}\right) \tag{32}$$

depends deterministically on the parameters θ through the natural parameters of the encoder; the hyperparameter τ controls the steepness of the logistic function, and consequently the bias-variance trade-off for the gradient; simulations with values of $\tau = 10^{-2}$ yield results comparable to the ones presented here.

The second issue pertains to the form of the rate. Its expression can be simplified as

$$\begin{aligned} D_{\mathrm{KL}}\left(p_\theta(\mathbf{r}|x)||q_\psi(\mathbf{r})\right) &= \left\langle (\boldsymbol{\eta}(x) - \mathbf{h})\,\mathbf{r} - \mathbf{r}^T J \mathbf{r} \right\rangle_{p_\theta(\mathbf{r}|x)} - \sum_{i=1}^{N} \log\left(1 + e^{\eta_i(x)}\right) + \log Z \\ &= (\boldsymbol{\eta}(x) - \mathbf{h})\,\mathbf{p}(x) - \mathbf{p}^T(x) J \mathbf{p}(x) - \sum_{i=1}^{N} \log\left(1 + e^{\eta_i(x)}\right) + \log Z, \end{aligned} \tag{33}$$

where $\mathbf{p}(x) = \mathcal{S}\left(\boldsymbol{\eta}(x)\right)$ is the vector of mean parameters of the encoding distribution (i.e., the spiking probability of neurons). In the expectation of the quadratic form, $\langle \mathbf{r}^T J \mathbf{r} \rangle_{p_\theta(\mathbf{r}|x)} = \mathrm{tr}(K_{\mathbf{rr}} J) + \mathbf{p}^T(x) J \mathbf{p}(x)$, we have that $\mathrm{tr}(K_{\mathbf{rr}} J) = 0$, as the covariance matrix of the activity patterns, $K_{\mathbf{rr}}$, is proportional to the identity, and the diagonal elements of J vanish. Here, the numerical load is in computing the gradient of the log-partition function, $\log Z$; this can be done by Monte Carlo methods [55].

S2 Appendix. Optimally heterogeneous allocation of neural resources

We provide an alternative derivation, based on scaling arguments, of the results in Ref. [6]. We consider a population of N neurons, in which neuron i responds to a continuous scalar stimulus, x, according to a bell-shaped tuning curve, $f_i(x)$. We consider a discretization of the stimulus space, $x = \{x_i\}_{i=1}^{L}$, and we denote by d_i the number of neurons whose preferred stimulus is x_i and by w_i their tuning width (Fig. S4A). The number of neurons encoding information about stimulus x_i scales as

$$\#\text{neurons} = M_i \sim d_i w_i, \tag{34}$$

as increasing the number of neurons and the tuning width (both of which, we assume, vary sufficiently smoothly with position) each increases the number of neurons that

'monitor' a given of the stimulus. We assume that neural responses, r, are corrupted by noise with standard deviation η. Through a simple geometric argument (Fig. S4B), we estimate the square of the difference between the stimulus estimate based on the activity of neuron j and the true stimulus, i.e., the squared error, as

$$(\hat{x}_i - x_i)^2 \equiv \Delta x_i^2 \approx \left(\frac{\eta}{f_j'(x_i)}\right)^2, \tag{35}$$

where $f_j'(x)$ denotes the slope of the tuning curve j at x_i. The derivative of a bell-shaped tuning curve scales as $f_i'(x) \sim f_i(x)/w_i$; if noise has a Poisson distribution, the variance of the response is equal to the mean, so that Eq. (35) can be written as

$$\Delta x_i^2 \sim \left(\frac{\text{const}}{w_i}\right)^{-2} \sim w_i^2. \tag{36}$$

As M independent neurons encode stimulus x_i, we can average the single estimates from each of the neuron to obtain a more faithful estimate. The variance of this population estimate, i.e., the MSE, for stimulus i, scales as

$$\begin{aligned} \varepsilon_i^2 = \text{Var}\left(\frac{1}{M_i}\sum_{j=1}^{M}(\Delta x_i)_j\right) &= \frac{1}{M^2}\sum_{j=1}^{M_i}(\Delta x_i^2)_j \\ &\approx \frac{w_i^2}{M_i} \\ &\approx \frac{w_i}{d_i}, \end{aligned} \tag{37}$$

