# Efficient coding explains neural response homeostasis and stimulus-specific adaptation

Edward James Young
Computational and Biological Learning Lab
Department of Engineering
University of Cambridge
ey245@cam.ac.uk
&
Yashar Ahmadian
Computational and Biological Learning Lab
Department of Engineering
University of Cambridge
ya311@cam.ac.uk

**Abstract**

In the absence of adaptation, the average firing rate of neurons would rise or drop when changes in the environment make their preferred stimuli more or less prevalent. However, by adjusting the responsiveness of neurons, adaptation can yield firing rate homeostasis and stabilise the average rates of neurons at fixed levels, despite changes in stimulus statistics. In sensory cortex, adaptation is typically also stimulus specific, in that neurons reduce their responsiveness to over-represented stimuli, but maintain or even increase their responsiveness to stimuli far from over-represented ones. Here, we present a normative explanation of firing rate homeostasis grounded in the efficient coding principle, showing that this homeostasis yields an optimal trade-off between coding fidelity and the metabolic cost of neural firing. Unlike previous efficient coding theories, we formulate the problem in a computation-agnostic manner, enabling our framework to apply far from the sensory periphery. We then apply this general framework to Distributed Distributional Codes, a specific computational theory of neural representations serving Bayesian inference. We demonstrate how homeostatic coding, combined with such Bayesian neural representations, provides a normative explanation for stimulus-specific adaptation, widely observed across the brain, and how this coding scheme can be accomplished by divisive normalisation with adaptive weights. Further, we develop a model within this combined framework, and by fitting it to previously published experimental data, quantitatively account for measures of stimulus-specific and homeostatic adaption in the primary visual cortex.

## 1 Introduction

Neural population responses, and thus the computations on sensory inputs represented by them, are corrupted by noise. The extent of this corruption can, however, be modulated by changing neural gains. When the gain of a neuron increases, the strength of its response noise, or trial-to-trial variability, typically grows sublinearly and more slowly than its average response; for example, for Poisson-like firing, response noise grows as the square root of the mean response. The neuron can therefore increase its signal-to-noise ratio by increasing its gain or responsiveness. However, this increase in coding fidelity comes at the cost of an elevated average firing rate, and thus higher metabolic energy expenditure. From a normative perspective, coding fidelity and metabolic cost are thus two conflicting forces.

In this article, we ask what is the optimal way of adjusting neural gains in order to combat noise with minimal metabolic cost, *irrespective* of the computations represented by the neural population. As we will see, the answer depends on the prevailing stimulus statistics. We will thus address the following more specific question: given a neural population with *arbitrary* tuning curve shapes and noise distribution, how should neurons optimally adjust their gains depending on the stimulus statistics prevailing in the environment?

To address this question, we use efficient coding theory as our normative framework (Attneave, 1954; Nadal and Parga, 1999; Laughlin, 1981; Barlow, 2012; Linsker, 1988; Ganguli and Simoncelli, 2014; Wei and Stocker, 2015; Atick and Redlich, 1990). Efficient coding theories formalise the problem of optimal coding subject to biological constraints and costs, and by finding the optimal solution make predictions about the behavior of nervous systems (which are postulated to have been approximately optimised, *e.g.*, via natural selection). Concretely, the Infomax Principle (Linsker, 1988; Atick and Redlich, 1990; Ganguli and Simoncelli, 2014; Wei and Stocker, 2015) states that sensory systems optimise the mutual information

between a noisy neural representation and an external stimulus (or alternatively, the ideal noise-free result of computations on external stimuli), subject to metabolic constraints. In order to do so, the neural system should exploit the statistics of stimuli in its local environment (Simoncelli and Olshausen, 2001). Efficient coding theories can therefore predict how sensory systems should optimally *adapt* to changes in environmental stimulus statistics.

Suppose an environmental shift makes a stimulus feature more prevalent. Neurons sensitive to that feature will then see their preferred feature more often, and thus their average firing rate will initially increase. However, typically, the neurons do gradually adapt by reducing their responsiveness or gain (Solomon and Kohn, 2014; Clifford et al., 2007; Benucci et al., 2013). A special case of such adaptation is *firing rate homeostasis* (Desai, 2003; Turrigiano and Nelson, 2004; Maffei and Turrigiano, 2008; Hengen et al., 2013), in which adaptation brings the average firing rate back to the level prior to the environmental shift (Fig. 1). Thus, under firing rate homeostasis, neuronal populations maintain a constant stimulus-averaged firing rate in spite of changes to the environment (Benucci et al., 2013; Hengen et al., 2013; Maffei and Turrigiano, 2008).

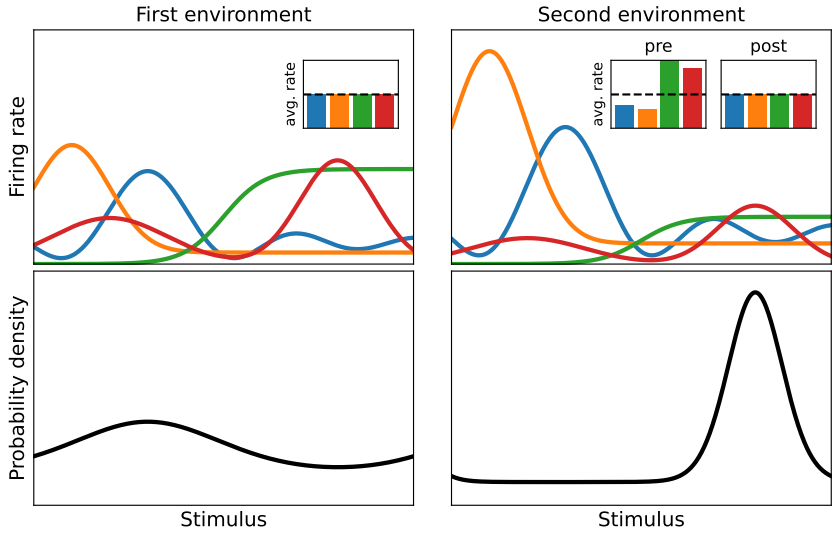


Figure 1: **Homeostatic adaptation via gain modulation (cartoon example).** Each column of panels corresponds to an environment as defined by its stimulus distribution (bottom row). The top panels show the post-adaptation tuning curves of the same four neurons in these two environments. Neurons adapt only by independently adjusting their gains, *i.e.*, by scaling their tuning curves up or down (correspondingly, tuning curves of the same color in the left and right columns, which belong to the same neuron, are scaled versions of each other and have identical shape). This adjustment is such that each neuron maintains the same stimulus-averaged firing rate in both environments. This is shown by the bar plots in the insets of the top row panels: the bar plot in the left column shows the average rates of the four neurons in the first environment, while the two bar plots in the right column show the average rates of these neurons before and after gain adaptation to the new statistics in the second environment. After adaptation the average rates return to their level before the environmental shift.

There are multiple levels at which homeostatic adaptation can be observed. Firstly, there is population homeostasis, in which the stimulus-average firing rate of an entire population of neurons remains constant, without individual neurons necessarily holding their rates constant (Slomowitz et al., 2015; Sanzeni et al., 2023). Secondly, there is what we term *cluster homeostasis*. In this form of homeostasis, stimulus-average firing rate of groups or clusters of neurons with similar stimulus tuning remains stable under environmental shifts (Benucci et al., 2013), but the firing rate of individual neurons within a cluster can change. Lastly, in the strongest form, homeostasis can occur at the level of individual neurons, in which case the firing rate of each individual neuron is kept constant under changes in environment statistics (Marder and Prinz, 2003). Previous normative explanations for firing rate homeostasis often focus on the necessity of preventing the network from becoming hypo- or hyperactive (Turrigiano and Nelson, 2004; Maffei and Turrigiano, 2008; Keck et al., 2013; Hengen et al., 2013). Although this might explain homeostasis at the population level, it does not adequately explain homeostasis at a more fine-grained level. As we will show in this paper, the answer to the normative question that we posed above provides an explanation for firing rate homeostasis at such a finer level. Specifically, we will show that, for a wide variety of neural noise models (and under conditions that we argue do obtain in the visual cortex), optimal adjustment of neural gains to combat the effect of coding noise predicts firing rate homeostasis at the level of clusters of similarly-tuned neurons.

Having shown the optimality of homeostatic codes without making assumptions on the nature of the neural representation and computation, we apply our framework to *Distributed Distributional Codes (DDC)* (Vertes and Sahani, 2018), as a specific computational theory for neural implementation of Bayesian inference. Combining DDC theory (which determines

tuning curve shapes) with our normative results (on optimal adjustments of neural gains) yields what we call *homeostatic DDC*. We will show that homeostatic DDC predict the typical finding that sensory adaptations are stimulus specific (Kohn, 2007; Schwartz et al., 2007). In stimulus specific adaptation (SSA), the suppression of response is greater for test stimuli that are closer to an over-represented adaptor stimulus than for stimuli further away (to which the neuron may even respond more strongly after adaptation). In particular, this causes a repulsion of tuning curves away from the adaptor (Movshon and Lennie, 1979; Müller et al., 1999; Dragoi et al., 2000, 2002; Benucci et al., 2013). We will show that homeostatic DDC are able to account for SSA effects which cannot be fully accounted for in previous efficient coding frameworks (Wei and Stocker, 2015; Ganguli and Simoncelli, 2014; Snow et al., 2016), and we will use them to quantitatively model homeostatic SSA as observed in V1, by re-analysing previously published data (Benucci et al., 2013). This model relies on a special class of homeostatic DDC which we term *Bayes-ratio coding*. We show that Bayes-ratio coding has attractive computational properties: it can be propagated between populations without synaptic weight adjustments, and it can be achieved by divisive normalisation with adaptive weights (Carandini and Heeger, 2012; Westrick et al., 2016).

We start the next section by introducing our efficient coding framework for addressing the normative question posed above. We show numerically that our framework predicts that the firing rate of clusters of similarly tuned neurons should remain constant despite shifts in environmental stimulus statistics, with individual neurons free to shuffle their average rates. This result is shown to hold across a diverse set of noise models, demonstrating the robustness of the optimality of homeostasis to the precise nature of neural response noise. We additionally demonstrate that our framework can account for a wide distribution of stimulus-averaged single-neuron firing rates as observed in cortex. We then provide an analytic argument that demonstrates why, within a certain parameter regime, homeostasis arises from our optimisation problem, and show that cortical areas, in particular the primary visual cortex (V1), are likely to be within this parameter regime. We then numerically validate the quality of our analytic approximations to the optimal solution. Lastly, we apply our theory to DDC representational codes, showing how homeostatic DDC can account for stimulus specific adaptation effects observed experimentally.

# 2 Results

## 2.1 Theoretical framework

The central question we seek to address is how the presence of response noise affects optimal neural coding strategies. Specifically, we consider how a neural population should distribute activity in order to best mitigate the effect of neural noise on coding fidelity, subject to metabolic constraints, irrespective of the computations it performs on sensory inputs (which we do not optimise, but take as given). We consider a population of $K$ neuronal units, responding to the (possibly high-dimensional) stimulus $\boldsymbol{s}$. We assume that our population engages in rate coding using time bins of a fixed duration, and denote the vector of joint population spike counts in a coding interval by $\boldsymbol{n} = (n_1, \ldots, n_K)$. Single-trial neural responses, $\boldsymbol{n}$, are taken to be noisy emissions from neural tuning curves, $\boldsymbol{h}(\boldsymbol{s})$, according to a noise distributions $\boldsymbol{n}|\boldsymbol{s} \sim P_{\text{noise}}(\boldsymbol{n}|\boldsymbol{s})$, subject to $\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}] = \boldsymbol{h}(\boldsymbol{s})$. We consider a number of different noise models, including correlated and power-law noise, showing that our main results are robust to specific assumptions about the nature of noise in neural systems. The sensory environment is characterised by a stimulus distribution $P(\boldsymbol{s})$; accordingly, an "environmental shift" corresponds to a change in $P(\boldsymbol{s})$ (making certain stimuli more or less prevalent).

We are interested in how changes in the stimulus distribution, $P$, affect the responsiveness of population neurons. We therefore adopt a shape-amplitude decomposition of the neural tuning curves. The tuning curve of the $a$-th unit, $h_a(\boldsymbol{s})$, is factorised into a *representational curve*, $\Omega_a(\boldsymbol{s})$, and a *gain*, $g_a$:

$$h_a(\boldsymbol{s}) = g_a \Omega_a(\boldsymbol{s}) = \text{gain} \times \text{representational curve}.$$

Fig. 2A demonstrates the effect of changing the gain while keeping the representational curve constant. We take the $\Omega_a$ to encode the computations represented by the neural population, and thus to be determined by those computational goals; we therefore treat them as given, and do not optimise them within our framework. Importantly, we do not make any assumptions on the shape of the representational curve: $\Omega_a$ can be any complex (*e.g.*, multi-modal, discontinuous) function of the possibly high-dimensional stimulus, and can thus represent *any* computation. In this respect, our treatment is more general than other efficient coding frameworks (*e.g.*, Ganguli and Simoncelli (2014)), which place tight constraints on the shape and configuration of the tuning curves (see the Discussion for a more detailed comparison of our approach with those studies). In particular, this generality enables our theory to apply to populations located deep in the processing pathway, and not just to primary sensory neurons.

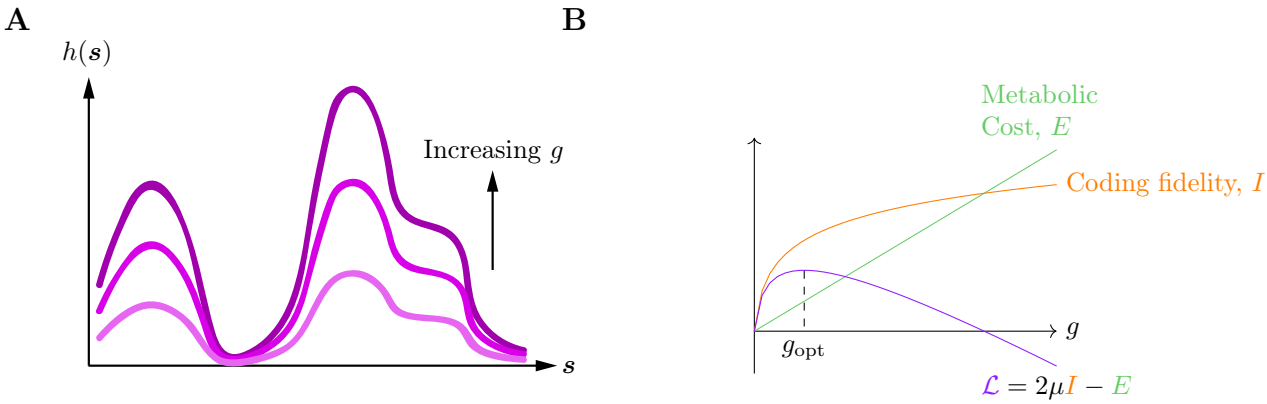


Figure 2: **The effect of changing the gain on tuning curves and different components of the objective function.** (A) As the gain, $g$, of a unit is increased, the shape of its tuning curve remains the same, but all firing rates are scaled up. (The cartoon shows a one-dimensional example, but our theory applies to general tuning curve shapes and joint configurations of population tuning curves, on high-dimensional stimulus spaces.) (B) Cartoon representation of the efficient coding objective function. Optimal neural gains maximise an objective function, $\mathcal{L}$, which is the weighted difference between mutual information, $I$, capturing coding fidelity, and metabolic cost, $E$, given by average population firing rate.

Our framework, relating environmental stimuli to neural tuning curves and finally to noise responses, is summarised by the diagram in Fig. 3.

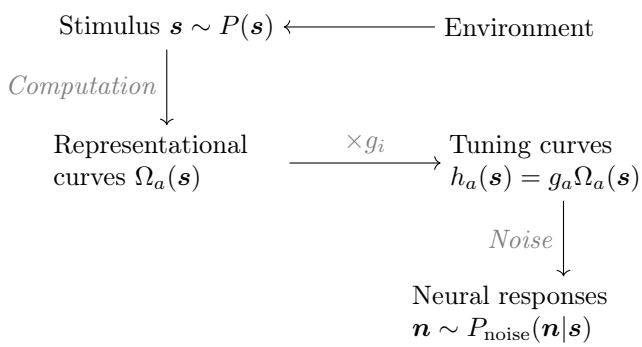


Figure 3: **Diagrammatic representation of the generative model underlying the theoretical framework.** The environment gives rise to a stimulus distribution $P$ from which the stimulus $\boldsymbol{s}$ is drawn. Conceptually, the brain performs some computation on $\boldsymbol{s}$, the result of which are represented by neurons via their representational curves $\Omega_a(\boldsymbol{s})$. These are multiplied by adaptive gains, $g_a$, to yields the actual tuning curves $h_a(\boldsymbol{s})$. The single-trial neural responses are noisy emissions based on $h_a(\boldsymbol{s})$.

We theorise that units adapt their gains to maximise an objective function that trades off the metabolic cost of neural activity with the information conveyed by the responses (Levy and Baxter, 1996; Ganguli and Simoncelli, 2014). Informally

$$\text{Objective} = \text{Coding fidelity} - \text{Metabolic cost}. \tag{1}$$

As is common in efficient coding frameworks (see *e.g.*, Atick and Redlich (1990); Ganguli and Simoncelli (2014)) we assume the main contributor to metabolic cost is the energy cost of emitting action potentials. Thus we take the metabolic cost term in the objective to be proportional to $\sum_{a=1}^{K} \mathbb{E}[n_a]$, the total average spike count fired by the population. On the other hand, a natural and common choice (up to approximations) for the coding fidelity is the mutual information between the stimulus and response, $I(\boldsymbol{n}; \boldsymbol{s})$. However, since in general the analytic maximisation of mutual information is intractable, in the tradition of efficient coding theory (Brunel and Nadal, 1998; Ganguli and Simoncelli, 2014; Linsker, 1988), we will instead optimise an approximate surrogate for mutual information. We decompose the mutual information as $I(\boldsymbol{s}; \boldsymbol{n}) = H[\boldsymbol{n}] - H[\boldsymbol{n}|\boldsymbol{s}]$. The marginal entropy term $H[\boldsymbol{n}]$ can be upper bounded by the entropy of a Gaussian with the same covariance,

$$H[\boldsymbol{n}] \leq \frac{1}{2} \ln \det(\text{Cov}[\boldsymbol{n}]) + \text{const}. \tag{2}$$

Substituting the right hand side for the marginal entropy term in the mutual information, gives us an upper bound for the

latter. Using this upper bound on mutual information as our quantification of "coding fidelity", yields the objective function

$$\mathcal{L}(\boldsymbol{g}) = \ln\det(\mathrm{Cov}[\boldsymbol{n}]) - 2\,H[\boldsymbol{n}|\boldsymbol{s}] - \mu^{-1}\sum_{j=1}^{N}\mathbb{E}[n_j], \tag{3}$$

where the constant $\mu > 0$ controls the information-energy trade-off. This trade-off is illustrated in Fig. 2B. To summarise, we assume that the optimal gains maximize the objective Eq. (3)

$$\boldsymbol{g}^{\mathrm{opt}} = \arg\max_{\boldsymbol{g}} \mathcal{L}(\boldsymbol{g}). \tag{4}$$

Note that because the $\Omega_a$ depend deterministically on the stimulus $\boldsymbol{s}$, and in turn fully determine the statistics of $n_a$, we have the identity $I(\boldsymbol{n};\boldsymbol{s}) = I(\boldsymbol{n};\boldsymbol{\Omega})$, which also holds for the upper bound on the mutual information that we are using. This means that the coding fidelity term in the objective function can equivalently be interpreted, without any reference to the low-level stimulus, as (an upper bound on) the mutual information between the population's noisy output and its ideal, noise-free outputs, $\boldsymbol{\Omega}$, as specified by the computational goals of the circuit.

## 2.2 Optimal adaptation of gains leads to rate homeostasis across a variety of noise models

We now investigate the consequences of optimal gain adaptation, according to Eqs. (3)–(4), on the statistics of population response. As our hypothesis is that gains adapt in order to optimally mitigate the effect of neural noise on coding fidelity (with minimal energy expenditure), ideally we would like our conclusions to be robust to the model of noise. We thus consider a broad family of noise models which allow for general noise correlation structure, as well as different (sub- or super-Poissonian) scalings of noise strength with mean response. Specifically, we assume that conditioned on the stimulus, the noisy population response, $\boldsymbol{n}$, has a normal distribution, with mean $\boldsymbol{h}(\boldsymbol{s})$, and covariance matrix

$$\text{noise covariance} = \mathrm{Cov}[\boldsymbol{n}|\boldsymbol{s}] = \sigma^2 h(\boldsymbol{s})^{\alpha}\Sigma(\boldsymbol{s})h(\boldsymbol{s})^{\alpha}, \tag{5}$$

where $h(\boldsymbol{s})$ is the diagonal matrix with the components of $\boldsymbol{h}(\boldsymbol{s})$ on its diagonal (more generally, we adopt the convention that the non-bold and index-free version of a bold symbol, representing a vector, denotes a diagonal matrix made from that vector), $\sigma$ scales the overall strength of noise, and $0 < \alpha < 1$ is a parameter that controls the power-law scaling of noise power with mean response $\boldsymbol{h}(\boldsymbol{s})$. Finally, $\Sigma(\boldsymbol{s})$ is the noise correlation matrix which can depend on the stimulus, but is independent of the gains. The noise model Eq. (5) can equivalently be characterised by (1) the noise variance of spike count $n_a$ is given by $\sigma^2 h_a(\boldsymbol{s})^{2\alpha}$, and (2) the noise correlation coefficient between $n_a$ and $n_b$ is given by $\Sigma_{ab}(\boldsymbol{s})$, independently of gains. Thus the strength (standard deviation) of noise scales sublinearly as gain to the power $\alpha$. Note that the case $\sigma = 2\alpha = 1$ corresponds to Poisson-like noise with variance of $n_a$ equal to its mean; in App. B.2, we show that, under mild conditions, a true Poisson noise model (with no noise correlations) leads to the same results as we derive below for the Gaussian noise model, Eq. (5), with $\sigma = 2\alpha = 1$ and $\Sigma(\boldsymbol{s}) = I$.

In general, the objective $\mathcal{L}$, Eq. (3), depends tacitly on the stimulus distribution $P(\boldsymbol{s})$ and the tuning curves $\boldsymbol{\Omega}(\boldsymbol{s})$. As shown in App. B.1, for our class of noise models, Eq. (5), $\mathcal{L}$ depends on $P$ and $\boldsymbol{\Omega}$ only through the following collection of parameters:

1. The units' pre-modulated average response (*i.e.*, the average spike count before multiplication by the adaptive gain $g_a$)

$$\omega_a := \mathbb{E}[\Omega_a(\boldsymbol{s})] = \int \Omega_a(\boldsymbol{s})P(\boldsymbol{s})d\boldsymbol{s}, \tag{6}$$

   These are related to the actual average responses, $r_a$, via $r_a = \mathbb{E}[h_a(\boldsymbol{s})] = g_a\omega_a$.

2. The coefficients of variation of $\Omega_a(\boldsymbol{s})$, which we denote by $\mathrm{CV}_a$.

3. The stimulus-averaged signal correlations, *i.e.*, the Pearson correlation coefficients between $\Omega_a(\boldsymbol{s})$'s. We denote the matrix of these signal correlations by $\rho$.

4. The matrix $W$ with elements defined by

$$W_{ab} = \mathbb{E}\left[\left(\frac{\Omega_a(\boldsymbol{s})}{\omega_a}\right)^{\alpha}\Sigma_{ab}(\boldsymbol{s})\left(\frac{\Omega_b(\boldsymbol{s})}{\omega_b}\right)^{\alpha}\right]. \tag{7}$$

   $W$ can be viewed as a stimulus-averaged normalised noise covariance matrix.[1]

[1] Note that for $\alpha = 1/2$ this matrix is a correlation matrix, in the sense that it is positive-definite and diagonal elements equal to 1.

Note that these quantities depend on the stimulus distribution $P$ and the functions $\Omega_a$. Note also that, except for $\omega_a$, the other three sets of quantities are invariant with respect to rescalings of the representational curves, and therefore can be equivalently defined in terms of the actual tuning curves, $h_a$, rather than $\Omega_a$. Thus, $\mathrm{CV}_a$'s are, equivalently, the coefficients of variation of the neurons' trial-averaged responses, and $\rho_{ab}$'s) are the correlation coefficients between pairs of trial-averaged responses (hence "signal correlation").

As shown in App. B.1, in terms of the above quantities, the objective function, Eq. (3), is given by

$$\mathcal{L}(\boldsymbol{g}) = \log\det\left(\sigma^2 W + (\omega g)^{1-\alpha}\mathrm{CV}\rho\mathrm{CV}(\omega g)^{1-\alpha}\right) - \mu^{-1}\sum_{a=1}^{K}\omega_a g_a \tag{8}$$

where, in line with the convention introduced above, CV, $\omega$, and $g$ denote $K \times K$ diagonal matrices with $\mathrm{CV}_a$, $\omega_a$, and $g_a$ on their diagonals, respectively. We have obtained the expression Eq. (8) for the objective for the case of a Gaussian noise model, Eq. (5). This noise model is very general, in the sense that it can capture any power-law relationship between the mean and variance of the response (given a stimulus) through $\alpha$, and any noise strength and correlation structure through $\sigma^2$ and $\Sigma(\boldsymbol{s})$. However, spike counts are discrete rather than continuous, a commonly used noise model for which is the Poisson distribution. In App. B.2, we demonstrate that, under appropriate conditions, for an independent Poisson noise model the objective Eq. (3) is approximately given by Eq. (8) with $\alpha = 1/2$ and with $\sigma^2 W$ replaced by the identity matrix.

We now present the results of numerical optimisation of the objective Eq. (8) for different noise distributions within the general family of noise models, Eq. (5). To capture the notion of adaptation under environmental shift, we simulated a family of environment models smoothly parameterised by $\nu \in [0, 1]$. An environment corresponds to a stimulus density $P(\boldsymbol{s})$, which, as we have just seen, affects the objective only via $\omega$, $\rho$, $W$, and CV. We will therefore specify environments through these four quantities. For simplicity, we fixed $\mathrm{CV}_a^2$, $\sigma^2$, and $\mu$ at 10, 1, and $5/(1-\alpha)$, in all environments. (Below, in Sec. 2.5, we will discuss the biological basis for these parameter choices.) On the other hand, $\omega_a$ and $\rho$ depended smoothly on the environment parameter $\nu$, as follows. For each $a = 1, \ldots, K$, $\omega_a(\nu = 0)$ and $\omega_a(\nu = 1)$ are drawn independently from a wide distribution (specifically the beta distribution $\mathrm{Beta}(6, 1)$). The value of $\omega_a(\nu)$ at intermediate values of $\nu$ were then obtained by linear interpolation between the independently sampled boundary values. This method for sampling environments is illustrated in Fig. 4. Similarly, to construct the matrix $\rho(\nu)$, we first constructed $\rho(\nu = 0)$ and $\rho(\nu = 1)$ with independent random structures and smoothly interpolated between them (see Sec. 4.1 for the details). We did this in such a way that, for all $\nu$, the correlation matrices, $\rho(\nu)$, were derived from covariance matrices with a $1/n$ power-law eigenspectrum (*i.e.*, the ranked eigenvalues of the covariance matrix fall off inversely with their rank), in line with the findings of Stringer et al. (2019) in the primary visual cortex. Finally, the stimulus-averaged noise correlation matrix, $W$, is determined by our choice of the noise model subfamily used in each simulation. We now present simulation results for three 1-parameter subfamilies of the general family of noise models, Eq. (5), in turn.