where the last equality follows from Eq. (34). By taking the limit of an infinitely fine discretization, $L \to \infty$, and assuming that the population size is large enough so that the quantities d_i and w_i vary smoothly, we can consider a continuum limit with

$$d_i \to d(x), \tag{38}$$

the neural density,

$$w_i \to w(x), \tag{39}$$

the tuning width, and

$$M_i \to M(x) = d(x)w(x) \tag{40}$$

We will require an additional constraint to find optimal solutions. Different forms of constraint can be imposed. The constraint that reproduces the results of [6] ensures a 'uniform coverage' across stimuli, i.e., $M(x) = \text{constant}$, or

$$w(x) \sim \frac{1}{d(x)}. \tag{41}$$

The efficient coding hypothesis posits that neurons are arranged so as to maximize the mutual information between stimuli and neural responses. An approximation of the mutual information in terms of the Fisher information, $J(x)$, in the asymptotic limit, can be obtained as

$$I(r,x) = \int dx p(x) \log\left(J(x)\right) + \text{const}, \tag{42}$$

where $p(x)$ is the distribution of stimuli and const denotes terms that don't depend on the neural responses [20]. The Fisher information is a lower bound to the variance of any unbiased estimator; if we assume that the bound is tight, we have that

$$J(x) \approx \frac{1}{\varepsilon^2(x)} \sim d(x)^2, \tag{43}$$

where $\varepsilon^2(x)$ corresponds to the continuum limit of Eq. (37) and we used the scaling relation of Eq. (41).

We now maximize the mutual information subject to a constraint on the neural resources–here, merely, the number of neurons–by optimizing the sum of the two terms

$$\max_{d(x)} \left\{ \int dx p(x) \log \left(d(x)^2 \right) + \beta \int dx d(x) \right\}. \tag{44}$$

By taking a functional derivative with respect to $d(x)$ and setting it to zero, we obtain

$$d(x) \sim p(x), \tag{45}$$

and, consequently, the scaling of the MSE as

$$\varepsilon^2(x) \sim \frac{1}{p^2(x)}. \tag{46}$$

Main differences with our model. Our model is similar to the one presented above, but it differs from it in ways which complicate analytical calculations and give rise to more complex behaviors.

- The first difference is in the noise model: we assume binary neurons, while the above calculations are carried out with Poisson neurons, an assumption which allows the simplification in Eq. (36).
- The second difference is that, in our formulation, the tuning width and neural density are free to vary independently. As a result, we can achieve a non-uniform coverage across stimuli.
- The third difference is that we assume a finite population size, rather than he asymptotic $N \to \infty$ limit.
- Finally, our loss function is similar to that in Eq. (44) for what concerns the first term, which represents the mutual information between stimuli and neural responses (although in our case we have a lower bound, which depends also on the decoder), but the constraint is more intricate due to its dependence on the generative model.