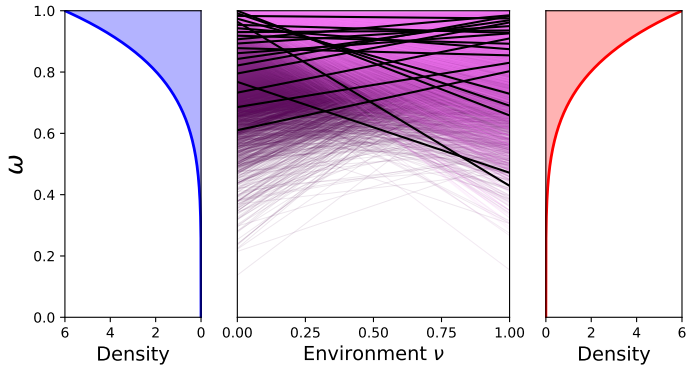


Figure 4: **The structure of pre-modulated average population responses in the simulated environments.** We simulated a one-parameter family of environments parameterised by $\nu \in [0, 1]$. In the environments on the two extremes, for each unit, the pre-modulated responses $\omega_a(\nu = 0)$ and $\omega_a(\nu = 1)$ are randomly and independently sampled from a $\mathrm{Beta}(1, 6)$ distribution, with density shown on the left and right panels. For intermediate environments (*i.e.*, for $0 < \nu < 1$), $\omega_a(\nu)$ is then obtained by linearly interpolating these values (middle panel). The middle panel shows 10000 such interpolations, with every 500th interpolation (ordered by initial value) coloured black.

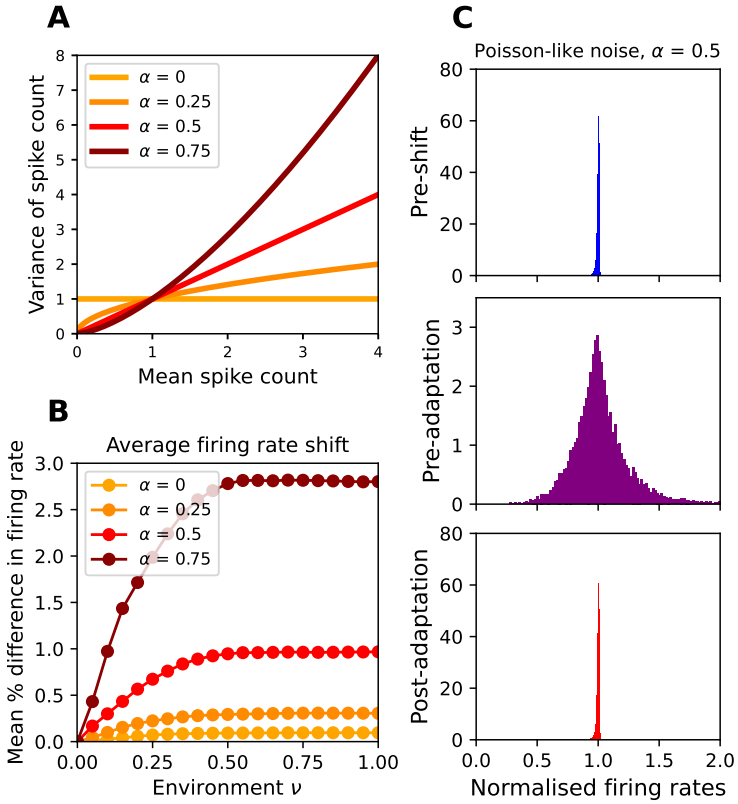


Figure 5: **Optimal gains under uncorrelated power-law noise lead to homeostasis of rates.** (A) Illustration of power-law noise. The trial-to-trial variance of spike counts, as a function of the trial-averaged spike count, is shown for different values of the noise scaling parameter $\alpha$. (C) Distribution of average firing rates before and after a discrete environmental shift from $\nu = 0$ to $\nu = 1$ for the case $\alpha = 1/2$. Top: the histogram of firing rates adapted to the original environment ($\nu = 0$) before the shift, as determined by the optimal gains Eq. (4) in this environment, $g_a^{\text{opt}}(\nu = 0)$. Middle: the histogram of firing rates immediately following the environmental shift to $\nu = 1$, but before gain adaptation; these are proportional to $\omega_a(\nu = 1)g_a^{\text{opt}}(\nu = 0)$. Bottom: the histogram of firing rates after adaptation to the new environment, proportional to $\omega_a(\nu = 1)g_a^{\text{opt}}(\nu = 1)$. We have normalised firing rates such that the pre-adaptation firing rate distribution has mean 1. (B) Deviation from firing rate homeostasis for different values of $\alpha$ as a function of environmental shift. For each environment (parametrised by $\nu$), we compute the optimal gains and find the adapted firing rate under these gains. The average relative deviation of the (post-adaptation) firing rates from their value before the environmental shift, $\frac{1}{K}\sum_a |r_a(\nu) - r_a(0)|/r_a(0)$, is plotted as a function of $\nu$.

**Uncorrelated power-law noise.** Consider first the case of no average noise correlations, in the sense that $W = I$.[2] We investigated the adaptation behavior of optimally gain-modulated population responses as a function of the noise scaling exponent $\alpha$. Fig. 5A shows the scaling of spike count variance as a function of mean spike count for different values of $\alpha$. By construction, in the absence of adaptation, an environmental shift ($\nu = 0 \to \nu = 1$) results in a wide range of average firing rates (Fig. 5C, middle panel). But following adaptation, the firing rates become tightly concentrated around a single value (Fig. 5C, bottom panel), which matches the value they were concentrated around in the original environment (Fig. 5C, top panel). Thus, adaptation results in (1) *uniformisation* of average firing rates across the population in a fixed environment, and (2) firing rate *homeostasis*, where following adaptation the mean firing rates of the units returns to their value before the environmental shift. These same uniformisation and homeostatic effects are seen for other values of $\alpha$ (Fig. 5B and Supp. Fig. 14). In Fig. 5B we quantify deviation from perfect homeostasis as the average % difference in the units' mean firing rates between the initial environment ($\nu = 0$) and environments at increasing $\nu$. Even under super-poissonian noise ($\alpha = 0.75$ in Fig. 5B), the average relative change in firing rates between environments never gets above 3%. Fig. 5B additionally shows that the slower spike count variance grows with mean spike count (*i.e.*, the smaller $\alpha$ is) the more precise the firing rate homeostasis.

[2] In terms of the objective function, Eq. (8), this is mathematically equivalent to having zero stimulus-conditioned noise correlations, *i.e.*, $\Sigma(\boldsymbol{s}) = 1$, but with CV's redefined not to denote the standard coefficient of variation, but rather $\text{SD}[\tilde{\Omega}]/\sqrt{\mathbb{E}[\tilde{\Omega}^{2\alpha}]}$ where $\tilde{\Omega} := \Omega/\omega$; for Poisson-like scaling, $\alpha = 1/2$, this reduces to the standard coefficient of variation.

**Uniform noise correlations.** In the second sub-family of noise models we allow for noise correlations, but consider only Poisson-like scaling of noise power (*i.e.*, $\alpha = 1/2$). More precisely, we assume that the effective noise correlation coefficients between all pairs of units (corresponding to off-diagonal elements of the matrix $W$, Eq. (7)) are the same, equal to $p$ (Fig. 6A).[3] As shown in Fig. 6C, optimal gain adaptation leads to uniformisation and homeostasis of mean firing rates in this noise model as well. In fact, in the presence of uniform noise correlations, homeostasis is even stronger compared to the previous noise model with zero noise correlations; in this case, mean relative deviations in average firing rates never exceed 1% (Fig. 6B). Further, as noise correlations get stronger (*i.e.*, as $p$ increases) we see tighter homeostasis.

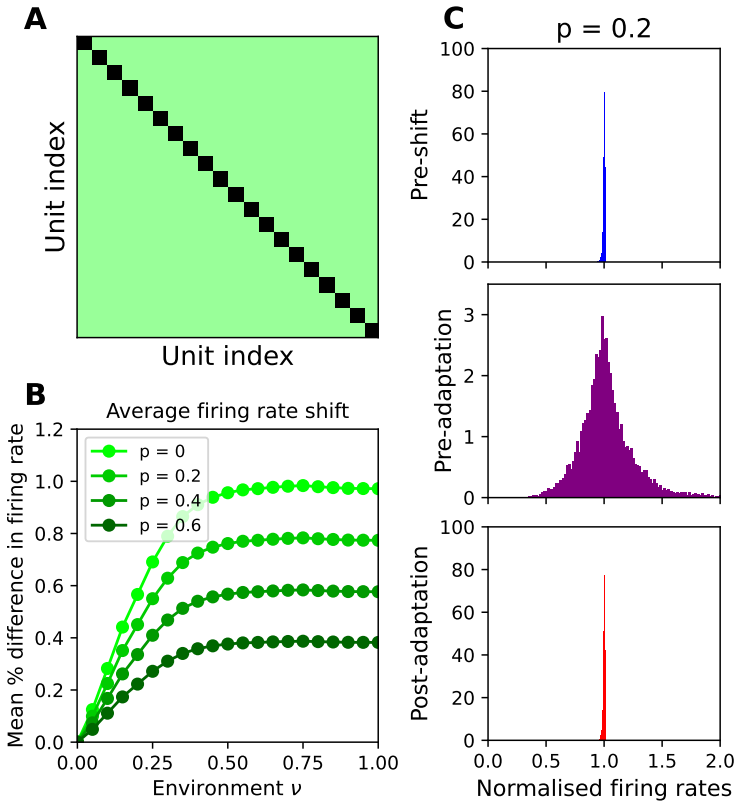


Figure 6: **Firing rate homeostasis under Poissonian noise with uniform noise correlations.** (A) Illustration of an effective noise correlation matrix with uniform correlation coefficients (off-diagonal elements). Panels B and C have the same format as in Fig. 5.

**Aligned correlated noise.** In the final noise model we considered, we once again fixed $\alpha = 1/2$ (Poisson-like scaling), but allowed for heterogeneous effective noise correlation coefficients. Specifically, we took the effective noise correlation matrix $W$ to have approximately the same eigenbasis as $\rho$, the signal correlation matrix —hence "aligned noise"—, but with a different eigenvalue spectrum. As described above, for all noise sub-families, $\rho(\nu)$ was obtained by normalising a covariance matrix with $1/n$ spectrum (corresponding to the findings of Stringer et al. (2019) in V1); in the current case, we obtained $W(\nu)$ by normalising a covariance matrix with the same eigenbasis but with a $1/n^{\gamma}$ spectrum (*i.e.*, with the $n$-th largest eigenvalue of $W$ scaling as $1/n^{\gamma}$); see Fig. 7A, and Sec. 4.1 for further details. Fig. 7C demonstrates that, as with the other noise sub-families, optimal gain adaptation in the presence of aligned correlated noise also leads to homeostasis and uniformisation of mean firing rates (see Supp. Fig. 16 for rate histograms for other values of $\gamma$). Across different values of $\gamma$, homeostasis is again tighter compared to the zero noise correlation case, with the average firing rate shift never exceeding 1% (Fig. 5B). Further, with the exception of $\gamma = 1$, we see more homeostasis for higher values of $\gamma$, corresponding to lower dimensional noise. The special case of $\gamma = 1$ corresponds to so-called information-limiting noise (Moreno-Bote et al., 2014; Kanitscheider et al., 2015), where noise and signal correlation structures are fully aligned: $W = \rho$. In this case, homeostasis is perfect, and the mean relative error in firing rates is 0. We provide an analytic proof that in this case we have perfect homeostasis and uniformisation in App. B.5, and discuss this situation further in Sec. 2.5 and Sec. 2.6.

[3] The positivity of the noise correlation matrix, $W$, requires that $p > -\frac{1}{K-1}$. Assuming large $K$, the right side of the inequality goes to zero; hence we only simulated cases with positive $p$.

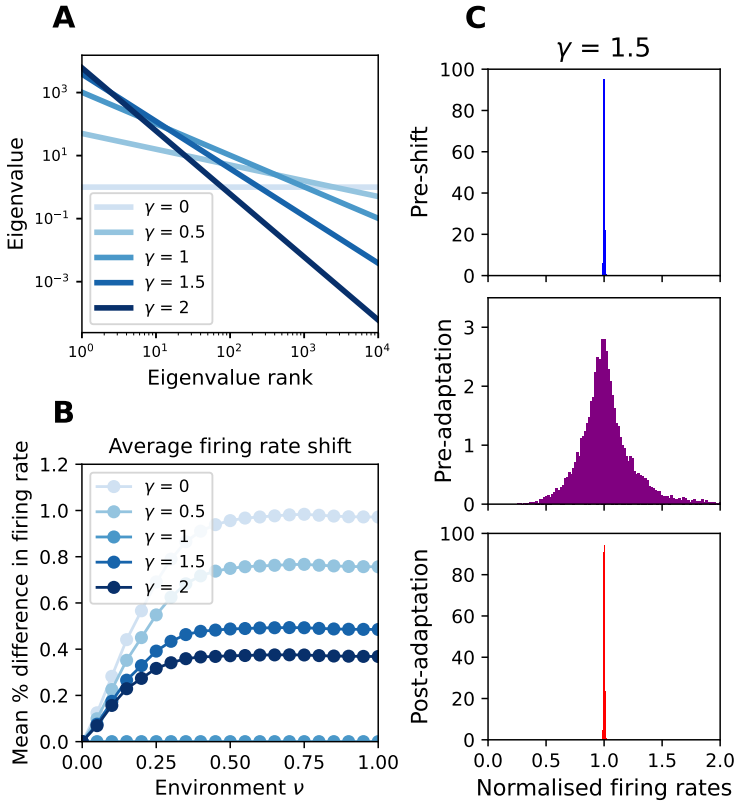


Figure 7: **Firing rate homeostasis under Poissonian, aligned noise.** (A) Eigenspectra of noise covariance matrix for the aligned noise model for various values of $\gamma$; in this model, the $n$-th eigenvalue is proportional to $1/n^{\gamma}$. Panels B and C have the same format as in Fig. 5.

We have shown that, according to our efficient coding framework, optimal adaptation of response gains (with the goal of combating neural noise at minimal metabolic cost) can robustly account for the firing rate homeostasis of individual units within a population, despite environmental shifts, under a diverse family of noise models. To shed light on these results, in Sec. 2.4, we analytically investigate sufficient conditions on the parameters, Eq. (6)–(7), of our model under which we expect homeostasis to emerge. Following that, in Sec. 2.5, we will show that biological estimates for these parameters obtained in cortex (which are determined by natural stimulus statistics, as well as the stimulus dependence of high-dimensional population responses) are consistent with these conditions.

In addition to homeostasis, our framework also predicts uniformisation of average firing rates across the population, in any given environment. By contrast, cortical neurons display a wide range of average firing rates spanning a few orders of magnitude (Buzsáki and Mizuseki, 2014). However, we can interpret the units in our preceding numerical experiments not as single neurons, but as clusters of similarly tuned neurons. In this interpretation, uniformisation occurs at the cluster level, with potentially a wide distribution of single-cell average firing rates within each cluster. In the next section, we show that our model is indeed consistent with such an interpretation. In particular, in the case of uncorrelated noise with Poisson-like mean-variance scaling, we show that a simple extension of our model gives rise to both diversity of firing rates as seen in cortex as well as homeostasis and uniformisation at the level of clusters of similarly tuned neurons.

## 2.3 A clustered population accounts for the diversity of firing rates seen in cortex

A cortical neuron can have very similar response tuning to other neurons in the same area (particularly with nearby cells). We will call a collection of neurons which have similar (but not necessarily identical) representational curves a *cluster*. To gain intuition and allow for analytic solutions, we will henceforth adopt a toy model in which the $N$ neurons in the population are sorted into $K$ such clusters. In this toy model, neurons within a cluster have very similar stimulus tuning, but neurons in different clusters are tuned differently; the model thus exaggerates a more realistic situation in which there is a more gradual transition from similar to dissimilar tuning curves within the entire population. More concretely, we will assume that the representational curves of the neurons, indexed by $i$, belonging to cluster $a$, are small perturbations to the same representational curve characterising that cluster; *i.e.*,

$$\hat{\Omega}_i(\boldsymbol{s}) = \Omega_a(\boldsymbol{s}) + \epsilon\delta\Omega_i(\boldsymbol{s}) \tag{9}$$

where $\epsilon$ is a small parameter controlling the deviation from perfect within-cluster similarity. Below, hatted symbols ($\hat{\bullet}$) denote variables defined analogously to Eq. (6)–(7) but for single neurons, with $\Omega_a(\boldsymbol{s})$ replaced by $\hat{\Omega}_i(\boldsymbol{s})$.

We assume single-neuron gains adapt to optimise an objective $\mathcal{L}_{\text{pop}}$, defined analogously to the objective $\mathcal{L}$, Eq. (8), but for individual neurons rather than clusters. In the case of uncorrelated noise with Poisson-like scaling, this objective is given by

$$\mathcal{L}_{\text{pop}}(\hat{\boldsymbol{r}}) = \ln\det\Big(\sigma^2 I_N + \hat{r}^{1/2}\hat{\text{CV}}\hat{\rho}\hat{\text{CV}}\hat{r}^{1/2}\Big) - \mu^{-1}\sum_{i=1}^{N}\hat{r}_i \tag{10}$$

where $\hat{r}_i$ are the mean firing rates of individual neurons, given by $\hat{r}_i = \hat{\omega}_i\hat{g}_i = \mathbb{E}[\hat{\Omega}_i(\boldsymbol{s})]\hat{g}_i$; note that maximising $\mathcal{L}_{\text{pop}}$ in terms of the single-neuron gains is equivalent to maximising it in terms of $\hat{r}_i$. In App. B.3, we show that this objective can be expanded in $\epsilon$ as

$$\mathcal{L}_{\text{pop}}(\hat{\boldsymbol{r}}) = \mathcal{L}(\boldsymbol{r}) + \epsilon^2\mathcal{L}_{\text{pert}}(\hat{\boldsymbol{r}}) + \mathcal{O}\left(\epsilon^3\right), \tag{11}$$

Here, the leading order term, $\mathcal{L}(\boldsymbol{r} = \omega\boldsymbol{g})$, is a cluster-level objective and is given by Eq. (8), now viewed as a function of the cluster rates $r_a$, *i.e.*, the total mean firing rate of neurons in a cluster; thus $r_a = \sum_{i\in C_a}\hat{r}_i$ where $C_a$ is the set of neurons in cluster $a$. Note that this term is a function only of the cluster firing rates. Thus, for $\epsilon = 0$ (perfect similarity within clusters), the efficient coding objective function is blind to the precise distribution of single-cell firing rates, $\hat{r}_i$, among the neurons of a cluster, as long as the total rate of the cluster is held fixed. This is because for clusters of identically-tuned and independently-firing neurons with Poisson-like noise, the cluster can be considered as a single coherent unit whose firing rate is given by the sum of the firing rates of its neurons. Thus at zeroth order, the distribution of individual neuron rates are free and undetermined by the optimisation, as long as they give rise to optimal cluster rates. At small but nonzero $\epsilon$, $\mathcal{L}_{\text{pop}}(\hat{\boldsymbol{r}})$ is still dominated by $\mathcal{L}$; we thus expect this term to play the dominant role in determining the cluster rates, with terms of order $\epsilon^2$ and higher having negligible effect. However, the perturbation term $\mathcal{L}_{\text{pert}}(\hat{\boldsymbol{r}})$ breaks the invariance of the loss with respect to changes in the distribution of single neuron gains (or rates) within clusters.

We will therefore approximate the maximisation of the total objective function $\mathcal{L}_{\text{pop}}$ as follows. We first specify the cluster rates by maximising $\mathcal{L}(\boldsymbol{r})$, obtaining the optimal cluster rates $r_a^{\text{opt}}$. We then specify the rates of individual neurons within the population by maximising the leading correction term in the objective, $\mathcal{L}_{\text{pert}}(\hat{\boldsymbol{r}})$, subject to the constraint that the single-neuron rates in each cluster sum to that cluster's optimal rate. We reinterpret the optimisation in the previous sections to correspond to the first stage of this two-stage optimisation, with population units there corresponding to clusters of similarly tuned neurons. As we have seen in Sec. 2.2, optimisation of the cluster-level objective, $\mathcal{L}$, can result in homeostasis and uniformisation across clusters.

As for the distribution of individual neuron rates, we show in App. B.3.4 (see the derivation of Eq. (91)) that, in the parameter regime (see Secs. 2.4 and 2.5) in which cluster-level homeostasis and uniformisation occurs, the perturbation objective $\mathcal{L}_{\text{pert}}$ decomposes into a sum of objectives over different clusters. Thus, in the second stage of optimisation the problem decouples across clusters, and (as shown in App. B.3.4) the optimal single-neuron rates in cluster $a$ maximise the objective

$$\arg\min \sum_{i,j\in C_a}\hat{r}_i\text{Cov}(\Delta\hat{\Omega}_i, \Delta\hat{\Omega}_j)\hat{r}_j - \mu\sum_{i\in C_a}\text{Var}(\Delta\hat{\Omega}_i)\hat{r}_i, \tag{12}$$

subject to the constraints

$$\sum_{i\in C_a}\hat{r}_i = r_a, \qquad \hat{r}_i \geq 0. \tag{13}$$

Here $\Delta\Omega_i$ is defined to be the centered, zero-mean version of $\delta\Omega_i$:[4]

$$\Delta\hat{\Omega}_i(\boldsymbol{s}) = \delta\hat{\Omega}_i(\boldsymbol{s}) - \mathbb{E}[\delta\hat{\Omega}_i(\boldsymbol{s})]\frac{\Omega_a(\boldsymbol{s})}{\omega_a}. \tag{14}$$

Note that the optimisation problem, Eqs. (12)–(13), is a quadratic program.

We solved the optimisation problem Eqs. (12)–(13) numerically for 10,000 randomly generated covariance matrices (note that different instances of the optimisation can be thought of as characterising different clusters within the same population). In general, the covariance matrix, $\text{Cov}(\Delta\hat{\boldsymbol{\Omega}})$ (which defines this optimization problem), depends on the stimulus distribution, the cluster tuning curve, and properties (*e.g.*, degree of smoothness) of the intra-cluster variations in single-cell tuning curves. In our simulations, instead of making specific arbitrary choices for these different factors, we used a minimal toy model where the covariance matrix was generated as the matrix of inner products of $k$ random vectors in a $D$-dimensional space (see Sec. 4.2 for details). In App. B.3.5 we show that $D$ can be interpreted as the dimensionality of the space of independent perturbations to the cluster-wide tuning curve $\Omega_a$ which vary significantly on the portion of stimulus space to which the cluster tuning curve responds (below, we will refer to $D$ as the effective tuning curve dimension).

In Fig. 8 (see also Supp. Fig. 18) we show the resulting firing rate distributions for neurons aggregated across all clusters (corresponding to the different optimisations), for different values of the cluster size $k$ and effective tuning curve

[4] We can interpret $\Delta\Omega_i$ as the direction of the perturbed tuning curve in excess of the cluster tuning curve $\Omega_a$.

dimension, $D$, of within-cluster tuning curve variations. Firstly, note that, depending on these parameters, a significant fraction of neurons can be silent, *i.e.*, with zero firing rate in the optimal solution; the figure shows the rate distributions for the active neurons, as well as their percentage of all neurons. Our theory therefore predicts that a significant fraction of cortical neurons are silent in any given sensory environment, and that which neurons are silent can shuffle following shifts in environmental stimulus statistics (specifically shifts that result in significant changes in the signal correlation structure of the cluster neurons). Secondly, note that the distributions span multiple orders of magnitude, in agreement with empirical observations in the cortex (see *e.g.* Slomowitz et al. (2015); Hengen et al. (2013); Maffei and Turrigiano (2008)), and is approximately log-normal (albeit with a skewed tail towards small log-rates), also consistent with empirical findings (Buzsáki and Mizuseki, 2014). Lastly, for fixed $k$, as the effective tuning curve dimension, $D$, increases, the fraction of silent neurons decreases (Supp. Fig. 18).

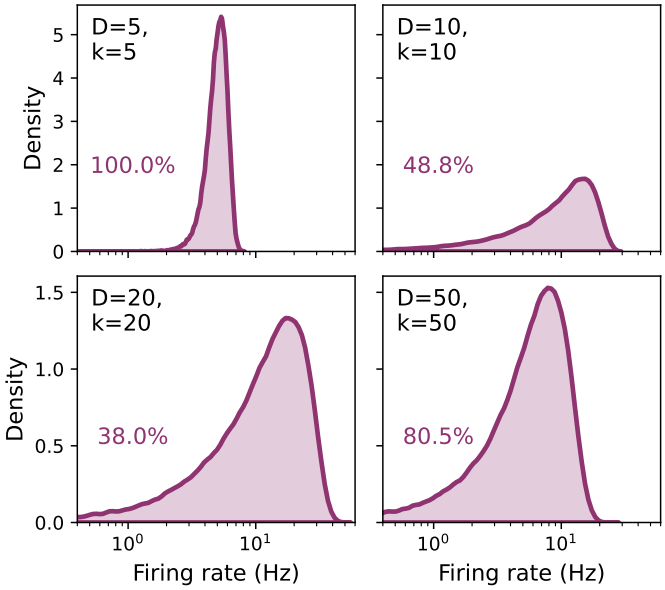


Figure 8: **The distribution of single neuron firing rates is consistent with empirical observations of approximately log-normal rates.** Each panel shows the distribution of log firing rates of the active single neurons for different choices of cluster size $k$ and effective tuning curve dimension, $D$, of the space of within-cluster tuning curve variations (see the main text for the definition, and Sec. 4.2 for further details). The distributions shown are for the active neurons with nonzero firing rate, with their percentage of the total population given in each panel (purple text). For all simulations we fixed the cluster firing rate to $\mu = k \times 5$ Hz, such that the mean rate of single neurons is 5 Hz on average. The plotted densities were obtained by averaging over $10,000$ optimisations of Eq. (12) based on different random draws of $\mathrm{Cov}(\Delta\mathbf{\Omega})$.

To summarise, we have shown that the objective $\mathcal{L}$ considered in Sec. 2.2, reinterpreted to correspond to the cluster level, can arise from optimisation of a similar objective, $\mathcal{L}_{\mathrm{pop}}$, defined at the level of single neurons (Eqs. (10)–(11)). Furthermore, the correction term to the cluster-level objective, $\mathcal{L}_{\mathrm{pert}}$, serves to break the symmetry within clusters, generating a diverse range of firing rates which matches the distribution of cortical single-neuron firing rates. In the rest of the paper, we will pursue the problem at the coarser level of clusters.

## 2.4 Analytical insight on the optimality of homeostasis

The numerical results of Sec. 2.2 showed that, across a range of neural noise models, optimal gain modulation for combating noise can give rise to firing rate homeostasis and uniformisation across neural clusters defined based on similarity of stimulus tuning. To shed light on these results, we analytically investigated how and in what parameter regimes these results arise. We found that homeostasis and uniformisation emerge in the regime where the matrix

$$\Delta := \frac{\sigma^2}{(\beta\mu)^\beta} \left(\mathrm{CV}\rho\mathrm{CV}\right)^{-1} W \tag{15}$$

is small, according to some notion of matrix norm; here $\beta = 2 - 2\alpha$. The intuitive meaning of this matrix is that its inverse, $\Delta^{-1}$, characterises the structure of signal-to-noise ratio (SNR) in the space of cluster responses (before gain modulation). Therefore the above condition can be roughly interpreted as the requirement that signal-to-noise ratio is strong. We will discuss this condition, and the constraints it imposes on various model parameters, further in Sec. 2.5, and will provide evidence that it is likely to hold in cortex. Here, we focus on its consequences.

We developed a perturbative expansion in $\Delta$, to obtain approximate analytical solutions for the cluster gains that maximise our objective $\mathcal{L}$, Eq. (8) (see App. B.4). In the first approximation, *i.e.*, to zeroth order in $\Delta$, this expansion yields

$$g_a \approx \frac{\beta\mu}{\omega_a}. \tag{16}$$

We denote this approximate solution by $g_a^{(0)}$. To see the significance of this result, note that the average response of cluster $a$, *i.e.*, its stimulus-averaged total spike count is given by $\mathbb{E}[h_a(\boldsymbol{s})] = g_a\omega_a$. Thus, in the leading approximation, the stimulus-average spike count of cluster $a$ is given by the constant $g_a^{(0)}\omega_a = \beta\mu$. Thus, since $\beta\mu$ is constant across both clusters and environments, the zeroth order solution yields homeostasis and uniformisation of cluster firing rates.

We also calculated the first order correction to this homeostatic solution. This is given by

$$g_a \approx \frac{\beta\mu}{\omega_a}(1 - \Delta_{aa}) := g_a^{(1)}. \tag{17}$$

which (assuming $\Delta_{aa}$ are small) yields approximate homeostasis, with small deviations depending on the SNR structure, as captured by $\Delta_{aa}$, and how that structure shifts between environments.

Improvements to the zeroth-order homeostatic solution can also be obtained in a different manner: by finding the best configuration of cluster gains, $g_a^{\text{hom}}$, that, by construction, are guaranteed to yield homeostasis and uniformisation. As we saw above, the latter arise if and only if $g_a\omega_a$ is constant (across clusters and environments). We denote this *a priori* unknown constant by $\chi$. In other words, we let

$$g_a^{\text{hom}} = \frac{\chi}{\omega_a}, \tag{18}$$

under which the average spike count of clusters in the coding interval is given by $\chi$. The variable $\chi$ is then chosen to optimise the expectation of $\mathcal{L}$, Eq. (8), across a relevant family of environments (note that we need to optimise the average objective across environments, since otherwise the optimal $\chi$ will in general depend on the environment, invalidating homeostasis). With such an optimal $\chi$, $\boldsymbol{g}^{\text{hom}}$ will, by definition, perform better than $\boldsymbol{g}^{(0)}$ in terms of our objective function. Further, it can be shown (see App. B.7) that $\chi < \beta\mu$, *i.e.*, the mean cluster firing rate under the optimal homeostatic solution is strictly smaller than that predicted by $\boldsymbol{g}^{(0)}$, and that the average spike-count of the entire population of clusters is approximately the same under this solution as it is under $g^{(1)}$. In Sec. 2.6, we numerically compare the performance of $\boldsymbol{g}^{(0)}, \boldsymbol{g}^{(1)}$, and $\boldsymbol{g}^{\text{hom}}$, and find that in many cases, the homeostatic $\boldsymbol{g}^{\text{hom}}$ performs close to the (numerically computed) true optimal gains.

Finally, we note the following exact analytical result regarding homeostasis that holds even for non-small $\Delta$. It follows form the structure of the objective function, Eq. (8), that, irrespective of how pre-modulation mean rates, $\omega_a$, change between two stimulus environments, as long as $\Delta$ and $W$, the effective noise covariance, remain invariant across the environments, optimal gains imply exact homeostasis of each cluster's average firing rate across the two environments. However, unless $\Delta$ is small, in general the optimal solution does not result in uniformization of firing rates across clusters.

## 2.5 Conditions for the validity of the homeostatic solution hold in cortex

In this section, we further examine the conditions under which the homeostatic solution, Eq. (16), provides a good approximation to the optimal gains, and, based on empirical estimates of relevant quantities, discuss whether these conditions obtain in cortical populations. As noted above, the general mathematical condition is that the matrix $\Delta$, defined in Eq. (15), is in some sense small (recall that through its inverse, $\Delta$ represents the structure of signal-to-noise ratio —before gain modulation— in the coding space). We can make this statement more concrete as follows. The homeostatic solution, and the first order correction to it, Eq. (17), arise from a perturbative asymptotic expansion in the small $\Delta$. When the leading order correction to the (zeroth-order) homeostatic solution becomes non-negligible, it indicates the breakdown of this expansion. From Eq. (17) we can thus reformulate the condition for the validity of the homeostatic approximation as

$$|\Delta_{aa}| = \frac{\sigma^2}{(\beta\mu)^\beta}\left|\sum_{b=1}^{K}\frac{\rho_{ab}^{-1}W_{ab}}{\text{CV}_a\text{CV}_b}\right| \ll 1 \tag{19}$$

for all $a$. Before proceeding to empirical estimates, let us first delineate the dependence of $\Delta$ (or its diagonal elements) on different model parameters. Firstly, note that $\Delta$ scales linearly with $\sigma^2$; thus, if noise is scaled up, so is $\Delta$. Next, consider increasing the noise scaling exponent $\alpha$. This decreases $(\beta\mu)^\beta$ provided $\alpha \leq 1 - e/2\mu$ (which holds in our simulations above). Moreover, increasing $\alpha$ increases the diagonal elements of $W$ (see Eq. (7)). Intuitively, this increases the magnitude of $W$. Through these two mechanisms, increasing $\alpha$ will generally increase the size of $\Delta$. Next, note that $\Delta$ has an inverse relationship with the coefficients of variation, CV, which increase as neurons become more stimulus selective; accordingly, the more selective the tuning curves, the smaller the $\Delta$. Lastly, $\Delta$ has a complex dependency on the spectra of both CV$\rho$CV and $W$ as well as the alignment of their eigenbases (which represent the geometric structure of signal and noise in

the neural coding space, respectively). In Sec. 2.6, we will numerically investigate these dependencies, and their effects on the validity of the homeostatic approximation.

Translated to biological quantities, it follows from the above dependencies that the smallness of $\Delta$ (or $\Delta_{aa}$) requires that (1) the cluster firing rates are sufficiently high, (2) noise variance scales sufficiently slowly with mean response, (3) individual clusters are sufficiently selective in responding to stimuli, and (4) the neural representation has a high-dimensional geometry in the population response space. We now review the available biological data on these variables in cortical populations.

**High firing rate.** We first estimate $\beta\mu$ which enters the prefactor in Eq. (19). We see from the zeroth order solution Eq. (16) that $\beta\mu$ is approximately the average total spike count of a cluster over the rate-coding time interval. Condition Eq. (19) thus requires that the stimulus-averaged firing rate of all clusters are sufficiently high. A wide range of mean firing rates for individual neurons have been reported in cortex. Here we focus on firing rates in rodent V1 during free behaviour (which tends to be lower compared to rates in cat or monkey cortex). Reported values tend to range from 0.4 Hz (Greenberg et al., 2008) to 14 Hz (Parker et al., 2022), depending on the cortical layer or area or the brain state, with other values lying in a tighter range of 4-7 Hz (Hengen et al., 2013; Szuts et al., 2011; Torrado Pacheco et al., 2019). Therefore, for rodent V1, we take a mean firing rate of 5 Hz to be a reasonable rough estimate. Assuming a coding interval of 0.1 seconds, and a cluster size of $k = 20$ (as a guess for the number of similarly coding V1 neurons), we obtain $\beta\mu \sim 10$. (Note that for lower or higher single neuron firing rates, a similar value of $\mu$ could be achieved by correspondingly scaling cluster sizes up or down.)

**Sufficiently slow scaling of noise variance.** We can obtain an estimate for the prefactor $\sigma^2/(\beta\mu)^\beta$, by fitting the mean-variance relationship for cluster spike counts: variance $= \sigma^2 \text{mean}^{2-\beta}$. A number of papers have fit such a relationship (Shadlen and Newsome, 1998; Gershon et al., 1998; Moshitch and Nelken, 2014; Koyama, 2015). We adjusted the values reported by those papers for our chosen coding interval length and cluster size (see Sec. 4.3, the plot in Fig. 13, and the values of $\beta$ and $\sigma^2$ thereby obtained). These (together with the estimate $\beta\mu \approx 10$) yield values for $\sigma^2/(\beta\mu)^\beta$ ranging from 0.05 (Shadlen and Newsome, 1998) to 0.31 (Koyama, 2015), with median value 0.14. Note that this is close to the value for Poissonian noise ($\sigma^2 = 1$, $\beta = 1$), which is 0.1.

**Sufficiently selective responses.** The coefficient of variation CV can be seen as a measure of neural selectivity or sparseness of neural responses. To see this, consider a toy model in which the cluster responds at a fixed level to a fraction $p$ of stimuli and is silent otherwise. In this case, $\text{CV}^2 = (1-p)/p \approx 1/p$ for small $p$. Our condition therefore requires that neurons are sufficiently selective in their responses and respond only to a small fraction of stimuli. Lennie (2003) places the fraction of simultaneously active neurons (which we use as a proxy for the response probability of a single cluster) at under 5%. For our toy model, this yields the estimate $\text{CV}^2 \approx 20$. The Bernoulli distribution in the toy model is particularly sparse, and so we take $\text{CV}^2 \approx 10$ as a more conservative estimate.

**High-dimensional signal geometry.** Assuming, for simplicity, that the coefficients of variation are approximately the same for all clusters, we see that the expression for $\Delta_{aa}$ are proportional to the diagonal elements of $\rho^{-1}W$. To estimate the latter, we will assume that these elements are all comparable in size,[5] and will therefore estimate their average value, which is given by the normalised trace $\text{tr}(\rho^{-1}W) = 1/K \sum_{a=1}^{K} (\rho^{-1}W)_{aa}$. By the invariance of trace, this can equally be characterised as the mean eigenvalue of $\rho^{-1}W$. We will estimate this trace for two sub-cases of interest.

Firstly, consider the case of zero noise correlations, *i.e.*, $W = I$. In this case, the trace is the mean of $[\rho^{-1}]_{aa}$. Recall that $\rho$ is the signal correlation matrix of neuron clusters. For a correlation matrix, it can be shown that $[\rho^{-1}]_{aa} \geq 1$, with equality if and only if cluster $a$ has zero signal correlation with every other cluster. $[\rho^{-1}]_{aa}$ can thus be seen as a measure of the extent to which cluster $a$ shares its representation with other clusters, *i.e.*, as a measure of representational redundancy. Many traditional efficient coding accounts predict zero signal correlation between neurons, at least in the low noise limit (Barlow, 2012; Atick and Redlich, 1990; Nadal and Parga, 1994, 1999), providing an additional normative justification for low signal correlations. However, complete lack of signal correlations is not necessary for our condition to hold; we merely require a sufficiently slow decay of the eigenvalue spectrum of the signal correlation matrix. This condition is geometrically equivalent to neural responses forming a representation with high intrinsic dimensionality. Stringer *et al.* (Stringer et al., 2019) found that, the signal covariance matrix of mouse V1 neurons responding to natural stimuli possesses an approximately $1/n$ spectrum. In this case, for $K$ large, we obtain the estimate $[\rho^{-1}]_{aa} \approx \text{tr}(\rho^{-1})/K \approx \ln(K)/2$ (see App. B.6). Given the slow logarithmic increase with $K$, we expect that our condition can hold even for very large neural populations. For example, suppose we take the entire human V1 as our neural population, to obtain an arguably generous upper bound for $K$. This contains roughly $1.5 \times 10^8$ neurons (Wandell, 1995), leading to $7.5 \times 10^6$ clusters of 20 neurons, and therefore an average value of $[\rho^{-1}]_{aa}$ just under 8. This, together with our other estimates above, yields the following

[5] This is expected to be valid when the eigenbasis of $\rho^{-1}W$ is not aligned with the standard (cluster) basis, which, in turn, corresponds to distributed coding.

overall estimate for $\Delta_{aa}$

$$\Delta_{aa} \sim \frac{\sigma^2}{(\beta\mu)^\beta}\frac{1}{\mathrm{CV}^2}\mathrm{tr}(\rho^{-1}) \sim 0.14 \times \frac{1}{10} \times 8 = 0.112. \tag{20}$$

We now consider another sub-case of interest, that of so-called *information-limiting noise* where directions of strong signal variations align with those of strong noise variation (Moreno-Bote et al., 2014; Kanitscheider et al., 2015). Specifically, we consider the extreme case $W = \rho$, *i.e.*, when noise and signal correlations are perfectly aligned. This case was considered numerically in Sec. 2.2, Fig. 7, and we saw that in this case the optimal solution displayed perfect homeostasis and uniformisation of rates across clusters. Moreover, in this case, $\rho^{-1}W = I$, and so the normalised trace $\mathrm{tr}(\rho^{-1}W)$ does not scale with $K$ at all. Note that this is an extreme of signal-noise alignment. In Sec. 2.2 we also considered less extreme forms of alignment, in which $W$ shares the eigen-bases with $\rho$ but has a $1/n^\gamma$ spectrum (as opposed to the $1/n$ spectrum for $\rho$). We show in App. B.6 that, for $0 < \gamma < 1$ (relatively high-dimensional and weakly correlated noise), we have $\mathrm{tr}(\rho^{-1}W) \approx \left(\frac{1-\gamma}{2-\gamma}\right)\ln(K) < \frac{1}{2}\ln(K)$. On the other hand, for $\gamma > 1$ (relatively low-dimensional and strongly correlated noise), $\mathrm{tr}(\rho^{-1}W)$ in fact decays like $K^{-\min(\gamma-1,1)}$ as $K$ grows. Thus, in this case, the estimate of $\Delta_{aa}$ is (potentially significantly) smaller than that of Eq. (20). Accordingly, even partial signal-noise alignment (especially for correlated, relatively low-dimensional noise) can strongly reduce the size of $\Delta$, encouraging homeostatic coding.

The above analysis makes it clear when we should expect homeostasis in general – when cluster firing rates are *high*, the noise scaling of cluster responses is sufficiently *slow*, responses are highly *selective*, and signal correlation structure corresponds to a *high-dimensional geometry* (Stringer et al., 2019), possibly with *information-limiting noise correlations*. On the other hand, when these conditions are violated, the optimal gain configuration can potentially deviate strongly from the homeostatic solution. Moreover, given the above estimates, we thus see that in V1 and possibly other cortical areas, we can expect the leading corrections to the homeostatic solution (see Eq. (17) and Eq. (20)) to not exceed 10%, in relative terms. Indeed, as we will show in the Sec. 2.6 using numerical validations, for values of the above parameters in their estimated biological range, the homeostatic solution provides a very good approximation to the optimal solution of Eq. (3).

## 2.6 Numerical comparison of homeostatic gains to optimal gains

We now return to the numerical simulations performed in Sec. 2.2, to compare the performance and accuracy of the different approximate solutions to optimally adapted gains, discussed in Sec. 2.4, with those of the optimal solution of Eq. (8). For each of the different noise models considered in that section we compare the performance of the homeostatic gains $g_a^{(0)} = \mu\beta/\omega_a$ and $g_a^{\mathrm{hom}} = \chi/\omega_a$ (where $\chi$ is chosen as detailed in App. B.7), as well as the first-order correction solution $\boldsymbol{g}^{(1)}$ (see Eq. (16)–(18)) to the numerically optimised gains $\boldsymbol{g}^{\mathrm{opt}}$. To examine the adaptation properties of these gains, we once again work with the sequence of environments indexed by $\nu \in [0, 1]$, with the same specifications as in the numerical simulations of Sec. 2.2.

To measure the accuracy of the approximations to the optimal gains, for each environment $\nu$, we calculated the mean relative errors

$$\frac{1}{K}\sum_{a=1}^{K}\frac{|g_a^{\mathrm{app}}(\nu) - g_a^{\mathrm{opt}}(\nu)|}{g_a^{\mathrm{opt}}(\nu)}, \tag{21}$$

for each of the approximate solutions, denoted by $\boldsymbol{g}^{\mathrm{app}}(\nu)$, with $\boldsymbol{g}^{\mathrm{opt}}(\nu)$, the numerically optimised solution in environment $\nu$. To measure the performance of an approximate solution, we use the following *relative improvement* measure, defined by

$$C^{\mathrm{app}}(\nu) = \frac{\mathcal{L}(\boldsymbol{g}^{\mathrm{app}}(\nu);\nu) - \mathcal{L}(\boldsymbol{g}^{\mathrm{app}}(0);\nu)}{\mathcal{L}(\boldsymbol{g}^{\mathrm{opt}}(\nu);\nu) - \mathcal{L}(\boldsymbol{g}^{\mathrm{app}}(0);\nu)}. \tag{22}$$

which can be interpreted as the improvement in the objective $\mathcal{L}(\cdot;\nu)$ (the efficient coding objective Eq. (8) according to the statistics of environment $\nu$) achieved by the adaptive gains $\boldsymbol{g}^{\mathrm{app}}(\nu)$ over the unadapted gains from the original environment $\boldsymbol{g}^{\mathrm{app}}(0)$, relative to the improvement obtained by the optimally adaptive gains $\boldsymbol{g}^{\mathrm{opt}}(\nu)$. The better the approximation the closer $C^{\mathrm{app}}(\nu)$ will be to 1. Note that, in general, the relative improvement worsens when $\nu$ goes to 0 (see Fig. 9B, D and F); this follows from the definition of $C(\nu)$ and the fact that when environmental change is weak, using gains optimized in the original environment without adapting them in the new environment is superior to adaptive but only approximately optimal gains. We now summarise the results for the different noise models discussed in Sec. 2.2 (see that section for the detailed specification of the three noise models).

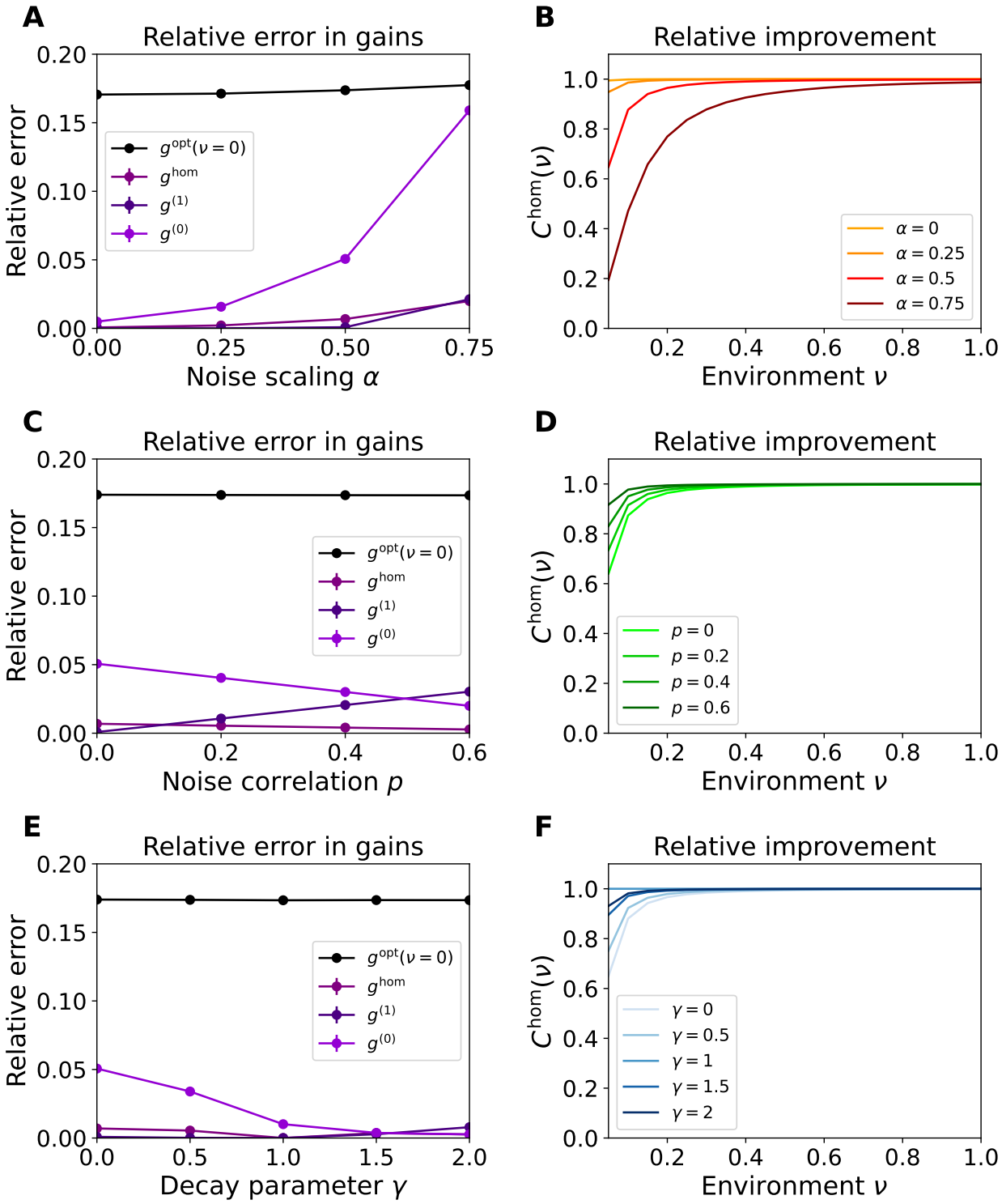


Figure 9: **Accuracy of homeostatic approximation to optimal gains.** (A) The relative error, Eq. (21), of different approximate solutions for the optimal gains, averaged across environments (the standard deviations of these values across environments were negligible relative to the average), as a function of the noise scaling parameter $\alpha$. To give a sense of the scale of variation of the optimal gains across environments, the black line shows the relative change in the optimal gains between the most extreme environments (a measure of effect size), $\frac{1}{K}\sum_{a=1}^{K}\frac{|g_a^{\text{opt}}(0)-g_a^{\text{opt}}(1)|}{g_a^{\text{opt}}(0)}$. (B) Relative improvement in the objective for the homeostatic approximation $\boldsymbol{g}^{\text{hom}}$ as measured by Eq. (22). Panels C and E (panels F and D) are the same as panel A (panel B), but for the uniformly correlated noise and the aligned noise models, respectively.

**Uncorrelated power-law noise.** First, we consider the uncorrelated noise model with general power-law scaling of noise strength with mean responses. In this case, the homeostatic solution $\boldsymbol{g}^{\text{hom}}$, and the first order correction, $\boldsymbol{g}^{(1)}$, had very

low ($< 3\%$) relative errors for all noise scaling exponents, $\alpha$, while the relative error in $\boldsymbol{g}^{(0)}$ becomes large for significantly super-Poissonian noise scaling, *i.e.*, $\alpha = 0.75$ (Fig. 9A).[6] The relative improvement measure $C^{\text{hom}}(\nu)$ for the homeostatic solution is also close to one, the optimal value, for all $\alpha$, except for 0.75 (Fig. 9B). For the other approximate solutions too (Supp. Fig. 17A,B), the relative improvement decreases with increasing noise scaling. This is consistent with our analysis, since increasing $\alpha$ increases the size of $\Delta$ (see Sec. 2.4) and hence the influence of terms of higher order in $\Delta$.

**Uniform noise correlation coefficients.** We now examine the effect of noise correlations on the accuracy of the approximate homeostatic solutions, first within the noise family with uniform noise correlation coefficients, denoted by $p$, and Poisson-like scaling ($\alpha = 1/2$). For all $p$, the relative deviation of the homeostatic solution $\boldsymbol{g}^{\text{hom}}$ from optimal gains was below 1% (Fig. 9C), while $\boldsymbol{g}^{(0)}$ and $\boldsymbol{g}^{(1)}$ were less accurate overall. Note that the accuracy of the homeostatic solutions $\boldsymbol{g}^{\text{hom}}$ and $\boldsymbol{g}^{(0)}$ increases with $p$, suggesting that positive noise correlations improve the optimality of homeostatic gains. The same trend can be seen in the relative improvement on the objective, $C^{\text{hom}}$, for homeostatic gain adaptation, which increases with $p$ (Fig. 9D).

**Aligned noise.** Lastly, we considered the algined noise model, in which noise has Poisson-like scaling (*i.e.*, $\alpha = 1/2$), and is correlated across clusters with the noise and signal covariance matrices sharing the same eigenbasis. In this noise family, the parameter, $\gamma$, controls the scaling of the eigenvalues of the noise covariance matrix with their rank, $n$, as $1/n^{\gamma}$. As in the previous cases, for all $\gamma$, the homeostatic solution $\boldsymbol{g}^{\text{hom}}$ was very close to the optimal gains $\boldsymbol{g}^{\text{opt}}$ (Fig. 9E), and preformed nearly equally (Fig. 9F). The case $\gamma = 1$, which corresponds to perfect alignment between signal and noise correlations, $W = \rho$, is particularly interesting. As noted in the previous section, this case corresponds to that of so-called information-limiting noise correlations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015). As we show in App. B.5, in this case the optimal gains, without any approximation, lead to homeostasis and uniformisation of cluster firing rates. Thus, in this case $\boldsymbol{g}^{\text{opt}} = \boldsymbol{g}^{\text{hom}}$ exactly, leading to zero relative error (Fig. 9E) for $\boldsymbol{g}^{\text{hom}}$, and perfect relative improvement $C^{\text{hom}}$ (Fig. 9F). For other cases, we can see that for increasing $\gamma$ (at least for the values we simulated), which corresponds to more correlated and lower-dimensional noise (as noise power falls off faster as a function of eigenvalue rank), the homeostatic gains $\boldsymbol{g}^{\text{hom}}$ tend to perform better.

Our numerical simulations across the three noise subfamilies (with varying noise power scaling, noise correlations, and noise spectrum) demonstrate that homeostatic strategies can - robustly and across a wide range of noise distributions - combat coding noise by effectively navigating the trade-off between energetic costs and coding fidelity. Finally, note that adaptation in $\boldsymbol{g}^{\text{hom}}$ uses only information local to each neuron, *i.e.*, its firing rate. This implies that good performance does not require the use of complex regulation mechanisms which take into account how the encoding of each cluster relates to the population at large; homeostatic regulation, implemented by purely local mechanisms, is sufficient.

## 2.7 Synaptic normalisation allows for propagation of homeostasis between populations

So far, we have considered homeostatic coding from a purely normative perspective, showing with both theoretical arguments and numerical simulations that homeostasis efficiently trades-off between coding fidelity and energetic costs. In this section, we begin to address the question of biological plausibility; specifically, what biological mechanisms are necessary for the homeostatic coding regime to be propagated between neural populations. Specifically, We will show that for the case of a linear feedforward network implementing the representation curves, there is a connection between homeostatic coding and synaptic scaling (Turrigiano, 2008).

Consider two neural populations, with an upstream population providing feedforward input to a downstream population. We will suppose that the upstream population is engaged in homeostatic coding, and ask what is necessary for the downstream population to be also engaged in homeostatic coding. We will denote the tuning and representational curves of the upsteam population by $h_a^{(i)}(\boldsymbol{s})$ and $\Omega_a^{(i)}(\boldsymbol{s})$, respectively, where $i = 1$ ($i = 2$) for the upstream (downstream) population. We drop corrections to the zeroth order solution for optimal gains, and work within the homeostatic coding regime; that is we assume the gains of the upstream population are given by $g_a = \chi/\omega_a^{(1)} = \chi/\mathbb{E}[\Omega_a^{(1)}(\boldsymbol{s})]$.