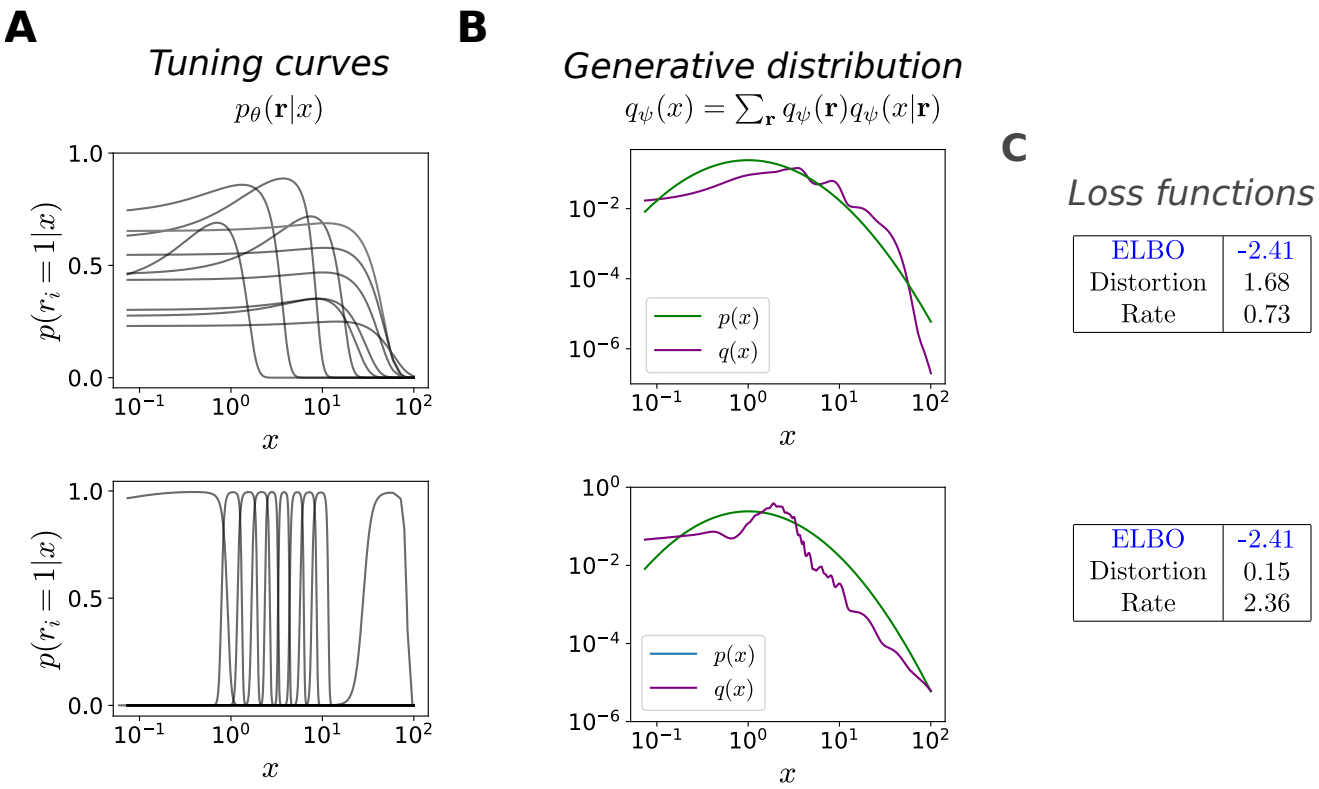

Fig S1. Qualitatively different optimal configurations. Same as Fig. 3, in the case of lognormal distribution over stimuli, $p(x) = \text{Lognormal}(1, 1)$. Top row: high-distortion, low-rate solution. Bottom row: low-distortion, high-rate solution. (**A**) Bell-shaped tuning curves of the encoder (probability of neuron i to emit a spike, as a function of x). (**B**) Comparison between the stimulus distribution, $p(x)$ (green curve), and the generative distribution, $q(x) = \sum_{\mathbf{r}} q(x|\mathbf{r})q(\mathbf{r})$ (purple curve). (**C**) Numerical values of the ELBO, and the distortion and rate terms.