The downstream population of tuning curves will be given by $h_m^{(2)}(\boldsymbol{s})$ for $m = 1, \ldots, M$. Let $W_{ma}$ be the synaptic weight from neuron $a$ in the upstream population to neuron $m$ in the downstream population. Working in a linearized rate model, this gives us that the tuning curve of neuron $m$ is

$$h_m^{(2)}(\boldsymbol{s}) = \sum_{a=1}^{K} W_{ma} h_a^{(1)}(\boldsymbol{s}). \tag{23}$$

Suppose that, as dictated by the computational goals of the circuit, the downstream population has representational curves $\Omega_m^{(2)}(\boldsymbol{s})$ for $m = 1, \ldots, M$. We similarly write each representational curve as a linear combination of the upstream cluster

[6] For the case of uncorrelated power-law noise model, as well as the aligned correlated noise model, we used an analytical approximation to the optimal $\chi$ (see App. B.7) used in the homeostatic approximation $g_a^{\text{hom}} = \chi/\omega_a$, Eq. (18). This means that the performance measures in the plots of Fig. 9B and F are lower-bounds for the performance of $g_a^{\text{hom}}$ with the truly optimal $\chi$.

representational curves,

$$\Omega_m^{(2)}(\boldsymbol{s}) = \sum_{a=1}^{K} w_{ma} \Omega_a^{(1)}(\boldsymbol{s}). \tag{24}$$

We call these weights $w_{ma}$ the *computational weights*, since they are determined by the computational goals of the downstream population. In line with our approach so far, we treat these computational goals, and hence also the computational weights, as given (*i.e.*, set independently of optimal gain adaptation).

We asked how the synaptic weights $W_{ma}$ should depend on the computational weights $w_{ma}$, and how they should adapt as stimulus statistics change in order for the downstream population to also engage in homeostatic coding. In App. B.8 we find that homeostatic coding in the downstream population is guaranteed if the synaptic weights are related to the computational weights via

$$W_{ma} = \frac{w_{ma}\omega_a^{(1)}}{\sum_{b=1}^{K} w_{mb}\omega_b^{(1)}} = \frac{w_{ma} g_a^{-1}}{\sum_{b=1}^{K} w_{mb} g_b^{-1}}, \tag{25}$$

where $\omega_b^{(1)} = \mathbb{E}\left[\Omega_b^{(1)}(\boldsymbol{s})\right]$ as before. Note that the total synaptic weight received by a downstream neuron is thus constant. In other words, this scheme requires the normalization of the total synaptic weights received by each downstream neuron to be constant, independently of adjustments of the gains of the pre-synaptic neurons or of possible changes in the computational weights.[7] Thus, homeostatic coding, applied sequentially to two populations, provides an additional normative interpretation of synaptic normalisation, in which synapses onto a neuron are jointly scaled to keep total input weights constant. This may be the computational reason for the synaptic scaling observed in certain studies of firing rate homeostasis (Turrigiano et al., 1998; Turrigiano, 2008) that occurs over relatively long timescales.

## 2.8 Homeostatic noisy DDCs

Up to this point we have made no assumptions about the nature of cortical representations, beyond rate coding (and the condition described by Eq. (19)), and thus our framework has been independent of the computational goals of the circuit. We now apply our framework to a specific theory of neural representation and computation, namely the distributive distributional code (DDC) (Vertes and Sahani, 2018), introducing what we term a *homeostatic DDC*. We will show below, in Sec. 2.10, that specific forms of homeostatic DDC are able to account for stimulus-specific adaptation effects, observed in V1 (Benucci et al., 2013), which have been difficult to account for under other efficient coding frameworks. While there is ample behavioural and psychophysical evidence for the hypothesis that animal perception and decision making approximate optimal Bayesian probabilistic inference (Kersten et al., 2004; Stocker and Simoncelli, 2006; van Bergen et al., 2015; Geurts et al., 2022), our knowledge of the neural implementation of computations serving Bayesian inference (such as computations of posterior distributions and posterior expectations) are rudimentary at best. The DDC framework is one proposal for neural implementations of Bayesian computations underlying perception. In a DDC model, neural responses directly encode posterior expectations of the inferred latent causes of sensory inputs according to an internal generative model. This internal model mathematically relates the latent causes or latent variables, which we denote by $\boldsymbol{z}$, to the sensory inputs or stimuli, $\boldsymbol{s}$, via a family of conditional distributions, $f(\boldsymbol{s}|\boldsymbol{z})$, as well as a prior distribution, $\pi(\boldsymbol{z})$, over the latent variables. The sensory system is assumed to implement a so-called *recognition model* corresponding to this generative model. Namely, given a stimulus $\boldsymbol{s}$, the task of the sensory system is to invert the generative process by calculating and representing the posterior distribution of latent variables given the current sensory input (see the schema in Fig. 10).

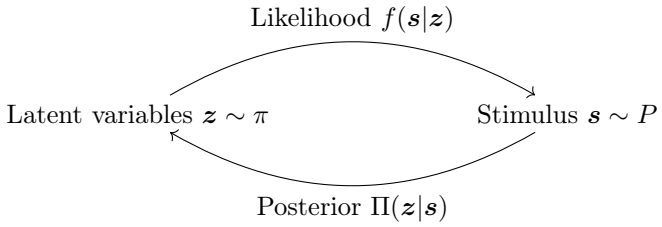


Figure 10: **Schema of the generative and recognition models underlying a DDC.** According to the internal generative model, sensory inputs, $\boldsymbol{s}$, are caused by latent causes, $\boldsymbol{z}$, that occur in the environment according to a prior distribution $\pi(\boldsymbol{z})$. A conditional distribution $f(\boldsymbol{s}|\boldsymbol{z})$, the so-called likelihood function, describes how the latent causes generate or give rise to the sensory inputs $\boldsymbol{s}$. The task of the brain is to invert this generative process by inferring the latent causes based on the current sensory input, which is done by computing the posterior distribution $\Pi(\boldsymbol{z}|\boldsymbol{s})$.

[7] Different normalisation factors are equivalent to different values of $\chi$ for different populations. Differences in the optimal choice of $\chi$ may arise from *e.g.*, different noise correlation statistics or different rate coding intervals between populations.

In a DDC, neural responses are specifically postulated to directly represent the posterior expectations of the latent variables, $\boldsymbol{z}$, or rather, the posterior expectations of a sufficiently rich (and fixed) class of so-called kernel functions of $\boldsymbol{z}$. Thus, each neuron, say neuron $a$, is assigned to a so-called kernel function of the latent variables $\phi_a(\boldsymbol{z})$, and its response represents the posterior expectation $\mathbb{E}[\phi_a(\boldsymbol{z})|\boldsymbol{s}]$. Here, we add two assumptions to this standard DDC scheme. First, we assume that the actual single-trial response, $n_a$, of the neuron is a noisy version of $\mathbb{E}[\phi_a(\boldsymbol{z})|\boldsymbol{s}]$, but such that the trial-averaged tuning curve of the neuron is $\mathbb{E}[\phi_a(\boldsymbol{z})|\boldsymbol{s}]$, up to a constant of proportionality. Second, we assume that these constants of proportionality are adaptive and are given by the optimal gains of our efficient coding framework, which we assume are the homeostatic gains given by $g_a^{\text{hom}} = \chi/\omega_a$. In the language of previous sections, this is equivalent to saying that we assume the representational curve of a neuron with kernel function ,$\phi_a(\boldsymbol{z})$, is given by $\Omega_a(\boldsymbol{s}) = \mathbb{E}[\phi_a(\boldsymbol{z})|\boldsymbol{s}]$. Under optimal homeostatic gains, our noisy homeostatic DDC predicts the following relationship between the neural tuning curves and Bayesian posterior and prior expectations of $\phi_a(\boldsymbol{z})$:

$$h_a(\boldsymbol{s}) = \chi \frac{\mathbb{E}[\phi_a(\boldsymbol{z})|\boldsymbol{s}]}{\mathbb{E}[\phi_a(\boldsymbol{z})]}. \tag{26}$$

Thus, this homeostatic DDC framework predicts that the neural tuning curves are the ratio of the posterior to the prior expectations of kernel functions. In deriving this result we have used the tower property to replace the trial-averaged posterior expectation $\mathbb{E}[\mathbb{E}[\phi_a(\boldsymbol{z})|\boldsymbol{s}]]$ with the prior expectation $\mathbb{E}[\phi_a(\boldsymbol{z})]$ under the internal model; strictly speaking, this is true under an ideal-observer internal model in which the generative internal model (as specified by $f(\boldsymbol{s}|\boldsymbol{z})$ and $\pi(\boldsymbol{z})$) provides a bona fide, objective description of the statistics of sensory stimuli $\boldsymbol{s}$. In Sec. 2.10 below, we will relax this assumption and discuss the consequences of using a non-ideal internal model instead.

## 2.9 Bayes-ratio coding

We now consider a special case of the homeostatic DDC in which the kernel functions are delta functions, $\phi_a(\boldsymbol{z}) = \delta(\boldsymbol{z} - \boldsymbol{z}_a)$. Here $\boldsymbol{z}_a$ is a point in latent variable space corresponding to neuron $a$'s preferred latent variable value. In this case, the homeostatic DDC becomes

$$h_a(\boldsymbol{s}) = \chi \frac{\Pi(\boldsymbol{z}_a|\boldsymbol{s})}{\pi(\boldsymbol{z}_a)} = \chi \frac{f(\boldsymbol{s}|\boldsymbol{z}_a)}{P(\boldsymbol{s})}, \tag{27}$$

where in the last equality we used Bayes' rule: $\Pi(\boldsymbol{z}|\boldsymbol{s}) = f(\boldsymbol{s}|\boldsymbol{z})\pi(\boldsymbol{z})/P(\boldsymbol{s})$. Thus, in this case, the tuning curve of a neuron is the ratio of the posterior distribution to the prior distribution both evaluated at that neuron's preferred latent variable value. Accordingly, we call this coding scheme *Bayes-ratio coding*.

Bayes-ratio coding can be achieved via divisive normalisation with adaptive weights. Divisive normalisation has been dubbed a canonical neural operation (Carandini and Heeger, 2012; Westrick et al., 2016), as it appears in multiple brain regions serving different computational goals. Given the feedforward inputs, $F_a(\boldsymbol{s})$, the divisive normalisation model computes the response (and thus the tuning curve) of neuron $a$ as

$$h_a(\boldsymbol{s}) = \chi \frac{F_a(\boldsymbol{s})^n}{\alpha^n + \sum_b w_b F_b(\boldsymbol{s})^n}. \tag{28}$$

Here, $w_b$ are a collection of so-called normalisation weights, and $\alpha \geq 0$ and $n \geq 1$ are constants. As we now show, Bayes-ratio coding can be naturally implemented by a divisive normalisation model in which $n = 1$. In the divisive normalisation implementation of Bayes-ratio coding, the feed-forward inputs, $F_b(\boldsymbol{s})$, are given by the likelihood function, $f(\boldsymbol{s}|\boldsymbol{z}_b)$, of the DDC's internal generative model, while the normalisation weights $w_b$ encode the prior probabilities, $w_b = \pi(\boldsymbol{z}_b)\delta\boldsymbol{z}_b$ (the volume element, $\delta\boldsymbol{z}_b$, is chosen such that the latent variable space is the disjoint union of volumes of size $\delta\boldsymbol{z}_b$ each containing their corresponding sample point $\boldsymbol{z}_b$). We then obtain

$$\begin{aligned}\sum_b w_b F_b(\boldsymbol{s}) = \sum_b f(\boldsymbol{s}|\boldsymbol{z}_b)\pi(\boldsymbol{z}_b)\delta\boldsymbol{z}_b &\approx \int f(\boldsymbol{s}|\boldsymbol{z})\pi(\boldsymbol{z})d\boldsymbol{z} \\ &= P(\boldsymbol{s}),\end{aligned}$$

which after substitution in Eq. (28) yields

$$h_a(\boldsymbol{s}) \approx \chi \frac{f(\boldsymbol{s}|\boldsymbol{z}_a)}{\alpha + P(\boldsymbol{s})}. \tag{29}$$

Taking the limit as $\alpha \to 0$, we obtain Bayes-ratio coding Eq. (27). Therefore, provided $\alpha$ is small compared to $P(\boldsymbol{s})$, divisive normalisation can be used to approximate Bayes-ratio coding. Not only does this result show that implementing Bayes-ratio coding is biologically plausible, it also provides a normative computational interpretation to both the feedforward inputs and the normalisation weights of the divisive normalization model, as the internal generative model's likelihoods and prior probabilities, respectively. Note also that the normalisation weights of this model are adaptive, and vary between environments with different latent variable marginal distributions, $\pi(\boldsymbol{z})$.

Finally, we consider the propagation of Bayes-ratio coding between populations, by using the scheme discussed in Sec. 2.7 for propagation of homeostatic codes. Suppose we now have a hierarchical generative model $\boldsymbol{z}^{(2)} \to \boldsymbol{z}^{(1)} \to \boldsymbol{s}$, specified by a prior distribution $\pi^{(2)}(\boldsymbol{z}^{(2)})$ and generative likelihood functions $f^{(2)}(\boldsymbol{z}^{(1)}|\boldsymbol{z}^{(2)})$ and $f^{(1)}(\boldsymbol{s}|\boldsymbol{z}^{(1)})$. Then consider a recognition model corresponding to this generative model, comprised of a feedforward network with two populations or layers that implement a Bayes-ratio encoding of the posterior distributions of the hierarchical latent variables $\boldsymbol{z}^{(1)}$, and $\boldsymbol{z}^{(2)}$, respectively, conditional on the stimulus $\boldsymbol{s}$. Let us assume that the first (upstream) layer is implementing Bayes-ratio coding for the lower-level variables $\boldsymbol{z}^{(1)}$ Then, as we show in App. B.9, a simple linear propagation of rates according to Eq. (23) with synaptic weights given by

$$W_{ma} = f^{(2)}\big(\boldsymbol{z}_a^{(1)}\big|\boldsymbol{z}_m^{(2)}\big)\delta\boldsymbol{z}_a^{(1)}, \tag{30}$$

would implement Bayes-ratio coding for $\boldsymbol{z}^{(2)}$ in the downstream layer. This result is significant for three reasons. Firstly, the synaptic weights make no reference to the prior distribution, $\pi^{(2)}$, over $\boldsymbol{z}^{(2)}$. There is therefore no need to adapt the weights when the environment changes, if the change only affects the statistics of the higher-level causes $\boldsymbol{z}^{(2)}$ but not the generative process. Secondly, while the downstream population, is part of the *recognition* model and (effectively) represents the *posterior* distribution of $\boldsymbol{z}^{(2)}$, the synaptic weights, Eq. (30), needed to implement this representation only require knowledge of the *generative* probabilities (specifically those given by $f^{(2)}$). Finally, we can see by induction that this scheme can be propagated through many layers, each representing posterior probabilities of higher-level latent variables further up a generative hierarchy.

## 2.10 Homeostatic DDCs account for stimulus specific adaptation

We found that homeostatic DDCs can provide a normative account of stimulus specific adaptation (SSA). Here, we start by showing this in the special case of Bayes-ratio coding, which is mathematically simpler and provides good intuitions for a more general subset of homeostatic DDCs. We then discuss this more general case and build a homeostatic DDC model of empirically observed SSA in the primary visual cortex (V1) (Benucci et al., 2013).

According to Eq. (27), for Bayes-ratio codes the tuning curve of neuron $a$ is given by $h_a(\boldsymbol{s}) = \chi f(\boldsymbol{s}|\boldsymbol{z}_a)/P(\boldsymbol{s})$. Suppose now that the stimulus distribution in the environment, $P(\boldsymbol{s})$, changes due to a change in the statistics of the latent causes, $\pi(\boldsymbol{z})$, while the causal processes linking the latent causes and observed stimuli, as captured by $f(\boldsymbol{s}|\boldsymbol{z})$, remain stable. In this case, by Eq. (27), the tuning curve of *all* neurons are modified by a multiplicative factor given by the ratio of the $P(\boldsymbol{s})$ in the old environment to that in the new environment. This will lead to a multiplicative suppression of the response of all neurons to stimuli that are over-represented in the new environment relative to the old one, in a way that is independent of the neurons' identity or preferred stimulus. Such a suppression thus constitutes a pure form of stimulus specific adaptation. In particular, neurons do not reduce their responsiveness to stimuli that are not over-represented in the new environment. Thus, tuning curves are suppressed only near over-represented stimuli (and may in fact be enhanced near under-represented stimuli), leading to a repulsion of tuning curve peak from over-represented stimuli. This repulsion is a typical manifestation of stimulus specific adaptation; in V1, for example, orientation tuning curves display a repulsion from an over-represented orientation (Movshon and Lennie, 1979; Müller et al., 1999; Dragoi et al., 2000, 2002; Benucci et al., 2013).

Heretofore we have implicitly made a so-called ideal observer assumption by assuming that the internal generative model underlying the DDC or Bayes-ratio code is a veridical model of the stimulus distribution in the environment. We now provide a significant generalization of the above results to the more general and realistic case of a non-ideal-observer internal model. As we will see, in this case, the effects of adaptation on tuning curves are captured by both a stimulus-specific factor as well as a neuron-specific factor. For a non-ideal observer model, there will be a discrepancy between the environment's true stimulus distribution, $P^E(\boldsymbol{s})$, and the marginal stimulus distribution predicted by the internal generative model, *i.e.*,

$$P^I(\boldsymbol{s}) \equiv \int f(\boldsymbol{s}|\boldsymbol{z})\pi(\boldsymbol{z})d\boldsymbol{z}. \tag{31}$$

In App. B.10, we demonstrate that the Bayes-ratio tuning curves in this case are given by

$$h_a(\boldsymbol{s}) = \chi\frac{f(\boldsymbol{s}|\boldsymbol{z}_a)}{P^I(\boldsymbol{s})F_a}, \tag{32}$$

(instead of Eq. (27)) where

$$F_a = \int \frac{P^E(\boldsymbol{s})}{P^I(\boldsymbol{s})} f(\boldsymbol{s}|\boldsymbol{z}_a)d\boldsymbol{s}. \tag{33}$$

Note that, for ideal-observer internal models, $P^E = P^I$, in which case the integral yields $F_a = 1$ due to the normalization of $f(\boldsymbol{s}|\boldsymbol{z})$, and we recover Eq. (27). As we can see, the tuning curves neatly decompose into a "base tuning curve" $\chi f(\boldsymbol{s}|\boldsymbol{z}_a)$ multiplied by a stimulus-specific factor $P^I(\boldsymbol{s})^{-1}$ and a neuron-specific factor $F_a^{-1}$. Consequently, if we consider a pair of environments indexed by $\nu = 0, 1$, and assume the likelihood function of the internal model does not change between

environments, the transformation of the tuning curves due to adaptation will be given by Eq. (34),

$$h_a(\boldsymbol{s};\nu=1) = \frac{P^I(\boldsymbol{s};\nu=0)}{P^I(\boldsymbol{s};\nu=1)} \times h_a(\boldsymbol{s};\nu=0) \times \frac{F_a(\nu=0)}{F_a(\nu=1)}. \tag{34}$$

Thus, adaptation from environment $\nu = 0$ to environment $\nu = 1$ causes the tuning curves to transform via multiplication by a *stimulus specific factor* $P^I(\boldsymbol{s};\nu=0)/P^I(\boldsymbol{s};\nu=1)$ and a *neuron specific factor* $F_a(\nu=0)/F_a(\nu=1)$. Now suppose an environmental change results in an increase in the prevalence of an adaptor stimulus. As long as the internal model is a sufficiently good (if not perfect) model of $P^E(\boldsymbol{s})$, it should adapt its prior distribution, $\pi(\boldsymbol{z})$, in a way that results in an increase of its predicted stimulus distribution, $P^I(\boldsymbol{s})$, around the adaptor stimulus. This results in a stimulus specific factor that goes below 1 around the adaptor $\boldsymbol{s}$, thus causing suppression and repulsion of tuning curves away from the adaptor. Now consider a neuron whose preferred stimulus (roughly represented by $\boldsymbol{z}_a$) is close to the adaptor stimulus. We further make the assumption that the change in the internal model, in response to the environmental change, is sufficiently conservative such that it leads to a smaller increase in $P^I(\boldsymbol{s})$ around the adaptor, compared with the increase in the veridical $P^E(\boldsymbol{s})$ (this is a reasonable assumption especially for artificially extreme adaptation protocols used in experiments). In this case, we expect the ratio $P^E(\boldsymbol{s})/P^I(\boldsymbol{s})$ to increase in the support of $f(\boldsymbol{s}|\boldsymbol{z}_a)$, which (by our assumption about the preferred stimulus $\boldsymbol{z}_a$) is near the adaptor. According to Eq. (33) this will lead to an increase in $F_a$, and hence a suppressive neuron-specific factor $F_a(v=0)/F_a(v=1)$ for neurons with preferred stimulus near the adaptor. In accordance with this picture, Benucci et al. (2013) found that homeostatic and stimulus-specific adaptation in V1 can indeed be modelled via separable stimulus-specific and neuron-specific factors.

Note that the assumption of constancy of the likelihood across the two environments need not be framed as a veridical assumption about the objective environmental change, but rather as an assumption about the inductive biases of the internal generative model, according to which it tends to model changes in the observed stimulus distribution, $P^E(\boldsymbol{s})$, as primarily resulting from changes in the statistics of the latent causes in the environment (which the model captures in $\pi(\boldsymbol{z})$), rather than in the causal process itself, as modelled by $f(\boldsymbol{s}|\boldsymbol{z})$. As long as this is good enough an assumption to result in a $P^I(\boldsymbol{s})$ which changes in similar ways to $P^E(\boldsymbol{s})$ (and in particular exhibits an increase around adaptor stimuli), the conclusions reached above will hold.

In the case of DDC codes that use more general kernel functions, other than the delta functions underlying Bayes-ratio codes, we arrive at an analogous transformation, given by Eq. (138) in App. B.10. However, in this case, neuron-specific and stimulus-specific effects are mixed, and the wider the DDC kernel the stronger the mixing. Nevertheless, for kernels that are unimodal and sufficiently narrow, we expect an approximate factorisation of the effect of adaptation into a stimulus-specific and a neuron-specific suppression factor, as seen in Eq. (34).

We now apply this result to the experiments performed by Benucci et al. (2013). These experiments examined the effects of adaptation on orientation-tuned neurons in the cat primary visual cortex. Anaesthetised cats were shown a sequence of oriented gratings chosen randomly from a distribution that had an increased prevalence (3 to 5 times more likely) of one particular orientation (arbitrarily defined to be to 0°). A control group was exposed to a uniform distribution of gratings. After about 2 seconds (on the order of 50 stimulus presentations), V1 tuning curves had adapted, and exhibited both the suppressive and repulsive effects mentioned above.

To model these findings we constructed a homeostatic DDC model as follows, and fit it to the data from Benucci et al. (2013). We took the model's stimulus and latent variable spaces to be the circular orientation space $[-90, 90)$, and we took the model's generative likelihoods to be given by translation invariant functions $f(s|z) = f(s-z)$, proportional to the Gaussian distribution with standard deviation $\sigma_f$, normalised over the circle.[8] Likewise, we take the DDC kernel functions $\phi_a(z) = \phi(z - z_a)$ proportional to the Gaussian distribution with standard deviation $\sigma_\phi$.

The distributions of stimulus orientations used in the experiment is highly artificial, in the way they jump discontinuously near the adaptor orientation. The internal prior distribution employed by the brain is more likely to be smooth, reflecting the brain's inductive bias adapted to natural environments. Thus in our model we chose the internal prior to be a smooth distribution. The orientation distribution used in the experiment can be understood as a mixture of a uniform distribution, and a spike concentrated at the adaptor (Fig. 11A, blue curve). To obtain the smooth internal prior in our model, we replace the spike component of the experimental orientation distribution with a Gaussian distribution centred at the adaptor with standard deviation $\sigma_\pi$ (Fig. 11A, red curve). Thus, our model has only three free parameters, $\sigma_f, \sigma_\phi$ and $\sigma_\pi$, *i.e.*, the widths of the generative likelihood, kernel functions, and the internal orientation prior, respectively.

[8] More rigorously, we ought to use a wrapped (periodic) normal distribution. However, since the standard deviations are small compared to the length of the circle, the normalisation constant is approximately 1, and we can treat the density as a normal Gaussian.

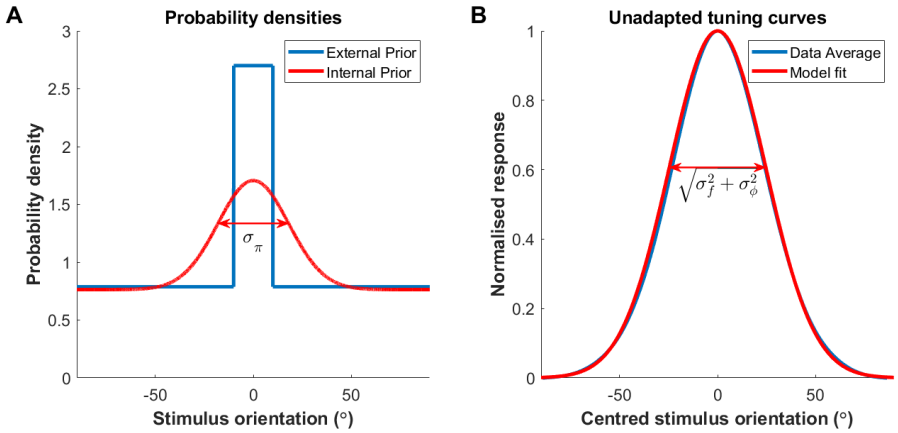


Figure 11: **Components of the model for stimulus-specific adaptation in V1.** (A) The distribution of stimulus orientations used in the experiments (Benucci et al., 2013), for an adaptor probability of 30% is shown in blue. We assume V1's internal model use continuous prior distributions and thus we used instead a mixture of the uniform distribution with a smooth Gaussian, with standard deviation $\sigma_\pi$, centered at the adaptor orientation (red curve). (B) The baseline tuning curves, *i.e.*, the tuning curves adapted to the uniform orientation distribution. The blue curve was obtained by averaging the experimentally measured tuning curves adapted to the uniform orientation distribution, after centering those curves at 0°. The model's baseline tuning curves (red) are given by the convolution of the Gaussian likelihood function, with width $\sigma_f$, of the postulated internal generative model, and the Gaussian DDC kernel, with width $\sigma_\phi$, and thus are themselves Gaussian with width $\sqrt{\sigma_f^2+\sigma_\phi^2}$. We fit this quantity to the experimentally observed baseline curves as described in Sec. 4.4.

Using the dataset from Benucci et al. (2013), we found the baseline tuning curves, *i.e.*, the tuning curves of the population adapted to the uniform orientation distribution, by assuming that neural response is a function only of the difference between the preferred orientation and stimulus orientation (see Sec. 4.4). We then calculate the changes in the neurons' preferred orientations (due to adaptation and the resulting tuning curve repulsion) in each of the experimental conditions (corresponding to different levels of over-representation of the adaptor stimulus). We then fit our model's three parameters to both the unadapted tuning curves (Fig. 11B) and the changes in preferred orientation (see Sec. 4.4 for further details). The best fit values for the model parameters were found to be $\sigma_\pi = 18°$, $\sigma_f = 14.2°$, and $\sigma_\phi = 20°$. The predictions of our model are compared to experimental data in Fig. 12. We found that the adaptive tuning curves within our model display suppression and repulsion away from the adaptor stimulus (Fig. 12A,D). In both the experimental data and our model, unadapted tuning curves exhibit a heightened average firing rate near the adaptor stimulus (Fig. 12B,E, blue curves). However, adaptation suppresses responses near the adaptor (Fig. 12A,D, red curves) to just the right degree to return the firing rates to the value before the environmental shift, leading to homeostasis of firing rates (as demonstrated by the uniformity of the red curves in Fig. 12B,E). Lastly, our model recapitulates the repulsion of preferred orientations found experimentally (Fig. 12C,F). Repulsion here is reflected by the change in preferred orientation having the same sign as the difference between the pre-adaptation preferred orientation and the adaptor orientation. As seen, in both the experiment and the model, repulsion is strongest for neurons with pre-adaptation preferred orientation about 30° from the adaptor.

In sum, the application of our homeostatic coding framework to a specific DDC model of Bayesian representations (which approximates what we have termed Bayes-ratio coding) explains quantitatively the empirical observations of both stimulus-specific and homeostatic adaptation in V1 (Benucci et al., 2013).