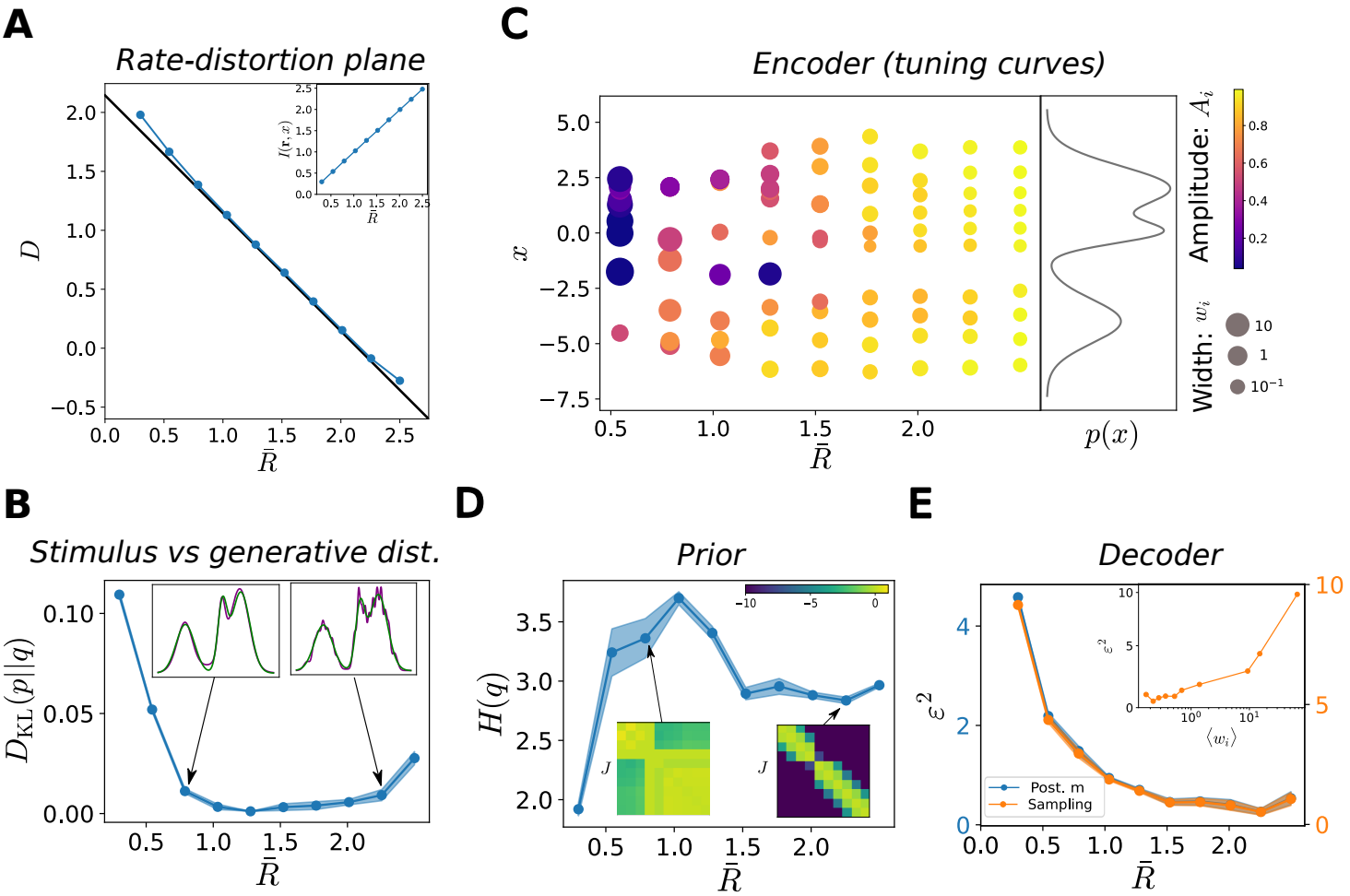

Fig S2. Characterization of the optimal solutions as functions of the target rate. Same as Fig. 4, but with $p(x)$ a multimodal distribution: a mixture of three Gaussians with means $\{-4, 0, 2\}$; variances $\{1, 0.5, 1\}$; and mixture coefficients $\{0.3, 0.2, 0.5\}$. (**A**) Solutions of the ELBO optimization problem as a function of target rate, $D(\bar{R})$ (blue curve), and theoretical optimum, $D = H(p) - R$ (black curve), in the rate-distortion plane. Values of $\bar{R}$ where the solutions coincide with the theoretical optimum (grey region). Solutions depart from the optimal line when the rate is very low (poor generative model) or very high (saturated distortion). Inset: mutual information between stimuli and neural responses as a function of $\bar{R}$. (**B**) D_{KL} divergence between the stimulus and the generative distributions, as a function of $\bar{R}$. Insets: two examples of comparison between stimulus (green curve) and generative distribution (purple curve). (**C**) Optimal tuning curves for different values of $\bar{R}$. Each dot represents a neuron: the position on the y-axis corresponds to its preferred stimulus, the size of the dot is proportional to the tuning width, and the color refers to the amplitude (see legend). The curve on the right illustrates the data distribution, $p(x)$. (**D**) Entropy of the prior distribution over neural activity, $q_\psi(\mathbf{r})$, as a function of $\bar{R}$. Insets show two configurations of the coupling matrices, with rows ordered according to the neurons' preferred stimuli, and coupling strengths colored according to the legend. (**E**) MSE of the stimulus estimate, obtained as the mean of the posterior (blue curve, scale on the left y-axis), or from samples (orange curve, scale on the right y-axis), as a function of $\bar{R}$. Inset: MSE (sampling) as a function of the average tuning width.