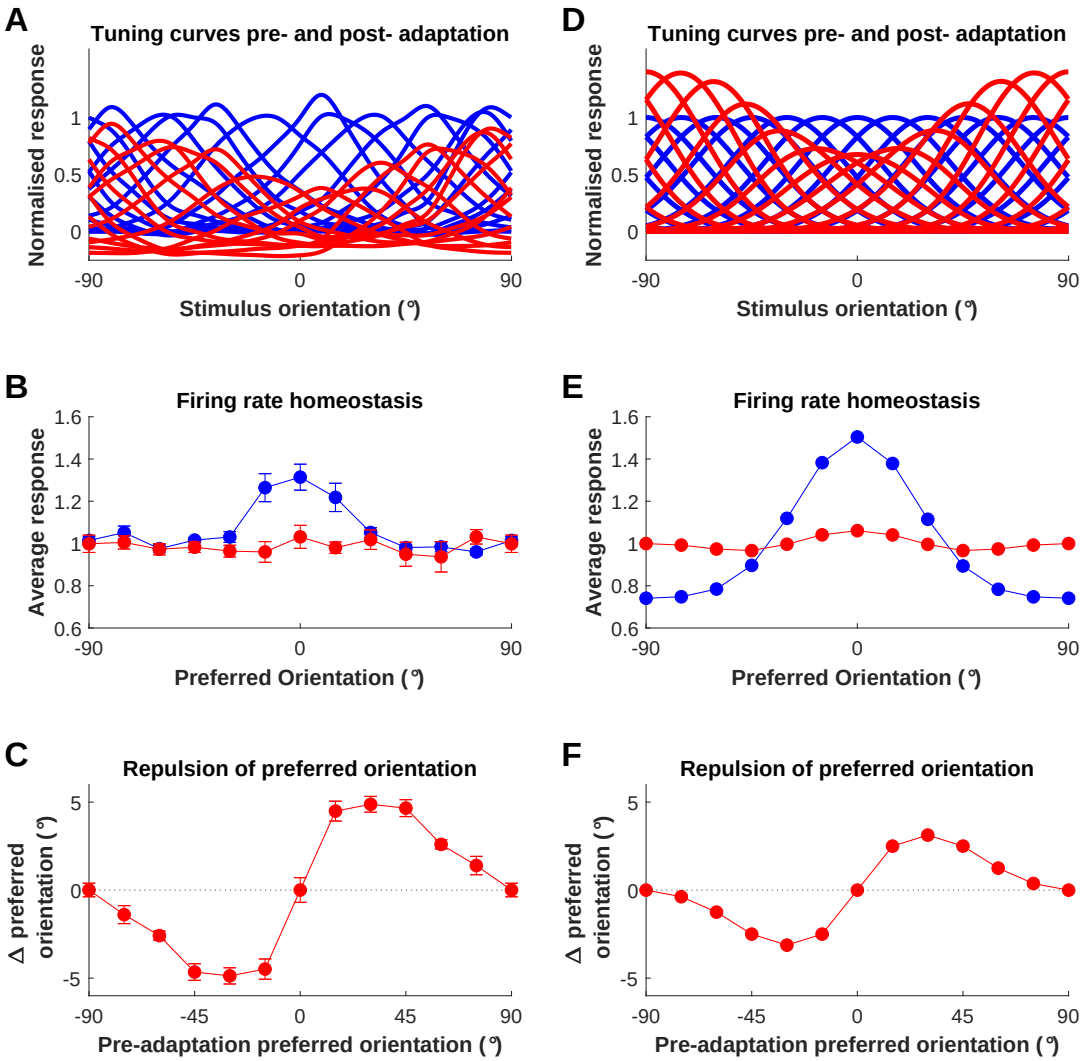


Figure 12: **A homeostatic DDC model accounts for the observations of stimulus-specific adaptation in V1 Benucci et al. (2013).** Panels A-C (left column) and E-F (right column) correspond to the experimental results and the predictions of our model, respectively. (A, D) The tuning curves for the adapted (red) and unadapted (blue) populations, averaged across experimental conditions. (B, E) The trial-averaged firing rates of the adapted (red) and unadapted (blue) populations exposed to the non-uniform stimulus ensemble. The trial-averaged population responses to the uniform stimulus ensemble were normalised to 1. (C, F) The repulsion of preferred orientations, obtained from the average tuning curves in panel A, as a function of the deviation of the neurons' preferred orientation from the adaptor orientation.

# 3 Discussion

We developed a theory of optimal gain modulation for combating noise in neural representations serving arbitrary computations. We demonstrated that — when mean neural firing rates are not too small; neurons are sufficiently selective and sparsely activated; and their responses form a high-dimensional geometry — the trade-off between coding fidelity and metabolic cost is optimised by gains that react to shifts in environmental stimulus statistics to yield firing rate homeostasis. Specifically, our theory predicts the homeostasis of the mean firing rate for clusters of similarly tuned neurons, while, at the same time, it accounts for the optimality of a broad distribution of firing rates within such clusters. By examining empirical estimates of parameters characterising cortical representations, we argued that the conditions necessary for the homeostatic adaptation to be optimal are expected to hold in the visual cortex and potentially other cortical areas. We further validated our approximation by demonstrating numerically that it performs well compared to the calculated optimal gains. Having developed a normative theory of neural homeostasis divorced from the computational aims of the neural population, we next showed how our theory can account for stimulus specific adaptation when coupled with distributed distributional codes (DDC). In particular, we focused on a special case of homeostatic DDC codes which we termed Bayes-ratio coding and showed that this model can account for stimulus specific adaptation effects empirically observed in the primary visual

cortex. In the following, we will situate our work within the broader context of efficient coding theory, Bayesian neural representations, and alternative theories of stimulus specific adaptation. We will also discuss the shortcomings of the present work and possible directions for extension.

Within efficient coding theory, our work is closely related to that of Ganguli and Simoncelli (2014), but differs from it in important ways. Ganguli and Simoncelli (2014) considered a population of tuning curves, defined on a one-dimensional stimulus space, which was homogeneous and translationally invariant up to a warping (*i.e.*, reparametrization) of the stimulus space and possible variations in gain. Their normative framework then optimised the neural gains as well as the stimulus warping (*i.e.*, the reparametrization mapping from the original stimulus parameter to one over which the tuning curve population is homogeneous) to maximise a similar objective function to ours. Specifically, their objective employed the Fisher Information lower bound on the mutual information (Brunel and Nadal, 1998), as opposed to the Gaussian upper bound we used; our choice of the Gaussian upper bound was motivated by increased analytic tractability in the multidimensional stimulus case covered by our theory. In the case of unimodal (and one-dimensional) tuning curves, they showed that the optimal warping amounts to a concentration of tuning curves around stimulus values that are over-represented in the environment. On the other hand, in their setting, the optimal gains stay constant and do not adapt. Nevertheless, their solution also exhibits homeostasis of single-neuron firing rates.

Our framework extends and generalises that of Ganguli and Simoncelli (2014) in different ways. In their framework, the tuning curves are one-dimensional, and rigid constraints (most importantly, either unimodal or sigmoidal tuning curves, and homogeneity up to stimulus reparametrization) are placed on the shape and arrangement of the tuning curve population. In contrast, our framework is applicable to tuning curves with heterogeneous (in particular multi-modal) shapes and arbitrary arrangement over a high-dimensional stimulus space; our homeostatic solution only requires sufficiently high signal-to-noise ratio (which is also the condition for the tightness of the mutual information lower-bound employed in Ganguli and Simoncelli (2014)). On the other hand, we only optimised the neural gains, and not the shapes or arrangement of the tuning curves, as we assumed the latter are determined by the computational goals of the circuit and not by the aim of optimally combating coding noise; in this way, our framework is agnostic to the computational goals of the population. Additionally, Ganguli and Simoncelli (2014) only considered the case of uncorrelated Poissonian noise. In contrast, our framework can handle correlated noise with general (non-Poissonian) power-law scaling. Furthermore, we analytically calculated first-order corrections to the homeostatic solution that arises in the high signal-to-noise ratio limit.

As noted above, the theory of Ganguli and Simoncelli (2014) predicts that tuning curves optimised to provide an efficient code under a stimulus distribution with over-represented stimuli cluster around those stimuli. This has been used to explain asymmetries in neural selectivities, such as the over-abundance of V1 neurons with preferred orientations near cardinal orientations, which are more abundant in natural scenes; or more generally, the aggregation of tuning curves around stimuli that are more prevalent in the natural environment. Many of these asymmetries have likely resulted from adaptation at very long (*e.g.*, ontogenetic, or possibly even evolutionary) timescales; however, long-term exposure on the order of few minutes to adaptor stimuli has also been found to result in the attraction of tuning curves towards adaptor stimuli (Ghisovan et al., 2009). This is in contrast to the repulsive effects seen in short-term adaptation (Movshon and Lennie, 1979; Müller et al., 1999; Dragoi et al., 2000, 2002; Benucci et al., 2013). It is likely that the effects of long-term exposure are the results of mechanisms which operate in parallel to those underlying short-term adaptation, with both types co-existing at different timescales. We can therefore interpret Ganguli and Simoncelli (2014) as a model of adaptation to natural stimulus statistics at long timescales. Our results obtained here are applicable to adaptation on shorter timescales (Müller et al., 1999; Dragoi et al., 2000, 2002; Benucci et al., 2013). In particular, when married with Bayesian theories of neural representation, our framework predicts repulsive effects around an adaptor stimulus, as shown in Sec. 2.10.

Westrick et al. (2016) also propose a model of the experimental findings of Benucci et al. (2013). Their model (which is not explicitly normative or based on efficient coding) uses divisive normalisation (Carandini and Heeger, 2012) with adaptive weights to achieve homeostasis and stimulus specific repulsion. As discussed above in Sec. 2.9, Bayes-ratio coding (a special case of homeostatic DDCs, which we showed can account for the findings of Benucci et al. (2013)) can be accomplished by such a divisive normalisation scheme. Our framework therefore yields a normative interpretation of the model of Westrick et al. (2016), and links divisive normalisation with Bayesian representations and efficient coding theory. In particular, our theory provides an interpretation of the normalization weights of the divisive normalization model as the internal prior over latent causes of sensory inputs. This prior should naturally adapt as stimulus statistics change across environments. Additionally, the feedforward inputs to the divisive normalisation model are interpreted as the likelihood of stimuli given latent causes.

Snow et al. (2016) developed two normative models of temporal adaptation effects in V1, based on statistics of dynamic natural visual scenes (natural movies). These models are based on generative models of natural movies in the class of mixtures of Gaussian scale mixture (MGSM) distributions. In MGSMs the outputs of linear oriented filters applied to video frames are assumed to be Gaussian variables, multiplied by positive scale variables. These scale variables can be shared within pools of filter outputs at different times (video frames) and orientations. The two models of Snow et al. (2016) differ according to which pools of filter outputs are able to share scale variables. In both models, V1 neural responses are assumed to represent the inferred Gaussian latent variables of the corresponding MGSM. While each model was able to account for some of the findings of Benucci et al. (2013), neither one could account for all. In particular, each of

their two models could only account for either stimulus-specific or neuron-specific adaptation factors found by Benucci et al. (2013). In particular, only the model that accounted for neuron-specific adaptation was able to produce firing rate homeostasis. As we showed in Sec. 2.10, our framework, which combines homeostatic gain modulation with the DDC theory of representation, robustly accounts both for stimulus-specific adaptation and firing rate homeostasis. Furthermore, when there is a discrepancy between the internal model's stimulus prior and the true environmental stimulus distribution, it additionally accounts for a neuron-specific adaptation factor as well (see Eq. (34)).

Our approach has relied on a number of assumptions and simplifications; relaxing or generalizing these assumptions provide opportunities for future research. Firstly, our toy model (see Sec. 2.3) imposed a cluster structure on the neural tuning curves, whereby neurons within a cluster are similarly tuned, while neurons in different clusters have distinct tuning. We found that total firings rates are equalized across clusters (Sec. 2.2), but the optimal rates of individual neurons within clusters span a broad range (Sec. 2.3). This results from the fact that the efficient coding objective function depends sharply on total cluster firing rates, but changes only slightly when the partition of the cluster firing rate among the cluster's neurons changes. In reality, tuning similarity has a more graded distribution. As the distinction between clusters with different tuning (or similarity of tuning within clusters) weakens, higher order terms in our expansion of the objective Eq. (11) cannot be neglected, and hence the justification for dividing the problem into two separate, cluster-level vs. within-cluster, problems breaks down. However, arguably, in a more realistic model with a more graded variation in neural signal correlations, the optimization landscape will still contain shallow directions (*i.e.*, directions over which the objective changes slowly), corresponding to partitions of firing among similarly tuned neurons, and non-shallow directions corresponding to total firings of such groups of neurons; if so, we would expect a similar solution to emerge, whereby similarly tuned neurons display heterogeneous firing rates, with the total firing rate of such a group of neurons displaying approximate homeostasis.

Secondly, we showed analytically that, at the level of cluster responses, homeostasis emerges universally in the high signal-to-noise regime. In general, firing rate homeostasis is no longer optimal in the low signal-to-noise ratio limit, as correction terms to our solutions become large. However, as we argued in Sec. 2.5, based on empirical values we expect the signal-to-noise ratio within cortex to be sufficiently high for our results to hold. Generally, within efficient coding theory (except for simple cases such as linear Gaussian models), the regime of arbitrary signal-to-noise ratio is analytically intractable. Many approaches are therefore limited to the regime of high signal-to-noise ratio. For example, the Fisher Information Lower Bound used frequently within efficient coding theory breaks down outside of this regime (Yarrow et al., 2012; Bethge et al., 2002; Wei and Stocker, 2016). Thus, inevitably, exploring the low signal-to-noise ratio regime would require numerical simulation. However, such numerical simulations require concrete models of signal and noise structure. The manner in which optimal coding deviates from homeostatic coding will not be universal across these models, but depend on each model's specific details. This limits our ability to draw general conclusions from numerical simulations, which is why we have chosen not to pursue that strategy extensively here.

Thirdly, our analysis here considered a specific class of noise models. However, these noise models include those of particular biological relevance, such as information-limiting noise correlations (Moreno-Bote et al., 2014; Kanitscheider et al., 2015), signatures of which have been found in the cortex (Rumyantsev et al., 2020), and power-law variability (Goris et al., 2014) (see Sec. 2.6). We note that our homeostatic solution is particularly well suited to the case of information-limiting noise correlations, performing optimally in that case (see App. B.5). The core ideas of our framework and derivation could potentially be applied to an even wider class of noise models.

Finally, up to section 2.8, we make no assumptions about what the neural population is attempting to represent via the representational curves. We then apply our framework to a specific theory of Bayesian encoding (namely DDCs), and develop the new idea of Bayes-ratio coding. There is therefore an opportunity to apply our framework to other theories of neural representation, and in particular to alternative theories of Bayesian representations, such as Probabilistic Population Codes (PPCs) (Beck et al., 2007).

In summary, we showed that homeostatic coding can arise from optimal gain modulation for fighting corruption by noise in neural representations. Based on this coding scheme, we derived a novel implementation of probabilistic inference in neural populations, known as Bayes-ratio coding, which can be achieved by divisive normalisation with adaptive weights to invert generative probabilistic models in an adaptive manner. This coding scheme accounts for adaptation effects that are otherwise hard to explain exclusively based on efficient coding theory. These contributions provide important connections between Bayesian representation, efficient coding theory, and neural adaptation.

# 4 Methods

## 4.1 Specification of environments for simulations

To generate the $K \times K$ signal correlation matrix, $\rho(\nu)$, for the different environments in our numerical simulations, in Sec. 2.2, we first generate a covariance matrix $\Sigma(\nu)$, and let $\rho(\nu)$ be the corresponding correlation matrix. The covariance matrix $\Sigma(\nu)$ is in turn constructed such that it has (1) the same eigenspectrum for all $\nu$, with eigenvalues that decrease inversely with rank, as observed in V1 (Stringer et al., 2019), and (2) an eigenbasis that varies smoothly with $\nu$. Thus we

defined $\Sigma(\nu) = U(\nu)\Lambda_1 U(\nu)^{\mathrm{T}}$, where $\Lambda_1 = \mathrm{diag}(1, 1/2, 1/3, \ldots, 1/K)$ and $U(\nu)$ is an orthogonal matrix that was generated as follows. We randomly and independently sample two $K \times K$ random iid Gaussian matrices $R_0, R_1 \sim \mathcal{N}_{K\times K}(0, I_{K\times K})$ and obtain symmetric matrices $S_0 = R_0 + R_0^{\mathrm{T}}$ and $S_1 = R_1 + R_1^{\mathrm{T}}$. As is well known, the eigen-basis (represented by an orthogonal matrix) of such a random Gaussian matrix is distributed uniformly (*i.e.*, according to the corresponding Haar measure) over the orthogonal group $O(K)$. For the extreme environments, $\nu = 0$ and 1, we let $U(\nu)$ be the matrix of eigenvectors of $S_0$ and $S_0$, when ordered according to eigenvalue rank. For $0 < \nu < 1$, we let $U(\nu)$ be the matrix of the eigenvectors (again ordered according to eigenvalue rank) of the interpolated symmetric matrix $S(\nu) = (1-\nu)S_0 + \nu S_1$. Thus, for the extreme environments the eigenbases of the covariance matrices $\Sigma(0)$ and $\Sigma(1)$ are sampled independently and uniformly (*i.e.*, from the Haar measure), and the eigenbases of $\Sigma(\nu)$ for intermediate environments (*i.e.*, for $0 < \nu < 1$) smoothly interpolate between these.

In the case of simulations for aligned noise, we also generate a noise correlation matrix $W(\nu)$ which has an approximately $1/n^\gamma$ spectrum. Ideally, $W$ and $\rho$ would have the same eigen-basis. However, this is impossible, since $W$ and $\rho$ are both correlation matrices. Instead, we generate $W(\nu)$ by normalising the positive definite matrix $U(\nu)\Lambda_\gamma U(\nu)^{\mathrm{T}}$ where $\Lambda_\gamma = \mathrm{diag}(1, 1/2^\gamma, 1/3^\gamma, \ldots, 1/K^\gamma)$

## 4.2 Within-cluster optimisation problem

In Sec. 2.3 we showed that in our clustered population model, the firing rates of individual neurons within a cluster optimise the objective Eq. (12). For a cluster of $k$ neurons, this problem is defined by the $k \times k$ covariance matrix $\mathrm{Cov}(\Delta\hat{\boldsymbol{\Omega}})$. In our simulations, we generated this matrix as the the Gram matrix of $k$ random $D$-dimensional vectors with isotropic Gaussians distributions, *i.e.*, $\mathrm{Cov}(\Delta\hat{\Omega}_i, \Delta\hat{\Omega}_j) = \delta\boldsymbol{w}_i^{\mathrm{T}}\delta\boldsymbol{w}_j$, where $\delta\boldsymbol{w}_i \sim \mathcal{N}(0, I_D)$. (A justification for our construction of the covariance matrix, and an interpretation of the dimensionality $D$, can be found in App. B.3.5.) Additionally, we fixed the cluster rate to $\mu = k \times 5$Hz, such that the average firing rate per single neuron is 5Hz, and fixed $\sigma^2 = 1$ (corresponding to a Fano factor of 1). For various values of $k$ and $D$, we used the above method to generate $10,000$ samples of the covariance matrix $\mathrm{Cov}(\Delta\Omega)$. For each such sample, we then solved the problem (12–13) numerically using the `cvxpy` package in Python (Diamond and Boyd, 2016). Neurons with firing rates less than $10^{-4}$Hz had their firing rate set to 0Hz. We aggregated the set of nonzero rates obtained from each of these optimisations to create the histograms in Fig. 8.

## 4.3 Estimating noise scaling parameters from biological data

In this section we discuss biological estimates for the parameters of the general form of mean-variance relationship of spike count cluster responses that we have assumed in our noise models,

$$\text{variance} = \tilde{\sigma}^2 \text{mean}^{2-\beta}.$$

Four papers estimate this scaled power-law relationship: the findings of those papers are summarised in the first five columns of Table 1.

| Paper | Area | $\tilde{\sigma}^2$ | $\beta$ | Time interval (s) | $\sigma^2/(\beta\mu)^\beta$ |
|---|---|---|---|---|---|
| Shadlen and Newsome (1998) | MT (rhesus monkey) | 0.4 | 0.7 | 0.5 | 0.05 |
| Gershon et al. (1998) | V1 (rhesus monkey) | 1.30 | 0.57 | 0.27 | 0.15 |
| Gershon et al. (1998) | V1 (rhesus monkey) | 1.82 | 0.82 | 0.27 | 0.19 |
| Gershon et al. (1998) | IT (rhesus monkey) | 1.36 | 1.18 | 0.3 | 0.13 |
| Moshitch and Nelken (2014) | Auditory cortex (cat) | 1.37 | 0.88 | 0.05 | 0.12 |
| Koyama (2015) | MT (rhesus monkey) | 4.57 | 1.43 | 0.5 | 0.31 |

Table 1: Table summarising empirical findings from four papers which estimate a scaled power-law relationship relationship between mean and variance of cortical spike count responses. To find the values in the last column we adjust the reported $\tilde{\sigma}^2$ to account for different coding intervals and cluster sizes used in each experiment (see Eq. (35)), and assume $\beta\mu = 10$.

Note that the time interval of these papers differs from the our assumed coding interval of $0.1s$. We therefore must adjust the reported values to be for an interval of this size. We do this by assuming temporal homogeneity over the timescales of interest; specifically we assume that mean and variance are linear in the time interval. We must also adjust the data to account for the fact that our responses are for clusters of $k = 20$ similarly tuned neurons. We do this by assuming uncorrelated response noise within each cluster, making mean and variance both linear functions of cluster size. Our adjustments do not affect the value of $\beta$, but do change the value of the pre-factor $\tilde{\sigma}^2$. Specifically, we perform the transformation:

$$\sigma^2 = \tilde{\sigma}^2 \left(\frac{\delta t}{k\Delta t}\right)^{1-\beta} \tag{35}$$

where $\delta t$ is the reported time interval, and $\Delta t$ is our coding interval, $\Delta t = 0.1$s. Fig. 13 shows the adjusted values on a scatter plot. Using the adjusted values, $\sigma^2$, and assuming that $\mu\beta = 10$ (to maintain a firing rate of 5Hz per neuron), we can find the value of $\sigma^2/(\beta\mu)^\beta$. These are reported in the furthest right column of Table 1.

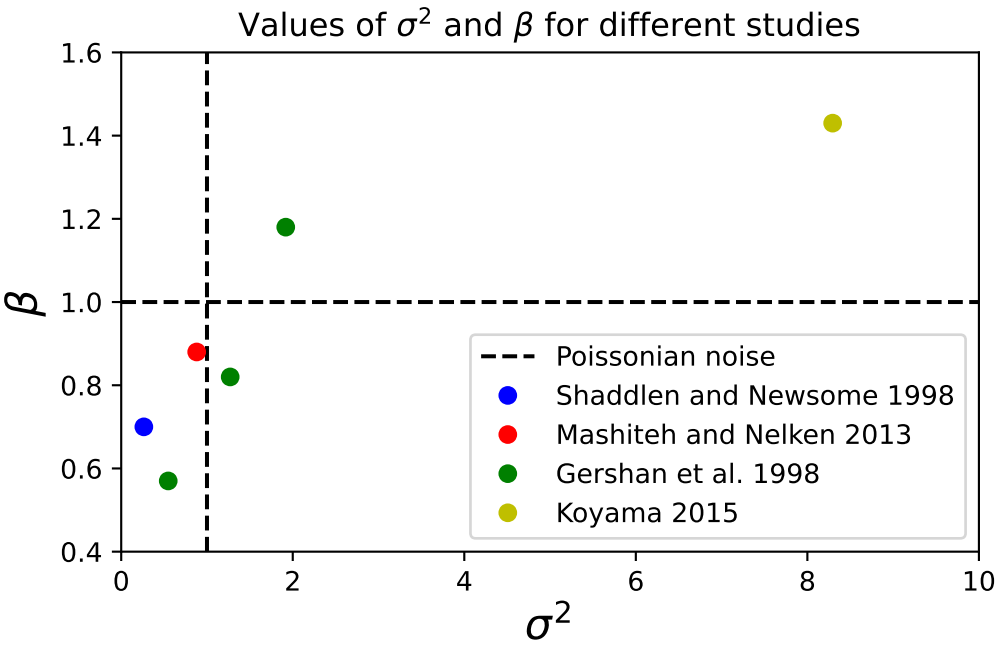


Figure 13: Empirically estimated values of the noise scaling parameter $\beta$ and the base noise-level $\sigma^2$, after adjusting for coding interval duration and cluster size using the formula Eq. (35). The dashed lines show the values for Poissonian noise ($\sigma^2 = 1$, $\beta = 1$).

### 4.4 Fitting a homeostatic DDC to Benucci *et al.* 2013

The data we obtained from Benucci et al. (2013) comprised 11 datasets. In each dataset, the neural population was clustered into 12 groups based on preferred orientation. These neural populations were then exposed (in different contexts) to a distribution of 6 to 12 oriented gratings. This distribution was either uniform or biased; in the latter case, one particular grating (arbitrarily defined as 0°) had higher prevalence, by either $30\%, 35\%, 40\%$, or $50\%$. We discarded all datasets with a 50% prevalence, since (Benucci et al., 2013) report that homeostasis was not obtained in this case, and they similarly discarded it from their main analysis. After exposure to the distribution of gratings, the responses (*i.e.*, spike count) of each cluster to a test set of 20 oriented gratings were then measured. Following Benucci et al. (2013) (see that reference for details), we normalized responses via an affine transform so that after exposure to the uniform stimulus ensemble, the normalised response ranged from 0 to 1.

For each dataset, we obtained the average tuning curve in the uniform ensemble context (Fig. 11B, blue) we used a cubic spline to interpolate and up-sample all the population tuning curves in that context. These were then centered at 0° and averaged across populations with different preferred orientations, yielding a single average tuning curve. To fit the resultant curve with a Gaussian (Fig. 11B, red), the width of this curve at height $e^{-2}$ was divided by 4 to obtain the standard deviation of the Gaussian. In subsequent fitting of the model, $\sqrt{\sigma_f^2 + \sigma_\phi^2}$ was constrained to equal this standard deviation. This reduces the number of free parameters to 2, which we choose to be $\sigma_\pi$ and $\sigma_f$. We found the best value of these parameters by fitting the curves for repulsion of tuning curves (adaptation-induced change in preferred orientation vs. pre-adaptation preferred orientation – Fig. 12C and F) between the model and experimental data. Specifically, we performed a grid search over values of $\sigma_\pi$ and $\sigma_f$ and chose the pair of values which minimised the sum of the absolute value of the difference between the empirical and model curves.

We obtained the empirical repulsion curves (Fig. 12C) as follows. For each preferred orientation cluster and context (*i.e.*, exposure to a uniform or biased ensemble), a smoothing kernel was applied before using cubic interpolation to generate up-sampled tuning curves. The smoothing kernel was applied to ensure that none of the tuning curves were multimodal (multimodality was a minor effect — reasonably attributable to finite-trial averaging noise — but could have nevertheless introduced relatively large noise in the estimated preferred orientations). The argmax of these tuning curves was found to give the preferred orientation. The preferred orientation was then compared across conditions (uniform ensemble vs biased ensemble) to give a change in preferred orientation. These were then averaged across contexts with the same adaptor

probability to obtain the repulsion curve for each adaptor probability.

## 5 Acknowledgements

We thank Máté Lengyel for helpful discussions and comments on the manuscript. EY was supported by the UKRI Engineering and Physical Sciences Research Council Doctoral Training Program grant EP/T517847/1. YA was supported by UKRI Biotechnology and Biological Sciences Research Council research grant BB/X013235/1.