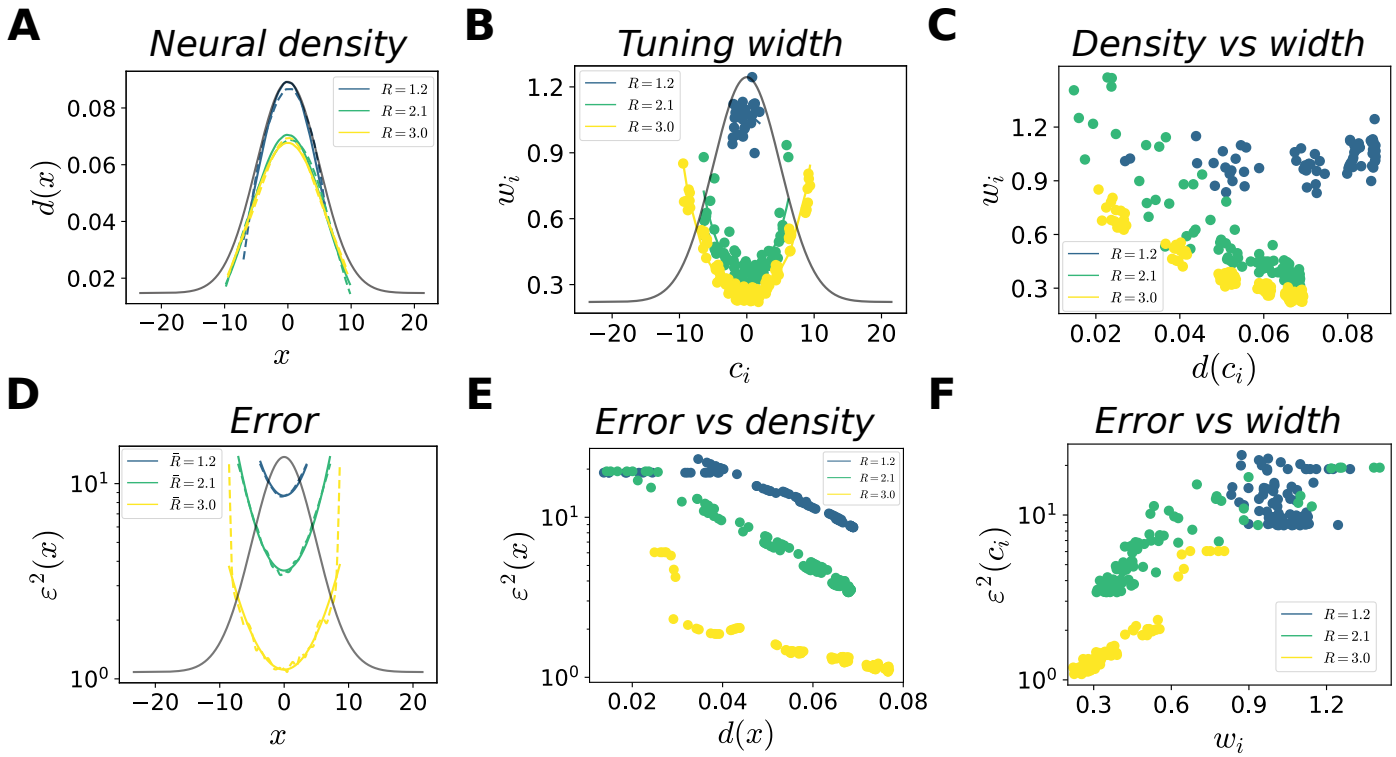

Fig S3. Optimal allocation of neural resources and coding performance. Same as Fig. 6,7, in the case of a Gaussian distribution $p(x) = \mathcal{N}(0, 5)$ (same of Fig. 3). (**A**) Neural density as a function x (dashed curves) and power-law fits (solid curves, $R^2 = (0.96, 0.99, 0.99)$, $\gamma_d = (0.99, 0.71, 0.64)$), for three values of $\bar{R}$ (low, intermediate, and high); the grey curve illustrates the stimulus distribution. The density is computed by applying kernel density estimation to the set of the preferred positions of the neurons. (**B**) Tuning width, w_i, as a function of preferred stimuli, c_i (dots), and power-law fits (solid curves, $R^2 = (0.10, 0.74, 0.92)$, $\gamma_w = (-, 0.87, 0.65)$) for three values of $\bar{R}$; the grey curve illustrates the stimulus distribution. (**C**) Tuning width, w_i, as a function of the neural density, $d(c_i)$, for three values of $\bar{R}$; Pearson correlation coefficient $\rho = (0.30, -0.91, -0.97)$. (**D**) MSE (estimate obtained through sampling) as a function of x (dashed curves), and power-law fits (solid curves, $R^2 = (0.96, 0.99, 0.90)$, $\gamma_e = (1.47, 1.32, 0.83)$), for three values of $\bar{R}$. (**E**),(**F**) MSE as a function of the neural density (E) and tuning width (F), for three values of $\bar{R}$; Pearson correlation coefficient $\rho_{density} = (-0.91, -0.97, -0.8)$, $\rho_{width} = (0.31, 0.93, 0.88)$