## A Extended data figures

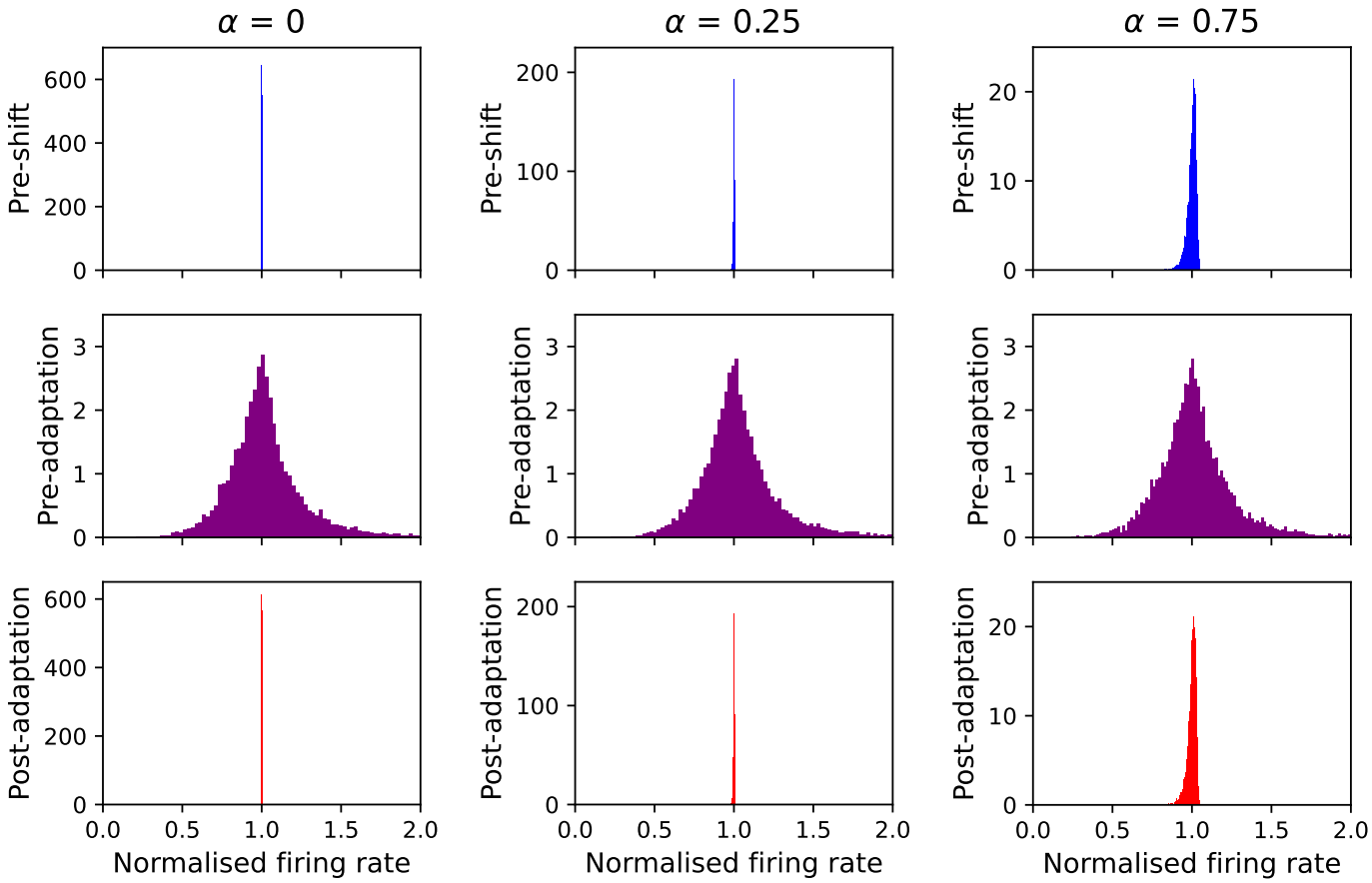


Figure 14: Distribution of average firing rates before and after a discrete environmental shift for the uncorrelated power-law noise subfamily. Each panel has the same format as in Fig. 5C, but for different values of the noise scaling parameter $\alpha$. The case $\alpha = 0.5$ is shown in Fig. 5C.

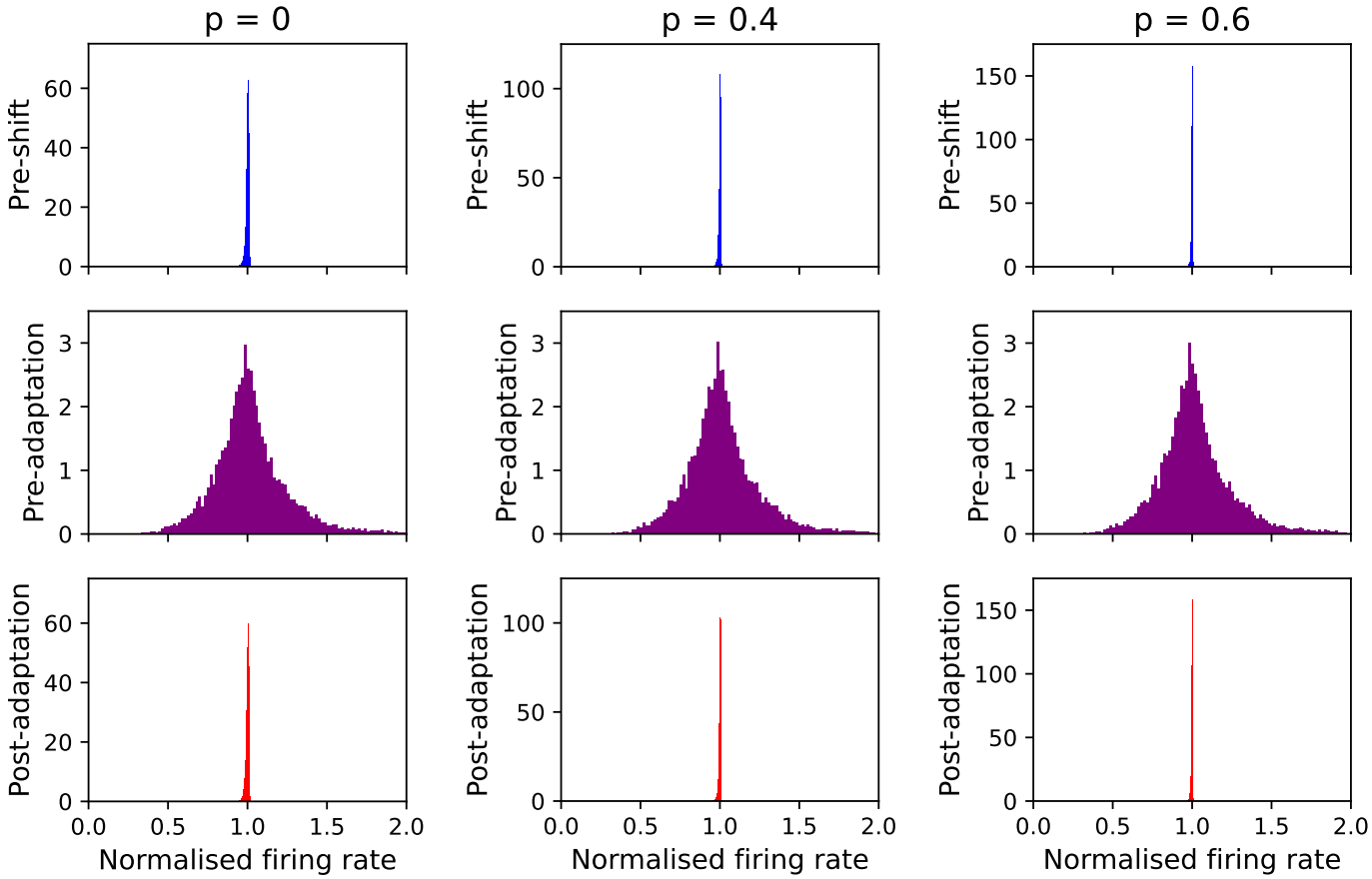


Figure 15: Distribution of average firing rates before and after a discrete environmental shift for the Poissonian noise subfamily with uniform noise correlations. Each panel has the same format as in Fig. 6C, but for different values of the noise correlation coefficient $p$. The case $p = 0.2$ is shown in Fig. 6C.

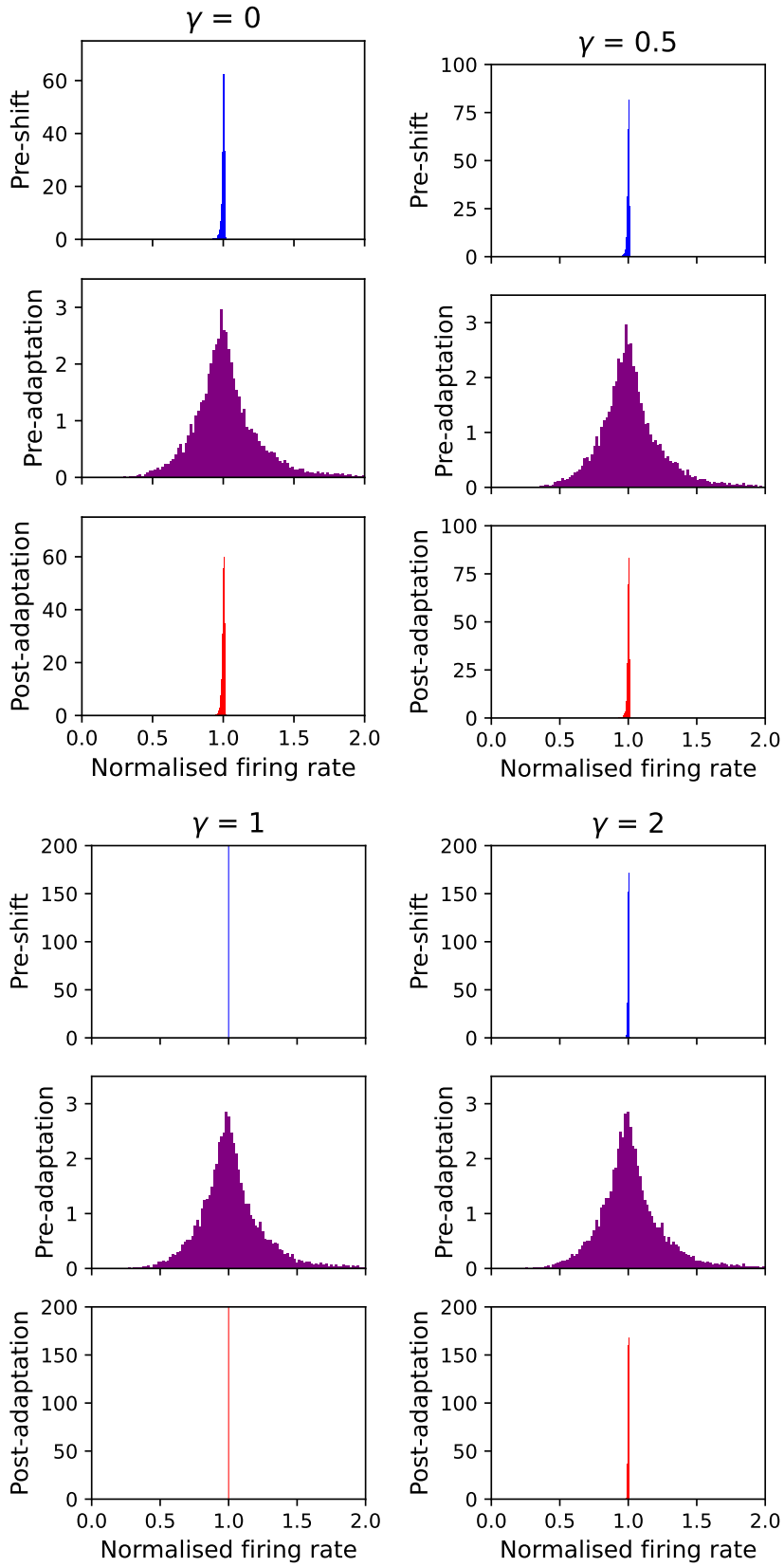


Figure 16: Distribution of average firing rates before and after a discrete environmental shift for the aligned Poissonian noise subfamily. Each panel has the same format as in Fig. 7C, but for different values of the noise spectrum decay parameter $\gamma$. The case $\gamma = 1.5$ is shown in Fig. 7C.

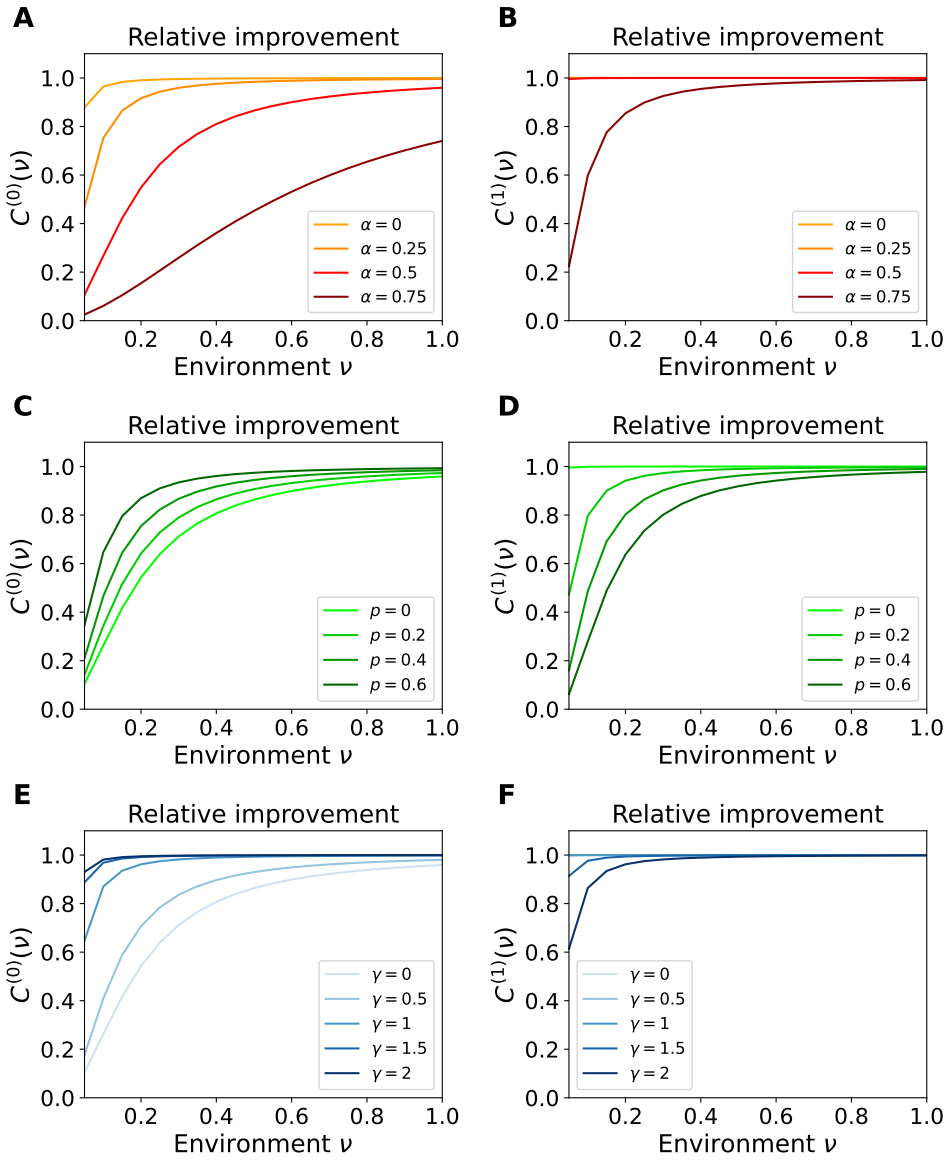


Figure 17: **Accuracy of homeostatic approximation to optimal gains.** The plots in the panels A, C, and E show the relative improvement in the efficient coding objective for the zeroth-order approximation for the optimal gains $\boldsymbol{g}^{(0)}$, Eq. (16), for the three noise sub-families; see the caption of Fig. 9 for a description of panels B, D, and F there, respectively. Plots in panels B, D, and F here similarly show the relative improvement in the efficient coding objective for the first-order approximation $\boldsymbol{g}^{(1)}$, Eq. (17), for the three noise sub-families.

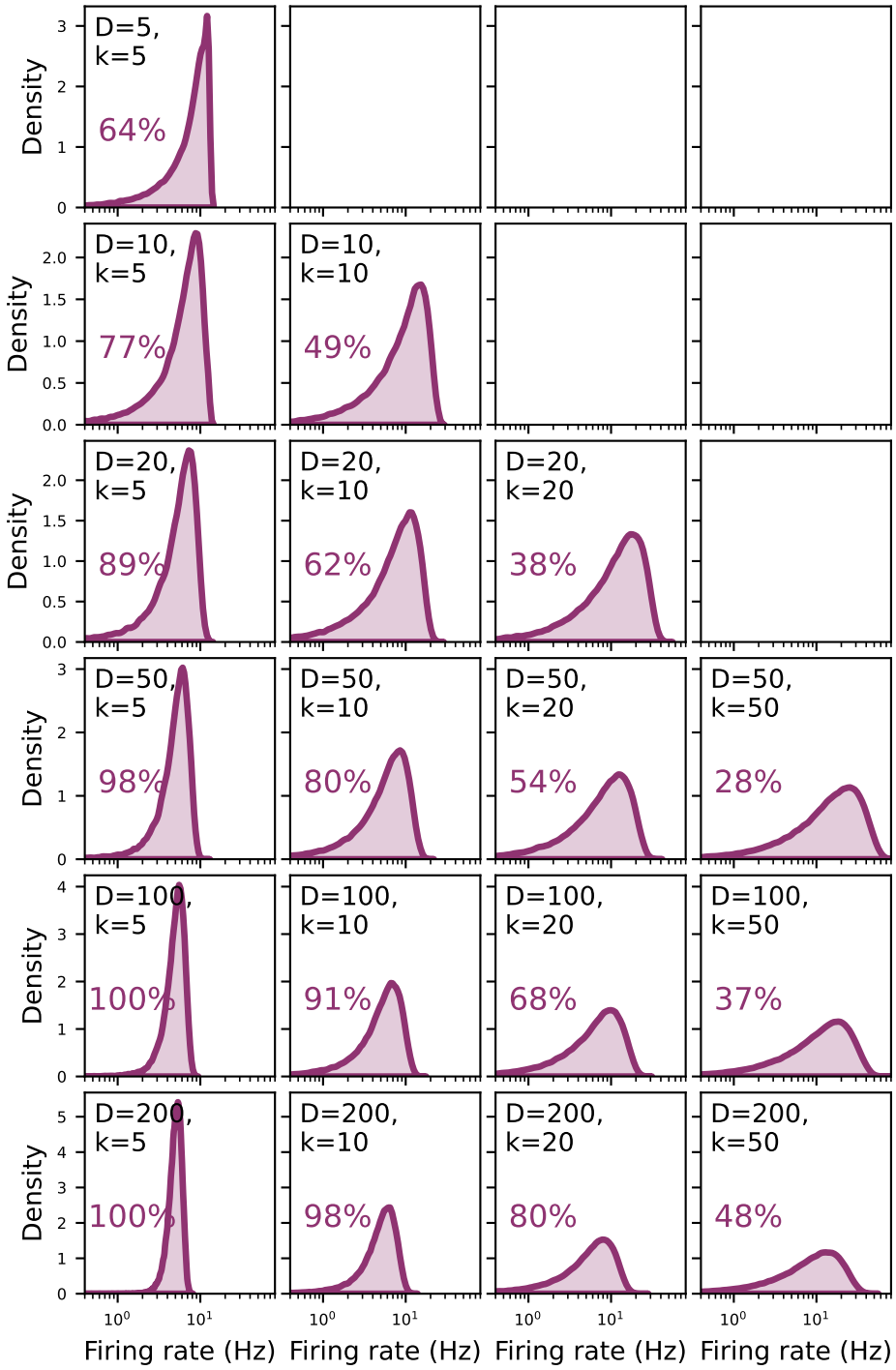


Figure 18: **Systematic exploration of how the firing rate distribution of individual neurons varies as the statistical properties of cluster responses change.** The panels here have the same description as those in Fig. 8, but show the single-neuron firing rate distributions for different choices of cluster size $k$ and effective tuning curve dimension, $D$. (Note that the upper-right panels corresponding to cases in which $D < k$ are left empty; this is because in such cases the covariance matrix is singular, and therefore the quadratic program does not have a unique solution. Such cases are also less relevant biologically given realistic estimates for the size of the similarly tuned neurons vs. the effective dimensionality of tuning curves, noting that the nominal dimension of the tuning curve function space is infinite.)

# B Supplementary Information: mathematical derivations

## B.1 Analytic expression for $\mathcal{L}$

In this appendix we derive an analytic expression for the upper bound objective $\mathcal{L}$. We adopt the noise model Eq. (5), repeated here for convenience:

$$\boldsymbol{n}|\boldsymbol{s} \sim \mathcal{N}\left(\boldsymbol{h}(\boldsymbol{s}), \sigma^2 h(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) h(\boldsymbol{s})^\alpha\right), \tag{36}$$

where $\Sigma(\boldsymbol{s})$ is the stimulus-dependent noise correlation matrix, and $0 < \alpha < 2$ is a scaling parameter. Here we have made use of our convention that the non-bold version of a vector (without indices) denotes the diagonal matrix formed from that vector, so that, for example, $h(\boldsymbol{s})$ is the diagonal matrix with $aa$-th entry equal to $h_a(\boldsymbol{s})$.

We decompose the mutual information between the spike counts and the stimulus as

$$I(\boldsymbol{n}; \boldsymbol{s}) = H[\boldsymbol{n}] - H[\boldsymbol{n}|\boldsymbol{s}]$$

and then obtain $\mathcal{L}$ by replacing the term $H[\boldsymbol{n}]$ in $\mathcal{L}^0 = 2I(\boldsymbol{n}; \boldsymbol{s}) - \mu^{-1}\sum_{a=1}^K \mathbb{E}[n_a]$ with the entropy of a Gaussian with equal covariance to $\boldsymbol{n}$. Doing so gives us the expression

$$\mathcal{L} = \ln\det(2\pi e \mathrm{Cov}(\boldsymbol{n})) - 2H[\boldsymbol{n}|\boldsymbol{s}] - \mu^{-1}\sum_{a=1}^K \mathbb{E}[n_a] \tag{37}$$

We consider $\mathcal{L}$ as a function of the rates $r_a = \omega_a g_a$ in order to keep the derivation clean.

We start by deriving an expression for the noise entropy $H[\boldsymbol{n}|\boldsymbol{s}]$. We use $\doteq$ to denote equality up to additive pure constants or terms independent of $r$ and $g$. We have

$$\begin{aligned}
2H[\boldsymbol{n}|\boldsymbol{s}] &= \int 2H\left[\mathcal{N}\left(\boldsymbol{h}(\boldsymbol{s}), \sigma^2 h(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) h(\boldsymbol{s})^\alpha\right)\right] P(\boldsymbol{s}) d\boldsymbol{s} \\
&\doteq \int \ln\det\left(\sigma^2 h(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) h(\boldsymbol{s})^\alpha\right) P(\boldsymbol{s}) d\boldsymbol{s} \\
&= \ln\det\left(r^{2\alpha}\right) + \int \ln\det\left(\omega^{-\alpha}\Omega(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) \Omega(\boldsymbol{s})^\alpha \omega^{-\alpha}\right) P(\boldsymbol{s}) d\boldsymbol{s}
\end{aligned}$$

(where in deriving the third line we used $h = g\Omega = r\omega^{-1}\Omega$), hence

$$2H[\boldsymbol{n}|\boldsymbol{s}] \doteq \ln\det\left(r^{2\alpha}\right). \tag{38}$$

Next, we find an expression for $\mathrm{Cov}(\boldsymbol{n})$, using the orthogonal decomposition $\mathrm{Cov}(\boldsymbol{n}) = \mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}])$. For the first term have

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s}) &= \sigma^2 h(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) h(\boldsymbol{s})^\alpha \\
&= \sigma^2 r^\alpha \left(\frac{h(\boldsymbol{s})}{\mathbb{E}[h(\boldsymbol{s})]}\right)^\alpha \Sigma(\boldsymbol{s}) \left(\frac{h(\boldsymbol{s})}{\mathbb{E}[h(\boldsymbol{s})]}\right)^\alpha r^\alpha
\end{aligned}$$

where we used $r = \mathbb{E}[h(\boldsymbol{s})]$. Thus

$$\mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})] = \sigma^2 r^\alpha W r^\alpha, \tag{39}$$

where we have defined (c.f. Eq. (7) of main text) the normalized stimulus-averaged noise covariance matrix $W$ as

$$W_{ab} = \mathbb{E}\left[\left(\frac{h_a(\boldsymbol{s})}{\mathbb{E}[h_a(\boldsymbol{s})]}\right)^\alpha \Sigma_{ab}(\boldsymbol{s}) \left(\frac{h_b(\boldsymbol{s})}{\mathbb{E}[h_b(\boldsymbol{s})]}\right)^\alpha\right] = \mathbb{E}\left[\left(\frac{\Omega_a(\boldsymbol{s})}{\omega_a}\right)^\alpha \Sigma_{ab}(\boldsymbol{s}) \left(\frac{\Omega_b(\boldsymbol{s})}{\omega_b}\right)^\alpha\right].$$

Next, we find an expression for $\mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}])$. Using $\boldsymbol{h}(\boldsymbol{s}) = \mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]$, we find

$$\begin{aligned}
\mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) &= \mathrm{Cov}\left(\boldsymbol{h}(\boldsymbol{s})\right) \\
&= \mathbb{E}[h(\boldsymbol{s})] \frac{\sqrt{\mathrm{Var}(h(\boldsymbol{s}))}}{\mathbb{E}[h(\boldsymbol{s})]} \left(\frac{1}{\sqrt{\mathrm{Var}(h(\boldsymbol{s}))}} \mathrm{Cov}(\boldsymbol{h}(\boldsymbol{s})) \frac{1}{\sqrt{\mathrm{Var}(h(\boldsymbol{s}))}}\right) \frac{\sqrt{\mathrm{Var}(h(\boldsymbol{s}))}}{\mathbb{E}[h(\boldsymbol{s})]} \mathbb{E}[h(\boldsymbol{s})],
\end{aligned}$$

or

$$\mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) = r\, P\, r, \tag{40}$$

where we defined

$$P \equiv \mathrm{CV}\rho\mathrm{CV}, \tag{41}$$

and introduced the coefficients of variation (CV) of trial-averaged responses, given by $h(\boldsymbol{s})$ (or equivalently, the CV's of the representational curves)

$$\mathrm{CV}_a := \frac{\sqrt{\mathrm{Var}(\Omega_a(\boldsymbol{s}))}}{\mathbb{E}[\Omega_a(\boldsymbol{s})]} = \frac{\sqrt{\mathrm{Var}(h_a(\boldsymbol{s}))}}{\mathbb{E}[h_a(\boldsymbol{s})]}, \tag{42}$$

and the Pearson's correlation matrix of $h$ (or equivalently of the $\Omega$'s), defined by

$$\rho_{ab} := \frac{\mathrm{Cov}[\Omega_a(\boldsymbol{s}), \Omega_b(\boldsymbol{s})]}{\sqrt{\mathrm{Var}[\Omega_a(\boldsymbol{s})]\ \mathrm{Var}[\Omega_b(\boldsymbol{s})]}} = \frac{\mathrm{Cov}[h_a(\boldsymbol{s}), h_b(\boldsymbol{s})]}{\sqrt{\mathrm{Var}[h_a(\boldsymbol{s})]\ \mathrm{Var}[h_b(\boldsymbol{s})]}}. \tag{43}$$

Thus $\rho$ is the matrix of signal correlations.

Putting Eq. (39) and Eq. (40) together, we obtain

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{n}) &= \mathbb{E}[\mathrm{Cov}[\boldsymbol{n}|\boldsymbol{s}]] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) \\ &= \sigma^2 r^\alpha W r^\alpha + r\,P, r.\end{aligned}$$

Finally, plugging this result and Eq. (38) into Eq. (37), up to an additive constant, we obtain

$$\begin{aligned}\mathcal{L} &= \ln\det\left(\sigma^2 r^\alpha W r^\alpha + rPr\right) - \ln\det(r^{2\alpha}) - \mu^{-1}\sum_{a=1}^{K}\mathbb{E}[n_a] \\ &= \ln\det\left(\sigma^2 W + r^{1-\alpha}Pr^{1-\alpha}\right) - \mu^{-1}\sum_{a=1}^{K} r_a,\end{aligned} \tag{44}$$

as required. Substituting back in $r_a = \omega_a g_a$ gets us Eq. (8).

## B.2 Upper bound for Poissonian noise

In this appendix, we consider the following model for cluster spike counts.

$$n_a|\boldsymbol{s} \sim \mathrm{Poisson}\left(g_a\Omega_a\right). \tag{45}$$

Here we derive a Gaussian upper bound to the mutual information, and show that an approximation to it leads to the same expression Eq. (8) for $\mathcal{L}$ derived for the Gaussian noise case, for $\sigma^2 W = I$ and $\alpha = 1/2$ (*i.e.*, uncorrelated unit Fano factor noise). We start with the objective

$$\mathcal{L}^0(\boldsymbol{g}) = 2I(\boldsymbol{n};\boldsymbol{s}) - \mu^{-1}\sum_{a=1}^{K} g_a\omega_a. \tag{46}$$

We decompose $I(\boldsymbol{n};\boldsymbol{s}) = H[\boldsymbol{n}] - H[\boldsymbol{n}|\boldsymbol{s}]$. We will once again upper bound the marginal entropy $H[\boldsymbol{n}]$. A difficulty arises from the fact that a Poisson random variable is discrete and the Gaussian upper bound we previously used is for a continuous random variables. We address this problem as follows: Consider a random variable $\boldsymbol{U}$ which is uniformly distributed on $[0,1)^K$, independent of $\boldsymbol{n}$. Then $\boldsymbol{n}+\boldsymbol{U}$ is a continuous random variable. We apply the Gaussian bound to

this. Let $p$ be the p.d.f. of $\boldsymbol{n}+\boldsymbol{U}$, $\mathcal{P}$ the p.m.f. of $\boldsymbol{n}$, and $u$ the p.d.f. of $\boldsymbol{U}$.

$$\begin{aligned}
H[\boldsymbol{n}+\boldsymbol{U}] &\leq \frac{K}{2}\ln(2\pi e) + \frac{1}{2}\ln\left(\det\left(\mathrm{Cov}(\boldsymbol{n}+\boldsymbol{U})\right)\right) \\
\mathrm{Cov}(\boldsymbol{n}+\boldsymbol{U}) &= \mathrm{Cov}(\boldsymbol{n}) + \mathrm{Cov}(\boldsymbol{U}) \\
\mathrm{Cov}(\boldsymbol{U}) &= \frac{1}{12}I_K \\
H[\boldsymbol{n}+\boldsymbol{U}] &= -\int_{\mathbb{R}^K} p(\boldsymbol{x})\ln(p(\boldsymbol{x}))d\boldsymbol{x} \\
S(\boldsymbol{y}) &= \{\boldsymbol{x}\in\mathbb{R}^K | y_i \leq x_i < y_i+1\} \\
H[\boldsymbol{n}+\boldsymbol{U}] &= -\sum_{\boldsymbol{y}\in\mathbb{N}^K}\int_{S(\boldsymbol{y})} p(\boldsymbol{x})\ln(p(\boldsymbol{x}))d\boldsymbol{x} \\
&= -\sum_{\boldsymbol{y}\in\mathbb{N}^K}\int_{S(\boldsymbol{y})} \mathcal{P}(\boldsymbol{y})u(\boldsymbol{x}-\boldsymbol{y})\ln(\mathcal{P}(\boldsymbol{y})u(\boldsymbol{x}-\boldsymbol{y}))d\boldsymbol{x} \\
&= -\sum_{\boldsymbol{y}\in\mathbb{N}^K}\int_{S(\boldsymbol{y})} \mathcal{P}(\boldsymbol{y})\ln(\mathcal{P}(\boldsymbol{y}))d\boldsymbol{x} \\
&= H[\boldsymbol{n}]
\end{aligned}$$

Putting this together gives us the upper bound

$$H[\boldsymbol{n}] \leq \frac{1}{2}\ln\left((2\pi e)^K \det\left(\frac{1}{12}I_K + \mathrm{Cov}(\boldsymbol{n})\right)\right). \tag{47}$$

We now address the problem of the marginal entropy $H[\boldsymbol{n}|\boldsymbol{s}]$. By conditional independence, we have that

$$H[\boldsymbol{n}|\boldsymbol{s}] = \sum_{a=1}^{K}\int H[\mathrm{Poisson}(g_a\Omega_a(\boldsymbol{s}))]P(\boldsymbol{s})d\boldsymbol{s}. \tag{48}$$

We now make use of an additional assumption, namely that the representational curves $\Omega_a$ have a baseline, and in particular

$$\Omega_a \gg 5\omega_a/\mu \tag{49}$$

everywhere. Under this condition we can make the approximation that for fixed $\boldsymbol{s}$,

$$H[\mathrm{Poisson}(g_a\Omega_a(\boldsymbol{s}))] \approx H[\mathcal{N}(g_a\Omega_a(\boldsymbol{s}), g_a\Omega_a(\boldsymbol{s}))]. \tag{50}$$

This means we can obtain the approximate upper bound:

$$\begin{aligned}
\mathcal{L}^0(\boldsymbol{g}) &\leq \ln\left((2\pi e)^K \det\left(\frac{1}{12}I_K + \mathrm{Cov}(\boldsymbol{n})\right)\right) - 2H[\boldsymbol{n}|\boldsymbol{s}] - \mu^{-1}\sum_{a=1}^{K} g_a\omega_a \\
&\approx \ln\left((2\pi e)^K \det\left(\mathrm{Cov}(\boldsymbol{n})\right)\right) - 2\sum_{a=1}^{K} H[\mathcal{N}(g_a\Omega_a(\boldsymbol{s}), g_a\Omega_a(\boldsymbol{s}))] - \mu^{-1}\sum_{a=1}^{K} g_a\omega_a \\
&= \mathcal{L}(\boldsymbol{g}) + \text{const.},
\end{aligned}$$

where $\mathcal{L}(\boldsymbol{g})$ is the same functional that we defined earlier for Gaussian random variables, in the case of uncorrelated, unit-fano factor noise, Eq. (8). We have made an additional approximation by neglecting the $I_K/12$ term, which relevant parameter conditions is negligible compared the covariance.

## B.3 Mathematical treatment of the clustered population

In this section we consider a clustered population in which the representational curves are a small perturbation to cluster-wide representational curves. Let $C_a$ denote cluster $a$ (or rather the set of indices of neurons belonging to that cluster). Then for $i \in C_a$,

$$\hat{\Omega}_i = \Omega_a + \epsilon\delta\Omega_i \tag{51}$$

We use hats to denote single-neuron quantities, and corresponding no-hat symbols for either cluster averages or zeroth-order values of tuning curves, etc, that are uniform across neurons in a cluster and hence only depend on cluster indices. In

particular, we use $P = \mathrm{CV}\rho\mathrm{CV}$, defined in Eq. (41), at cluster level (hence a $K \times K$ matrix), and use $\hat{P}$ to denote the equivalent matrix defined for individual neurons (an $N \times N$ matrix). As we noted in the main text Sec. 2.3, for this analysis, we limit ourselves to the case of uncorrelated noise, $\Sigma(\boldsymbol{s}) = I$, with Poisson-like scaling $\alpha = 1/2$. Mirroring Eq. (44), the population level objective in this case can be written as

$$\begin{aligned}\mathcal{L}_{\mathrm{pop}}(\hat{\boldsymbol{r}}) &= \ln\det\left(\sigma^2 I_N + \hat{r}^{1/2}\hat{P}\,\hat{r}^{1/2}\right) - \mu^{-1}\sum_{i=1}^{N}\hat{r}_i \\ &= \ln\det\left(I_N + \sigma^{-2}\hat{r}^{1/2}\hat{P}\,\hat{r}^{1/2}\right) - \mu^{-1}\sum_{i=1}^{N}\hat{r}_i + \text{const.}\end{aligned} \tag{52}$$

In the rest of this section we will set $\sigma$ to 1 by absorbing it into $\hat{P}$; at the end, we can replace all $\hat{P}$ —or factors proportional to them— by $\sigma^{-2}\hat{P}$ to recover the general case).

To simplify the mathematical derivation, we will assume that clusters are the same size $k = N/K$, and that w.l.o.g. the population is sorted so that neurons in the same cluster appear adjacent to each other in the ordering. However, the assumption of equal size clusters is not essential, and our final results are valid for the case of clusters of variable size as well. To zeroth order in $\epsilon$, the elements of $\hat{P}$ are constant over blocks corresponding to clusters; in other words

$$\hat{P} = P \otimes \mathbf{1}_k \mathbf{1}_k^{\mathrm{T}} + \delta\hat{P} \tag{53}$$

for some deviation $\delta\hat{P}$ that is $O(\epsilon)$ (here $\mathbf{1}_k$ denotes the $k$-dimensional vector with all components equal to 1). However, we will find that the first nonzero corrections to the loss (for $\epsilon > 0$) arise from $O(\epsilon^2)$ corrections to $\hat{P}$ (or equivalently to $\delta\hat{P}$ as defined by Eq. (53)). We will therefore (1) expand $\mathcal{L}$ to second order in $\delta\hat{P}$, (2) (in the next section) expand $\delta\hat{P}$ to second order in $\epsilon$, and (3) substitute $\delta\hat{P}$ in the loss and eventually neglect terms that are $o(\epsilon^2)$.

We start by expanding $\mathcal{L}$. Using Eq. (60), and plugging Eq. (53) into Eq. (52) and expanding to second order in $\delta\hat{P}$, we obtain

$$\mathcal{L} = \log\det(I_N + \mathcal{A} + \mathcal{B}) - \mu^{-1}\sum_{i=1}^{N}\hat{r}_i = \mathcal{L}^{(0)} + \mathcal{L}^{(1)} + \mathcal{L}^{(2)} + o(\delta\hat{P}^2) \tag{54}$$

where

$$\mathcal{L}^{(0)} = \log\det(I_N + \mathcal{A}) - \sum_{i=1}^{N}\hat{r}_i \tag{55}$$

$$\mathcal{L}^{(1)} = \mathrm{Tr}((I_N + \mathcal{A})^{-1}\mathcal{B}) \tag{56}$$

$$\mathcal{L}^{(2)} = -\frac{1}{2}\mathrm{Tr}((I_N + \mathcal{A})^{-1}\mathcal{B}(I_N + \mathcal{A})^{-1}\mathcal{B}). \tag{57}$$

and we defined the following matrices:

$$\mathcal{A} \equiv \hat{r}^{1/2}(P \otimes \mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}^{1/2}, \tag{58}$$

$$\mathcal{B} \equiv \hat{r}^{1/2}\delta\hat{P}\hat{r}^{1/2}. \tag{59}$$

Here and in the rest of this appendix we use $\mathbf{1}$, instead of $\mathbf{1}_k$, to denote the $k$-dimensional vector with all components equal to 1.

### B.3.1 Cluster level problem at the zeroth order

We now show that the zeroth order loss, Eq. (55), depends only on the cluster rates $r_a = \sum_{i \in C_a}\hat{r}_i$. This objective has two components: the information term and the energetic term. For the energetic term the claim is obvious, as we have

$$\sum_{i=1}^{N}\hat{r}_i = \sum_{a=1}^{K}\sum_{i \in C_a}\hat{r}_i = \sum_{a=1}^{K} r_a \tag{60}$$

where $r_a$ denotes the total rate of cluster $a$. As for the information term, and will show that

$$\ln\det\left(I_N + \mathcal{A}\right) = \ln\det\left(I_N + \hat{r}^{1/2}\left(P \otimes \mathbf{1}\mathbf{1}^{\mathrm{T}}\right)\hat{r}^{1/2}\right)$$

can be re-written (up to a constant) as

$$\ln\det\left(I_K + r^{1/2}\,P\,r^{1/2}\right), \tag{61}$$

reducing the problem to a cluster-level one (the second determinant is of a $K \times K$ matrix, rather than $N \times N$, and only depends on cluster rates $r_a$ and the zeroth-order signal correlations encoded in $P$). Let us denote the Cholesky factor of (the positive definite) $P$ by $U$, such that $P = UU^{\mathrm{T}}$. We thus have

$$P \otimes \mathbf{1}\mathbf{1}^{\mathrm{T}} = (U \otimes \mathbf{1})(U \otimes \mathbf{1})^{\mathrm{T}}.$$

Further defining $U_{\hat{r}}$ to be the $N \times K$ matrix $\hat{r}^{1/2}(U \otimes \mathbf{1})$, we find

$$\mathcal{A} = \hat{r}^{1/2}(P \otimes \mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}^{1/2} = U_{\hat{r}}U_{\hat{r}}^{\mathrm{T}}. \tag{62}$$

By the matrix determinant lemma we then obtain

$$\det(I_N + \mathcal{A}) = \det\left(I_K + U_{\hat{r}}^{\mathrm{T}}U_{\hat{r}}\right). \tag{63}$$

Finally,

$$U_{\hat{r}}^{\mathrm{T}}U_{\hat{r}} = (U \otimes \mathbf{1})^{\mathrm{T}}\hat{r}(U \otimes \mathbf{1}) = U^{\mathrm{T}}rU = U^{\mathrm{T}}r^{1/2}r^{1/2}U. \tag{64}$$

Using this and making use of the matrix determinant lemma one more time, we see that the determinant on the right hand side of Eq. (63) can be written as Eq. (61).

Putting together the information and energetic terms (and momentarily returning the $\sigma^2$), we obtain that (up to an additive constant)

$$\mathcal{L}^{(0)} = \ln\det\left(\sigma^2 I_K + r^{1/2}P\,r^{1/2}\right) - \mu^{-1}\sum_{a=1}^{K} r_a. \tag{65}$$

### B.3.2 Expansion of $\delta\hat{P}$ in $\epsilon\delta\Omega$

We will denote $\epsilon\delta\Omega_i$ above by $V_i$ in this subsections (so $V = O(\epsilon)$), and will use $\delta$'s to denote deviation from expectation: $\delta X = X - \mathbb{E}[X]$. We have

$$\hat{\Omega}_a^i = \Omega_a + V_a^i \tag{66}$$

$$\hat{\omega}_a^i = \omega_a + v_a^i \tag{67}$$

where $v_a^i = \mathbb{E}[V_a^i]$. We also have

$$\mathrm{Cov}[\hat{\Omega}]_{ab}^{ij} = \mathbb{E}[\delta\hat{\Omega}_a^i\delta\hat{\Omega}_b^j] = \mathrm{Cov}[\Omega]_{ab} + \mathbb{E}[\delta V_a^i\delta\Omega_b] + \mathbb{E}[\delta\Omega^a\delta V_b^j] + \mathrm{Cov}[V]_{ab}^{ij} \tag{68}$$

or (keeping cluster indices explicit but hiding witihn-cluster ones)

$$\mathrm{Cov}[\hat{\Omega}]_{ab} = K_{ab}\mathbf{1}\mathbf{1}^{\mathrm{T}} + \boldsymbol{c}_{ab}\mathbf{1}^{\mathrm{T}} + \mathbf{1}\boldsymbol{c}_{ba}^{\mathrm{T}} + \mathrm{Cov}[V]_{ab} \tag{69}$$

where we defined the vector $\boldsymbol{c}_{ab}$ to have components

$$c_{ab}^i = \mathbb{E}[\delta V_a^i\delta\Omega_b]. \tag{70}$$

and

$$K_{ab} = \mathrm{Cov}[\Omega_a, \Omega_b]. \tag{71}$$

Since

$$\hat{P} = \hat{\mathrm{C}}\mathrm{V}\hat{\rho}\hat{\mathrm{C}}\mathrm{V} = \hat{\omega}^{-1}\mathrm{Cov}[\hat{\Omega}]\hat{\omega}^{-1}. \tag{72}$$

we also expand the diagonal factors to second order in $V$:

$$\begin{aligned}\hat{\omega}_a^{-1} &= \omega_a^{-1}I_k - \omega_a^{-1}v_a\omega_a^{-1} + \omega_a^{-1}v_a\omega_a^{-1}v_a\omega_a^{-1} + o(V^2),\\ &= \omega_a^{-1}\left(I_k - u_a + u_a^2\right) + o(V^2),\end{aligned} \tag{73}$$

where we defined

$$u_a = \omega_a^{-1}v_a, \qquad \boldsymbol{u}_a = \omega_a^{-1}\boldsymbol{v}_a, \tag{74}$$

(note that $\omega_a$ are numbers, but $\hat{\omega}_a$, $v_a$ and $u_a$ are $k \times k$ diagonal matrices).

First we consider the first order (in $V$ or equivalently in $\epsilon$) contributions to $\delta\hat{P}$. Using Eq. (69) and Eq. (73) in Eq. (72) we have

$$\delta\hat{P}_{ab}^{(1)} = (\omega_a\omega_b)^{-1}\left[\boldsymbol{c}_{ab}\mathbf{1}^{\mathrm{T}} + \mathbf{1}\boldsymbol{c}_{ba}^{\mathrm{T}} - K_{ab}\boldsymbol{u}_a\mathbf{1}^{\mathrm{T}} - K_{ab}\mathbf{1}\boldsymbol{u}_b^{\mathrm{T}}\right] \tag{75}$$

$$= (\omega_a\omega_b)^{-1}\left[\boldsymbol{f}_{ab}\mathbf{1}^{\mathrm{T}} + \mathbf{1}\boldsymbol{f}_{ba}^{\mathrm{T}}\right] \tag{76}$$

where we defined

$$\boldsymbol{f}_{ab} \equiv \boldsymbol{c}_{ab} - K_{ab}\boldsymbol{u}_a. \tag{77}$$

Next, we obtain the second order contributions to $\delta\hat{P}$. Using Eq. (69) and Eq. (73) in Eq. (72), we obtain

$$\delta\hat{P}^{(2)}_{ab} = (\omega_a\omega_b)^{-1}\left(\mathrm{Cov}[V]_{ab} - u_a\boldsymbol{f}_{ab}\mathbf{1}^{\mathrm{T}} - \mathbf{1}\boldsymbol{f}_{ba}^{\mathrm{T}}u_b - \boldsymbol{u}_a\boldsymbol{c}_{ba}^{\mathrm{T}} - \boldsymbol{c}_{ab}\boldsymbol{u}_b^{\mathrm{T}} + K_{ab}\boldsymbol{u}_a\boldsymbol{u}_b^{\mathrm{T}}\right), \tag{78}$$

and in particular

$$\delta\hat{P}^{(2)}_{aa} = (\omega_a)^{-2}\left(\mathrm{Cov}[V]_{aa} - u_a\boldsymbol{f}_{aa}\mathbf{1}^{\mathrm{T}} - \mathbf{1}\boldsymbol{f}_{aa}^{\mathrm{T}}u_a - \boldsymbol{u}_a\boldsymbol{c}_{aa}^{\mathrm{T}} - \boldsymbol{c}_{aa}\boldsymbol{u}_a^{\mathrm{T}} + K_{aa}\boldsymbol{u}_a\boldsymbol{u}_a^{\mathrm{T}}\right). \tag{79}$$

#### B.3.3 $\delta\Omega_i$ corrections to the loss

We now consider the contributions of the first and second order corrections to $\delta\hat{P}$ to the loss, via Eqs. (56) and (57): to calculate corrections to the loss to $O(\epsilon^2)$, we need to plug $\delta\hat{P}^{(1)} + \delta\hat{P}^{(2)}$ into $\mathcal{L}^{(1)}$, and plug in $\delta\hat{P}^{(1)}$ (only) in $\mathcal{L}^{(2)}$. Firstly, consider the matrix inverse that appears in expressions Eqs. (56)–(57) for $\mathcal{L}^{(1)}$ and $\mathcal{L}^{(2)}$. Using $A = U_{\hat{r}}U_{\hat{r}}^{\mathrm{T}}$ from Eq. (62) and the Woodbury matrix identity, we can write

$$\begin{aligned}(I_N + \mathcal{A})^{-1} &= \left(I_N + U_{\hat{r}}U_{\hat{r}}^{\mathrm{T}}\right)^{-1} \\ &= I_N - U_{\hat{r}}\left(I_K + U_{\hat{r}}^{\mathrm{T}}U_{\hat{r}}\right)^{-1}U_{\hat{r}}^{\mathrm{T}}\end{aligned}$$

Using $U_{\hat{r}}^{\mathrm{T}}U_{\hat{r}} = U^{\mathrm{T}}rU$ –see Eq. (64)– we find

$$\begin{aligned}U_{\hat{r}}\left(I_K + U_{\hat{r}}^{\mathrm{T}}U_{\hat{r}}\right)^{-1}U_{\hat{r}}^{\mathrm{T}} &= \hat{r}^{1/2}\left(U\otimes\mathbf{1}\right)\left[I_K + U^{\mathrm{T}}rU\right]^{-1}\left(U^{\mathrm{T}}\otimes\mathbf{1}^{\mathrm{T}}\right)\hat{r}^{1/2} \\ &= \hat{r}^{1/2}\left[\left(U[I_K + U^{\mathrm{T}}rU]^{-1}U^{\mathrm{T}}\right)\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}}\right]\hat{r}^{1/2}.\end{aligned}$$

Assuming $P$ is full-rank (the generic case), we have $U[I_K + U^{\mathrm{T}}rU]^{-1}U^{\mathrm{T}} = \left(\left(U^{-1}\right)^{\mathrm{T}}U^{-1} + r\right)^{-1} = \left(P^{-1} + r\right)^{-1}$, yielding

$$(I_N + \mathcal{A})^{-1} = I_N - \hat{r}^{1/2}\left(\left[P^{-1} + r\right]^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)\hat{r}^{1/2}. \tag{80}$$

Up until this point, all expressions we have obtained have been exact (apart from the $\epsilon$-expansion itself). We now make an approximation to the full perturbation objective which is valid in the high signal-to-noise ratio regime as outlined in Sec. 2.4. Specifically, we will take the matrix

$$\Delta = (\mu P)^{-1} = (\mu\sigma^{-2}\mathrm{CV}\rho\mathrm{CV})^{-1}$$

to be small, and expand to zeroth order in this matrix. $\Delta$ represents the high-dimensional structure of signal-to-noise ratio. As we show in App. B.4, to zeroth order in this matrix, $r = \mu I_K$. Using the approximation

$$P^{-1} + r = \mu\left(\Delta + \mu^{-1}r\right) \approx r, \tag{81}$$

valid to zeroth-order in $\Delta$, in Eq. (80) we obtain

$$(I_N + \mathcal{A})^{-1} \approx I_N - \mathcal{C}, \tag{82}$$

where we defined

$$\mathcal{C} \equiv \hat{r}^{1/2}(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}^{1/2}. \tag{83}$$

Note that in writing Eq. (81) we have assumed that cluster rates are $O(\mu)$, so that the second term in the parenthesis is $O(1)$ and dominates $\Delta$; this will be justified *a posteriori*, by the homeostasis result for the inter-cluster problem (as the approximate optimiser of Eq. (65), in the small $\Delta$ regime).

We will show that $\mathcal{C}$ and hence $I - \mathcal{C}$ are projection operators and we will characterise the latter's kernel (*i.e.*, the vectors annihilated by it). First $\mathcal{C}$ is clearly a symmetric matrix. So we just need to show that $\mathcal{C}^2 = \mathcal{C}$. To prove this (and other statements), we note that products of $N\times N$ matrices (or their products with $N$-dimensional vectors) can be written in terms of products of their blocks (corresponding to the clustering of neurons) as follows: $(AB)_{ab} = \sum_{c=1}^{K} A_{ac}B_{cb}$ where the subscripts index blocks, with $a, b \in \{1,\dots K\}$, and $A_{ac}$ and $B_{cb}$ are multiplying as matrices. As we will see our results do not truly rely on the tensor product structure we have assumed for various matrices (here $\mathcal{A}$ and $\mathcal{C}$); but only on their

constancy within blocks defined by clusters. *Thus our results generalise to the case where clusters contain different number of neurons.* For simplicity, however, we will stick to the tensor structure, corresponding to the same number of neurons in different clusters. Using the above observation, we can write (note that below $r_a$ and $\delta_{ab}$ are scalars, while $\hat{r}_{aa}$ denotes the $a$-th diagonal block of $\hat{r}$ and is a diagonal matrix)

$$
\begin{aligned}
(\mathcal{C}^2)_{ab} &= \sum_{c=1}^{K} \hat{r}_{aa}^{1/2}(\delta_{ac} r_a^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}_{cc}(\delta_{bc} r_b^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}_{bb}^{1/2} = \delta_{ab}\,\hat{r}_{aa}^{1/2}(r_a^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}_{aa}(r_a^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}_{aa}^{1/2} \\
&= \hat{r}_{aa}^{1/2}\delta_{ab}\mathbf{1}\frac{\mathbf{1}^{\mathrm{T}}\hat{r}_{aa}\mathbf{1}}{r_a^2}\mathbf{1}^{\mathrm{T}}\hat{r}_{aa}^{1/2} = \hat{r}_{aa}^{1/2}(\delta_{ab} r_a{}^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}_{aa}^{1/2} = (\hat{r}^{1/2}(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}^{1/2})_{ab} = (\mathcal{C})_{ab}.
\end{aligned}
$$

where we used $\mathbf{1}^{\mathrm{T}}\hat{r}_{aa}\mathbf{1} = \sum_{i\in C_a}\hat{r}_i = r_a$.

Similarly for vectors of the form $\hat{\boldsymbol{w}} = \boldsymbol{w}\otimes\mathbf{1}$ (namely vectors with components that are constant over each cluster), we have $(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}\,\hat{\boldsymbol{w}} = \hat{\boldsymbol{w}}$, as

$$
[(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}\,\hat{\boldsymbol{w}}]_a = \sum_{b=1}^{K}(\delta_{ab} r_a^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}_{bb}(w_b\mathbf{1}) = w_a\mathbf{1}(r_a^{-1}\mathbf{1}^{\mathrm{T}}\hat{r}_{aa}\mathbf{1}) = w_a\mathbf{1} = \hat{\boldsymbol{w}}_a.
$$

(note that $w_a$ are scalars). Thus if $\hat{\boldsymbol{w}}$ is any such vector, then $\hat{r}^{1/2}\hat{\boldsymbol{w}}$ is left invariant by $\mathcal{C}$: $\mathcal{C}\hat{r}^{1/2}\hat{\boldsymbol{w}} = \hat{r}^{1/2}(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}\hat{\boldsymbol{w}} = \hat{r}^{1/2}\hat{\boldsymbol{w}}$. Finally, this means that $\hat{r}^{1/2}\hat{\boldsymbol{w}}$ is in the kernel of $I-\mathcal{C} = (I+\mathcal{A})^{-1}$ and is annihilated by it. Moreover, since $\mathcal{C}$ is symmetric, row vectors of the form $\boldsymbol{w}^{\mathrm{T}}\otimes\mathbf{1}^{\mathrm{T}}$ are annihilated when multiplied by $\hat{r}^{1/2}(I-\mathcal{C}) = \hat{r}^{1/2}(I+\mathcal{A})^{-1}$ on the right. Finally, matrices with cluster blocks that have uniform rows (columns) are annihilated when multiplied by $(I+\mathcal{A})^{-1}\hat{r}^{1/2}$ (by $\hat{r}^{1/2}(I+\mathcal{A})^{-1}$) on the left (right).