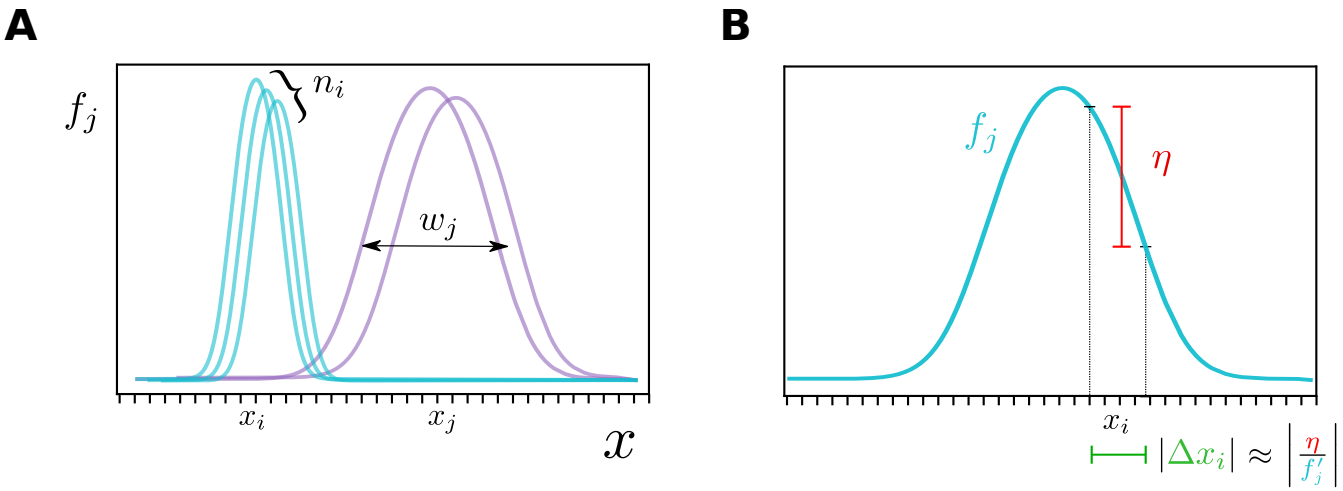

Fig S4. Population coding model with bell-shaped tuning curves. (**A**) A one-dimensional stimulus is encoded through bell-shaped tuning curves. The number of neurons whose preferred positions are a given stimulus, x_i, is denoted by n_i, while w_i denotes the tuning width. (**B**) Approximate scaling of the error in stimulus estimate, Δx_i, when the response of a neuron, with mean f_j, is affected by a noise of standard deviation η.

References

1. Barlow HB. Possible Principles Underlying the Transformations of Sensory Messages. Sensory Communication. 1961;1(1):216–234. doi:10.7551/mitpress/9780262518420.003.0013.

2. Atick JJ, Redlich AN. Towards a Theory of Early Visual Processing. Neural Computation. 1990;2(3):308–320. doi:10.1162/neco.1990.2.3.308.

3. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annual Review of Neuroscience. 2001;24(1):1193–1216. doi:10.1146/annurev.neuro.24.1.1193.

4. Lewicki MS. Efficient coding of natural sounds. Nature Neuroscience. 2002;5(4):356–363. doi:10.1038/nn831.

5. Laughlin SB, de Ruyter van Steveninck RR, Anderson JC. The metabolic cost of neural information. Nature Neuroscience. 1998;1(1):36–41. doi:https://doi.org/10.1038/236.

6. Ganguli D, Simoncelli EP. Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. Neural Computation. 2014;26(10):2103–2134. doi:10.1162/NECO_a_00638.

7. Park IM, Pillow J. Bayesian Efficient Coding. bioRxiv. 2020; p. 178418. doi:https://doi.org/10.1101/178418.

8. Von Helmholtz H. Helmholtz's treatise on physiological optics. Optometry and Vision Science. 1927;4(11). doi:10.1097/00006324-192711000-00011.

9. Dayan P, Hinton GE, Neal RM, Zemel RS. The Helmholtz Machine. Neural Computation. 1995;7(5):889–904. doi:https://doi.org/10.1162/neco.1995.7.5.889.

10. Dayan P, Abbott LF. Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems. MIT Press; 2001.

11. Wainwright MJ, Simoncelli E. Scale mixtures of Gaussians and the statistics of natural images. Advances in Neural Information Processing Systems. 1999;12:855–861.

12. Csikor F, Meszéna B, Szabó B, Orbán G. Top-down inference in an early visual cortex inspired hierarchical Variational Autoencoder. arXiv preprint arXiv:220600436. 2022;.

13. Vertes E, Sahani M. Flexible and accurate inference and learning for deep generative models. Advances in Neural Information Processing Systems. 2018;31:4166–4175.

14. Zemel RS, Dayan P, Pouget A. Probabilistic Interpretation of Population Codes. Neural Computation. 1998;10(2):403–430. doi:https://doi.org/10.1162/089976698300017818.

15. Lee TS, Mumford D. Hierarchical Bayesian inference in the visual cortex. JOSA A. 2003;20(7):1434–1448. doi:https://doi.org/10.1364/JOSAA.20.001434.

16. Hoyer PO, Hyvärinen A. Interpreting Neural Response Variability as Monte Carlo Sampling of the Posterior. Advances in Neural Information Processing Systems. 2003;15:293–300.

17. Kingma DP, Welling M. Auto-Encoding Variational Bayes. International Conference on Learning Representations. 2014;2. doi:https://doi.org/10.48550/arXiv.1312.6114.

18. Berkes P, Orbán G, Lengyel M, Fiser J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science. 2011;331(6013):83–88. doi:10.1126/science.1195870.

19. Alemi AA, Poole B, Fischer I, Dillon JV, Saurous RA, Murphy K. Fixing a Broken ELBO. International Conference on Machine Learning. 2018;35.

20. Brunel N, Nadal JP. Mutual Information, Fisher Information, and Population Coding. Neural Computation. 1998;10(7):1731–1757. doi:10.1162/089976698300017115.

21. Salinas E, Abbott LF. Vector reconstruction from firing rates. Journal of Computational Neuroscience. 1994;1(1):89–107. doi:10.1007/BF00962720.

22. Sharpee TO, Berkowitz JA. Linking neural responses to behavior with information-preserving population vectors. Current Opinion in Behavioral Sciences. 2019;29:37–44. doi:10.1016/j.cobeha.2019.03.004.

23. Schneidman E, Berry MJ, Segev R, Bialek W. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature. 2006;440(7087):1007–1012. doi:10.1038/nature04701.

24. Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nature Neuroscience. 2006;9(11):1432–1438. doi:10.1038/nn1790.

25. Walker EY, Cotton RJ, Ma WJ, Tolias AS. A neural basis of probabilistic computation in visual cortex. Nature Neuroscience. 2020;23(1):122–129. doi:10.1038/s41593-019-0554-5.

26. Sønderby CK, Raiko T, Maaløe L, Sønderby SK, Winther O. Ladder Variational Autoencoders. Advances in Neural Information Processing Systems. 2016;30:3745–3753. doi:https://doi.org/10.48550/arXiv.1602.02282.

27. Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, et al. beta-VAE: Learning basic visual concepts with a constrained variational framework. International Conference on Learning Representations. 2017;5.

28. Nedić A, Ozdaglar A. Subgradient methods for saddle-point problems. Journal of Optimization Theory and Applications. 2009;142(1):205–228. doi:10.1007/s10957-009-9522-7.

29. Boyd S, Boyd SP, Vandenberghe L. Convex optimization. Cambridge university press; 2004.

30. Lin T, Jin C, Jordan MI. On gradient descent ascent for nonconvex-concave minimax problems. International Conference on Machine Learning. 2020;37. doi:https://doi.org/10.48550/arXiv.1906.00331.

31. Arrow KJ, Azawa H, Hurwicz L, Uzawa H, Chenery HB, Johnson SM, et al. Studies in linear and non-linear programming. vol. 2. Stanford University Press; 1958.

32. Kingma DP, Ba JL. Adam: A Method for Stochastic Optimization. International Conference on Learning Representations. 2015;3. doi:https://doi.org/10.48550/arXiv.1412.6980.

33. Tomczak JM, Welling M. VAE with a VampPrior. International Conference on Artificial Intelligence and Statistics. 2017; p. 1214–1223.

34. Achille A, Soatto S. Information Dropout: Learning Optimal Representations Through Noisy Computation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018;40(12):2897–2905. doi:10.1109/TPAMI.2017.2784440.

35. Wei XX, Stocker AA. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. Advances in Neural Information Processing Systems. 2012;25:1304–1312.

36. Wei XX, Stocker AA. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. Nature Neuroscience. 2015;18(10):1509–1517. doi:10.1038/nn.4105.

37. Yerxa TE, Kee E, DeWeese MR, Cooper EA. Efficient sensory coding of multidimensional stimuli. PLoS computational biology. 2020;16(9):e1008146. doi:10.1371/journal.pcbi.1008146.

38. Ganguli D, Simoncelli EP. Neural and perceptual signatures of efficient sensory coding. arXiv preprint arXiv:160300058. 2016;.

39. Wei XX, Stocker AA. Mutual information, fisher information, and efficient coding. Neural Computation. 2016;28(2):305–326. doi:10.1162/NECO_a_00804.

40. Moore BCJ. Frequency difference limens for short-duration tones. The Journal of the Acoustical Society of America. 1973;54(3):610–619.

41. Seriès P, Stocker AA, Simoncelli EP. Is the homunculus "aware" of sensory adaptation? Neural Computation. 2009;21(12):3271–3304.

42. Dalal SR, Hall WJ. Approximating Priors by Mixtures of Natural Conjugate Priors. Journal of the Royal Statistical Society: Series B (Methodological). 1983;45(2). doi:10.1111/j.2517-6161.1983.tb01251.x.

43. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

44. Higgins I, Chang L, Langston V, Hassabis D, Summerfield C, Tsao D, et al. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. Nature Communications. 2021;12(1):1–14. doi:https://doi.org/10.1038/s41467-021-26751-5.

45. Orbán G, Berkes P, Fiser J, Lengyel M. Neural Variability and Sampling-Based Probabilistic Representations in the Visual Cortex. Neuron. 2016;92(2):530–543. doi:10.1016/j.neuron.2016.09.038.

46. Seung HS, Sompolinsky H. Simple models for reading neuronal population codes. PNAS. 1993;90(22):10749–10753. doi:10.1073/pnas.90.22.10749.

47. Zhang K, Sejnowski TJ. Neuronal Tuning: To Sharpen or Broaden? Neural Computation. 1999;11(1):75–84. doi:10.1162/089976699300016809.

48. Tkačik G, Prentice JS, Balasubramanian V, Schneidman E. Optimal population coding by noisy spiking neurons. PNAS. 2010;107(32):14419–14424. doi:10.1073/pnas.1004906107/-/DCSupplemental.

49. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996;381(6583):607–609. doi:10.1038/381607a0.

50. Barello G, Charles AS, Pillow JW. Sparse-Coding Variational Auto-Encoders. bioRxiv. 2018; p. 399246. doi:10.1101/399246.

51. Aitchison L, Hennequin G, Lengyel M. Sampling-based probabilistic inference emerges from learning in neural circuits with a cost on reliability. arXiv preprint arXiv:180708952. 2018;.

52. Aridor G, Grechi F, Woodford M. Adaptive Efficient Coding: A Variational Auto-encoder Approach. bioRxiv. 2020; p. 2020.05.29.124453. doi:10.1101/2020.05.29.124453.

53. Maddison CJ, Mnih A, Teh YW. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. arXiv preprint arXiv:161100712. 2016;.

54. Jang E, Gu S, Poole B. Categorical reparameterization with gumbel-softmax. arXiv preprint arXiv:161101144. 2016;.

55. Rolfe JT. Discrete Variational Autoencoders. arXiv preprint arXiv:160902200. 2016;doi:https://doi.org/10.48550/arXiv.1609.02200.