It follows immediately that, given the structure of Eqs. (56)–(58), any of the terms in the expressions Eqs. (76)–(89) for blocks of $\delta\hat{P}^{(1)}$ and $\delta\hat{P}^{(2)}$ that have $\mathbf{1}$ as an outer product factor contribute nothing to $\delta\mathcal{L}$. In particular, the $O(\epsilon)$ part, $\delta\hat{P}^{(1)}$, makes no contribution to the loss. Since we are interested in leading order corrections, it thus suffices to only consider the contribution of $\delta\hat{P}^{(2)}$ —after dropping terms involving $\mathbf{1}$ in Eq. (89)— as it enters $\mathcal{L}^{(1)}$ (in the $\Delta\ll 1$ limit). Denoting this correction by $\delta\mathcal{L}$, from Eqs. (56)–(58) and Eq. (82) we find

$$
\begin{aligned}
\delta\mathcal{L} &= \mathrm{Tr}((I_N-\mathcal{C})\mathcal{B}^{(2)}) && (84)\\
&= \mathrm{Tr}(\hat{r}\delta\hat{P}^{(2)}) - \mathrm{Tr}(\hat{r}(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}\delta\hat{P}^{(2)}). && (85)
\end{aligned}
$$

Traces of $N\times N$ matrix products can be written in terms of traces of products of their blocks as $\mathrm{Tr}(AB) = \sum_{ab}\mathrm{tr}(A_{ab}B_{ab})$ where tr denotes trace over $k\times k$ blocks. We get a further simplification because $\hat{r}$ is diagonal and $\hat{r}(r^{-1}\otimes\mathbf{1}\mathbf{1}^{\mathrm{T}})\hat{r}$ is block-diagonal (due to the diagonality of $\hat{r}$ and $r$), with the diagonal blocks given by $r_a^{-1}\hat{\boldsymbol{r}}_a\hat{\boldsymbol{r}}_a^{\mathrm{T}}$. But if $D$ is a block-diagonal matrix, then $\mathrm{Tr}(DB) = \sum_a \mathrm{tr}(D_{aa}B_{aa})$. Thus

$$
\begin{aligned}
\delta\mathcal{L}(\hat{\boldsymbol{r}}) &= \sum_{a=1}^{K}\delta\mathcal{L}_a(\hat{\boldsymbol{r}}_a) && (86)\\
\delta\mathcal{L}_a(\hat{\boldsymbol{r}}_a) &= \mathrm{tr}(\hat{r}_a\delta\hat{P}_{aa}^{(2)}) - r_a^{-1}\mathrm{tr}(\hat{\boldsymbol{r}}_a\hat{\boldsymbol{r}}_a^{\mathrm{T}}\delta\hat{P}_{aa}^{(2)}) && (87)\\
&= \mathrm{tr}(\hat{r}_a\delta\hat{P}_{aa}^{(2)}) - r_a^{-1}\hat{\boldsymbol{r}}_a^{\mathrm{T}}\delta\hat{P}_{aa}^{(2)}\hat{\boldsymbol{r}}_a. && (88)
\end{aligned}
$$

In other words, the correction to the loss is a sum of terms, each of which depends only on the rates in one of the cluster only. As a result, given the total cluster rates $r_a$, the optimisation of rates of single neurons decouples across clusters. Which means we can optimise the rates in cluster $a$ under the constraint that they sum up to $r_a$, to maximise Eq. (88). Using Eq. (89) we now express $\delta\mathcal{L}_a$ explicitly.

#### B.3.4 Within-cluster loss to leading order in $\epsilon$

From Eq. (79), after dropping the terms with factors of $\mathbf{1}$, we have

$$
\delta\hat{P}_{aa}^{(2)} \propto \mathrm{Cov}[V]_{aa} - \boldsymbol{u}_a\boldsymbol{c}_{aa}^{\mathrm{T}} - \boldsymbol{c}_{aa}\boldsymbol{u}_a^{\mathrm{T}} + K_{aa}\boldsymbol{u}_a\boldsymbol{u}_a^{\mathrm{T}}. \tag{89}
$$

We further ignored the factor $\omega_a^{-2}$ which multiplies both terms in the loss and is thus irrelevant to the optimisation; we will similarly ignore a factor of $\mu$ in the loss below. The right hand side of the above expression is the covariance of the zero-mean vector (of random variables)

$$
\Delta\hat{\Omega}_a^i = V_a^i - u_a^i\Omega_a = \delta V_a^i - u_a^i\delta\Omega_a \tag{90}
$$

(as can be easily checked: $\mathrm{Cov}[\Delta\hat{\Omega}^i_a, \Delta\hat{\Omega}^j_a] = \mathbb{E}[\Delta\hat{\Omega}^i_a \Delta\hat{\Omega}^j_a] = \mathrm{Cov}[V^i_a, V^j_a] - u^i_a c^j_{aa} - c^i_{aa} u^j_a + \mathbb{E}[\delta\Omega^2_a] u^i_a u^j_a$ as $c^i_{aa} = \mathbb{E}[\delta V^i_a \delta\Omega_a]$).

Since the problem has decoupled across clusters we will now drop the cluster index $a$; in particular we will denote the unperturbed $\Omega_a$ by $\Omega_0$, and we will also denote the given/fixed total cluster rate, $r_a$, by $R_0$. Substituting the above in Eq. (88) and ignoring irrelevant overall prefactors, we find that the single-neuron rates in a cluster with total rate $R_0$ are solutions of the following quadratic programming problem:

$$\max_{\substack{\hat{\boldsymbol{r}}: \\ \sum_i \hat{r}_i = R_0}} \left[ \hat{\boldsymbol{r}}^{\mathrm{T}} \mathrm{Var}(\Delta\boldsymbol{\Omega}) - R_0^{-1} \hat{\boldsymbol{r}}^{\mathrm{T}} \mathrm{Cov}(\Delta\boldsymbol{\Omega}) \hat{\boldsymbol{r}} \right], \tag{91}$$

where

$$\Delta\Omega^i = V^i - \frac{\Omega_0}{\omega_0} v^i. \tag{92}$$

This is equivalent (up to replacement of $R_0$ with $\mu$) to the optimisation problem, Eq. (12), of the main text. Finally, note the scaling property of this quadratic programming problem: if we scale the cluster rate, $R_0$, by some $\alpha$, the solution $\hat{\boldsymbol{r}}$, and thus the rates of individual neurons in the cluster, scale by the same factor. Note also that the scaling of $\Delta\Omega^i$ by a constant (within a cluster) factor does not make a difference, and we can make it "dimensionless" by dividing it by $\omega_0$ if helpful (in fact this amounts to returning the prefactor $\omega_a^{-2}$ that we dropped after Eq. (89)).

### B.3.5 A toy model for a cluster's tuning curves

Here we will develop a toy model for $\mathrm{Cov}(\Delta\boldsymbol{\Omega})$, *i.e.*, for the statistics of deviations of representational curves of neurons in the same cluster from the unperturbed curve $\Omega_0$. Assuming a degree of smoothness for $\Omega^i(\boldsymbol{s})$ (as a function on the stimulus space), we will adopt a basis of smooth functions $b_\mu(\boldsymbol{s})$ for $1 \leq \mu \leq N_b$, and assume that the log tuning curves are in the span of this set of functions. In other words, we assume

$$\Omega^i(\boldsymbol{s}) \propto \exp(\boldsymbol{w}^i \cdot \boldsymbol{b}(\boldsymbol{s})). \tag{93}$$

for some vector of coefficients $\boldsymbol{w}^i = (w^i_1, \ldots, w^i_{N_b})$. Similar to Eq. (51) we assume $\boldsymbol{w}^i = \boldsymbol{w}^i_0 + \delta\boldsymbol{w}^i$ (we absorb $\epsilon$ into $\delta\boldsymbol{w}^i$). Expanding the above equation we get

$$\Omega^i(\boldsymbol{s}) = \Omega^i_0(\boldsymbol{s})(1 + \delta\boldsymbol{w}^i \cdot \boldsymbol{b}(\boldsymbol{s}))$$

or equivalently

$$V^i(\boldsymbol{s}) = \Omega^i_0(\boldsymbol{s}) \delta\boldsymbol{w}^i \cdot \boldsymbol{b}(\boldsymbol{s}). \tag{94}$$

It follows that $v^i = \mathbb{E}[\Omega_0 \boldsymbol{b}] \cdot \delta\boldsymbol{w}^i$, and thus (from Eq. (92))

$$\begin{aligned} \Delta\Omega^i(\boldsymbol{s}) &= \delta\boldsymbol{w}^i \cdot \left( \Omega_0(\boldsymbol{s}) \boldsymbol{b}(\boldsymbol{s}) - \frac{\Omega_0(\boldsymbol{s})}{\omega_0} \mathbb{E}[\Omega_0 \boldsymbol{b}] \right) \\ &= \delta\boldsymbol{w}^i \cdot \Delta\boldsymbol{b}(\boldsymbol{s}) \Omega_0(\boldsymbol{s}) \end{aligned} \tag{95}$$

where we defined

$$\Delta\boldsymbol{b}(\boldsymbol{s}) = \boldsymbol{b}(\boldsymbol{s}) - \mathbb{E}\left[ \frac{\Omega_0}{\omega_0} \boldsymbol{b} \right]. \tag{96}$$

We thus obtain:

$$\mathrm{Cov}(\Delta\boldsymbol{\Omega})_{ij} = \delta\boldsymbol{w}_i^{\mathrm{T}} G \, \delta\boldsymbol{w}_j \tag{97}$$

$$G := \mathbb{E}\left[ \Omega_0^2 \Delta\boldsymbol{b} \Delta\boldsymbol{b}^{\mathrm{T}} \right] = \int \Delta\boldsymbol{b}(\boldsymbol{s}) \Delta\boldsymbol{b}(\boldsymbol{s})^{\mathrm{T}} \, \Omega_0^2(\boldsymbol{s}) P(\boldsymbol{s}) d\boldsymbol{s}. \tag{98}$$

We think of $G$ as a metric defining the inner product of the vectors $\delta\boldsymbol{w}_i$. Note also that given the final comment in the previous subsection, we can replace $\Omega_0(\boldsymbol{s})$ inside the integral expression for $G$ with $\Omega_0(\boldsymbol{s})/\omega_0$; this is nice as it means that if it happens that the cluster tuning curve $\Omega_0$ is mostly supported on regions with small stimulus frequency, the above metric does not go to zero. Alternatively, since the scaling of $\mathrm{Cov}(\Delta\boldsymbol{\Omega})$ is irrelevant to the within-cluster rate optimisation, we can assume that the expectation in the definition of $G$ is taken, not with respect to $P(\boldsymbol{s})$, but with respect to the *normalised* measure with density proportional to $\Omega_0^2(\boldsymbol{s}) P(\boldsymbol{s})$. If we denote expectation under this "tilted" measure by $\mathbb{E}_{\Omega_0^2}$ we can also write:

$$G := \mathbb{E}_{\Omega_0^2}\left[ \Delta\boldsymbol{b} \Delta\boldsymbol{b}^{\mathrm{T}} \right]. \tag{99}$$

Similarly we could have written

$$\Delta\boldsymbol{b}(\boldsymbol{s}) := \boldsymbol{b}(\boldsymbol{s}) - \mathbb{E}_{\Omega_0}[\boldsymbol{b}(\boldsymbol{s})] \tag{100}$$

(note that this last expectation is tilted by $\Omega_0$ and not by $\Omega_0^2$). To summarise, we have found that the quadratic form for our quadratic programming optimisation is given by the Gram matrix of the neurons' coefficients $\delta \boldsymbol{w}_i$ with respect to the inner product defined by $G$.

For our minimal toy model, we will assume that the perturbations $\delta \boldsymbol{w}_i$are isotropic, mean-zero Gaussians, *i.e.*, $\delta \boldsymbol{w} \sim \mathcal{N}(0, I_{N_b})$. We will additionally assume that the positive-definite matrix $G$ has $D$ eigenvalues which are roughly equal, and all significantly larger than the other eigenvalues. In this case, (up to scaling) we can approximate $G$ as a projection matrix onto $D$-dimensions. This allows us to generate $\mathrm{Cov}(\Delta\hat{\boldsymbol{\Omega}})_{ij}$ simply as $\delta \boldsymbol{w}_i^T \delta \boldsymbol{w}_j$, where the vectors $\{\delta \boldsymbol{w}\}_{i=1}^k$ were drawn independently from a $D$-dimensional isotropic Gaussian, $\delta \boldsymbol{w}_i \sim \mathcal{N}(0, I_D)$.

## B.4 First order maximisation of $\mathcal{L}$

In this section we compute the maximiser of $\mathcal{L}$, to first order in the parameter $\Delta$, defined by Eq. (15):

$$\Delta := \frac{\sigma^2}{(\beta\mu)^\beta} \left(\mathrm{CV}\rho\mathrm{CV}\right)^{-1} W = \frac{\sigma^2}{(\beta\mu)^\beta} P^{-1} W.$$

To simplify the derivation we will work with $\mathcal{L}$ as a function of the rates $r_a = g_a \omega_a$, rather than the gains. Writing Eq. (8) in terms of $r_a$, we have

$$\mathcal{L} = \log\det\left(\sigma^2 W + r^{1-\alpha}\mathrm{CV}\rho\mathrm{CV} r^{1-\alpha}\right) - \mu^{-1}\sum_{a=1}^K r_a.$$

We will also introduce the matrix

$$M = r^{1-\alpha}\mathrm{CV}\rho\mathrm{CV} r^{1-\alpha} = r^{1-\alpha} P\, r^{1-\alpha}, \tag{101}$$

so that

$$\mathcal{L} = \log\det\left(\sigma^2 W + M\right) - \mu^{-1}\sum_{a=1}^K r_a. \tag{102}$$

Note that

$$\frac{\partial M}{\partial r_a} = \frac{1-\alpha}{r_a}\left(E_a M + M E_a\right) \tag{103}$$

where $E_a$ is the diagonal matrix which is zero everywhere except having 1 as the $a$-th element along the diagonal.

Using Eq. (102) and Eq. (103), we obtain that

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial r_a} &= \frac{1-\alpha}{r_a}\mathrm{tr}\left(\left[\sigma^2 W + M\right]^{-1}\left(E_a M + M E_a\right)\right) - \mu^{-1}, && (104)\\
&= \frac{2(1-\alpha)}{r_a}\mathrm{tr}\left(\left[\sigma^2 W + M\right]^{-1} M E_a\right) - \mu^{-1}, && (105)\\
&= \frac{2(1-\alpha)}{r_a}\mathrm{tr}\left(\left[\sigma^2 M^{-1} W + I_K\right]^{-1} E_a\right) - \mu^{-1}, && (106)\\
&= \frac{2(1-\alpha)}{r_a}\left[I_K + \sigma^2 M^{-1} W\right]^{-1}_{aa} - \mu^{-1}, && (107)
\end{aligned}$$

where we used the cyclicity of trace and the symmetry of $M$ and $W$. Setting the derivative in Eq. (104) to zero and rearranging, and defining $\beta = 2(1-\alpha)$, we obtain

$$r_a = \mu\beta\left[(I_K + \sigma^2 M^{-1} W)^{-1}\right]_{aa}. \tag{108}$$

We now apply the ansatz

$$r_a = \mu\beta(1 - \Delta_{aa} + \mathcal{O}(\Delta^2)). \tag{109}$$

Plugging this into Eq. (101) we have that

$$\begin{aligned}
M = r^{1-\alpha} P\, r^{1-\alpha} &= (\mu\beta)^\beta P + \mathcal{O}(\Delta),\\
\sigma^2 M^{-1} W &= \Delta + \mathcal{O}\left(\Delta^2\right),
\end{aligned}$$

which after substitution in the right hand side of Eq. (108) yields

$$\begin{aligned}
\mathrm{RHS} &= \mu\beta\left[I + \Delta + \mathcal{O}\left(\Delta^2\right)\right]^{-1}_{aa}\\
&= \mu\beta\left[I - \Delta + \mathcal{O}\left(\Delta^2\right)\right]_{aa}\\
&= \mu\beta\left(1 - \Delta_{aa}\right) + \mathcal{O}\left(\Delta^2\right),
\end{aligned}$$

where on the penultimate line we have used the Neumann expansion of $\left[I + \Delta + \mathcal{O}\left(\Delta^2\right)\right]^{-1}$.

## B.5 Aligned noise leads to perfect homeostasis and uniformisation

In this appendix we demonstrate that (for constant coefficient of variation) if $\rho$ and $W$ are perfectly aligned *i.e.*, , $\rho = W$, then the optimal rates are all equal and invariant of the environment, *i.e.*, we have perfect homeostasis and uniformisation.

In App. B.4 we derived an equation for the maximiser of the objective function $\mathcal{L}(\boldsymbol{r})$, Eq. (108),

$$r_a = \mu\beta \left[\sigma^2 r^{\alpha-1}(\mathrm{CV}\rho\mathrm{CV})^{-1} r^{\alpha-1} W + I\right]_{aa}^{-1} \tag{110}$$

We now show that, when $W = \rho$ and CV is constant, this equation is satisfied by constant $\boldsymbol{r} = \chi \mathbf{1}_k$ for some $\chi$, and find an equation for $\chi$.

$$\begin{aligned}
r^{\alpha-1}(\mathrm{CV}\rho\mathrm{CV})^{-1} r^{\alpha-1} W &= \chi^{2(\alpha-1)}\mathrm{CV}^{-2}\rho^{-1} W \\
&= \chi^{2(\alpha-1)}\mathrm{CV}^{-2} I \\
\left[\sigma^2 r^{\alpha-1}(\mathrm{CV}\rho\mathrm{CV})^{-1} r^{\alpha-1} W + I\right]_{aa}^{-1} &= \left[\sigma^2 \chi^{-\beta}\mathrm{CV}^{-2} I + I\right]_{aa}^{-1} \\
&= \frac{1}{\sigma^2\chi^{-\beta}\mathrm{CV}^{-2} + 1} \\
\chi &= \frac{\mu\beta}{\sigma^2\chi^{-\beta}\mathrm{CV}^{-2} + 1} \\
\frac{\sigma^2}{\mathrm{CV}^2}\chi^{1-\beta} + \chi &= \mu\beta \\
f(\chi) &:= \frac{\sigma^2}{\mathrm{CV}^2}\chi^{1-\beta} + \chi
\end{aligned}$$

We will now show that this equation has a solution, provided certain conditions on the parameters hold. We divide into two cases. If $\beta \leq 1$ then $f(\chi)$ is strictly increasing and limits to 0 as $\chi \to 0$ and infinity as $\chi \to \infty$. By continuity we must have $f(\chi) = \mu\beta$ at some point.

If $\chi > 1$ then we still have $f(\chi) \to \infty$ as $\chi \to \infty$. We now find the minimum of $f$. Taking derivatives, we see that this occurs at

$$\chi = \left(\frac{\sigma^2}{\mathrm{CV}^2}(\beta - 1)\right)^{1/\beta} \tag{111}$$

At this point, we have

$$f(\chi) = \left(\frac{\sigma^2}{\mathrm{CV}^2}(\beta - 1)\right)^{1/\beta}\left(\frac{\beta}{\beta - 1}\right). \tag{112}$$

This is less than $\mu\beta$ provided that

$$\frac{\sigma^2}{\mathrm{CV}^2} \leq \mu^\beta(\beta - 1)^{\beta-1}. \tag{113}$$

So provided Eq. (113) holds, there is a solution by continuity. Recall that $\mu\beta$ is approximately the average firing rate of a cluster. Typically we are interested in holding the product of these parameters constant at a value $r$. If we do so, we can take the derivative of the right hand side of Eq. (113) in $\beta$, obtaining that this expression is minimised when $\beta = r/(r-1)$, where it attains a value of $r - 1$. Putting this together, we can say that there exists a uniform, homeostatic solution provided

$$\frac{\sigma^2}{\mathrm{CV}^2} \leq r - 1. \tag{114}$$

where $r$ is the product $\beta\mu$.

## B.6 Estimates for $\mathrm{tr}(\hat{\rho}^{-1}W)$

We define $A_\gamma$ to be the normalisation constant of a correlation matrix with an eigenspectrum proportional to $1/n^\gamma$. Since a $K \times K$ correlation matrix has trace $K$, we have

$$A_\gamma = \frac{K}{\sum_{n=1}^{K}\frac{1}{n^\gamma}}. \tag{115}$$

We start with the case $\gamma \neq 1$. Then we have that

$$\begin{aligned}
\frac{K}{A_\gamma} &= \sum_{n=1}^{K} \frac{1}{n^\gamma} \\
&\approx \int_1^{K+1} \frac{1}{x^\gamma} dx \\
&= \left[ \frac{x^{1-\gamma}}{1-\gamma} \right]_1^{K+1} \\
&= \frac{(K+1)^{1-\gamma} - 1}{1-\gamma}.
\end{aligned}$$

Now, taking the limit as $\gamma \to 1$, we obtain that

$$\frac{K}{A_1} \approx \ln(K). \tag{116}$$

Now when $\gamma < 1$ and $K$ is large, we obtain

$$\frac{K}{A_\gamma} \approx \frac{K^{1-\gamma}}{1-\gamma}. \tag{117}$$

Lastly, when $\gamma > 1$ and $K$ is large, we obtain

$$\frac{K}{A_\gamma} \approx \frac{1}{\gamma - 1}. \tag{118}$$

These approximations allow us to arrive at the following estimates. Firstly, consider $\mathrm{tr}(\rho^{-1})/K$.

$$\begin{aligned}
\mathrm{tr}(\rho^{-1})/K &= \frac{1}{K} \sum_{n=1}^{K} \frac{n}{A_1} \\
&= \frac{1}{KA_1} \frac{K(K+1)}{2} \\
&= \frac{K}{A_1} \frac{(K+1)}{K} \frac{1}{2} \\
&\approx \frac{\ln(K)}{2}.
\end{aligned}$$

Next, consider the case of aligned noise, when $W$ has a $A_\gamma/n^\gamma$ eigenspectrum and aligned eigenbases with $\rho$. In this case,

$$\begin{aligned}
\mathrm{tr}(\rho^{-1}W) &= \frac{A_\gamma}{A_1} \sum_{n=1}^{K} \frac{1}{n^{\gamma-1}} \\
&= \frac{A_\gamma}{A_1} \sum_{n=1}^{K} \frac{1}{n^{\gamma-1}} \\
&= \frac{A_\gamma}{A_1} \sum_{n=1}^{K} \frac{1}{n^{\gamma-1}} \\
&= \frac{A_\gamma}{K} \frac{K}{A_1} \frac{K}{A_{\gamma-1}} \\
&= \frac{A_\gamma}{K} \frac{K}{A_{\gamma-1}} \ln(K)
\end{aligned}$$

We can then derive the following approximations for $\mathrm{tr}(\rho^{-1}W)/K$ using Eqs. (116)–(118).

$$\begin{aligned}
\gamma > 2 &\implies \left(\frac{\gamma-1}{\gamma-2}\right)\frac{\ln(K)}{K} \\
\gamma = 2 &\implies \frac{1}{K} \\
2 > \gamma > 1 &\implies \left(\frac{\gamma-1}{2-\gamma}\right)\frac{\ln(K)}{K^{\gamma-1}} \\
\gamma = 1 &\implies 1 \\
1 > \gamma > 0 &\implies \left(\frac{1-\gamma}{2-\gamma}\right)\ln(K)
\end{aligned}$$

## B.7 Optimal homeostatic gains

We now consider the case where we enforce homeostasis on the gains. Prior to taking expectations across environments, our objective, considered as a function of $r_a = g_a\omega_a$ is (up to additive constants)

$$\mathcal{L} = \log\det\left(\sigma^2 W + r^{1-\alpha}Pr^{1-\alpha}\right) - \mu^{-1}\sum_{a=1}^{K} r_a,$$

where $P = \mathrm{CV}\rho\mathrm{CV}$. Enforcing homeostasis at the cluster level means setting $\hat{r}_a = \chi$ for all clusters. Substituting this in, we obtain the function

$$\begin{aligned}
\mathcal{L}(r = \chi\mathbf{1}_K) &= -K\chi/\mu + \log\det\Big(\sigma^2 W + \chi^{2(1-\alpha)}P\Big) \\
&= -K\chi/\mu + \log\det\big(I_K + \chi^{\beta}[\sigma^{-2}W^{-1}P]\big) + \text{const.} \\
&= -K\chi/\mu + \sum_n \log\left(1 + \chi^{\beta}\lambda_n\right) + \text{const.},
\end{aligned}$$

with $\lambda_n$ are the eigenvalues of $\sigma^{-2}W^{-1}P$. Note that this is only a function of the spectrum, and not of the eigenbasis. Therefore, under the assumption that the spectrum remains fixed across environments, we can drop the need to take expectations, and work with $\mathcal{L}$ as a function only of the spectrum. We work under this assumption going forward.

The optimal $\chi$, within this family of approximate solutions, obeys

$$\begin{aligned}
0 &= -K + \mu\sum_n \frac{\lambda_n\beta\chi^{\beta-1}}{1+\chi^{\beta}\lambda_n} \\
K &= \frac{\mu\beta}{\chi}\sum_n \frac{\lambda_n\chi^{\beta}}{1+\chi^{\beta}\lambda_n} \\
\chi &= \frac{\mu\beta}{K}\sum_n \left(1 - \frac{1}{1+\chi^{\beta}\lambda_n}\right) \\
\chi &= \mu\beta\left(1 - \frac{1}{K}\sum_n \frac{1}{1+\chi^{\beta}\lambda_n}\right).
\end{aligned} \tag{119}$$

We now demonstrate that, under appropriate conditions, $\chi \approx \frac{1}{K}\sum_a g_a^{(1)}\omega_a$.

$$\begin{aligned}
g_a^{(1)}\omega_a &= \mu\beta(1 - \Delta_{aa}) \\
&= \mu\beta\left(1 - \frac{\sigma^2}{(\beta\mu)^{\beta}}[P^{-1}W]_{aa}\right) \\
\sum_a g_a^{(1)}\omega_a &= \mu\beta\left(1 - \frac{1}{(\beta\mu)^{\beta}}\sum_a [\sigma^2 P^{-1}W]_{aa}\right) \\
&= \mu\beta\left(1 - \frac{1}{(\beta\mu)^{\beta}}\mathrm{tr}\left(\sigma^2 P^{-1}W\right)\right) \\
&= \mu\beta\left(1 - \frac{1}{(\beta\mu)^{\beta}}\sum_n \frac{1}{\lambda_n}\right) \\
\frac{1}{K}\sum_a g_a^{(1)}\omega_a &= \mu\beta\left(1 - \frac{1}{K}\sum_n \frac{1}{(\beta\mu)^{\beta}\lambda_n}\right)
\end{aligned}$$

Now, under the assumption that $(\beta\mu)^\beta\lambda_n \gg 1$ for each $n$, we can say that $\chi \approx \mu\beta$ and $(\beta\mu)^\beta\lambda_n \approx \chi^\beta\lambda_n + 1$. Substituting this in gives us the required result. Note that this condition is equivalent to the high signal-to-noise ratio condition we have used throughout.

We now consider the special cases discussed in Sec. 2.2 in which more precise solutions can be obtained. Note that in all three cases the analytics correspond to idealisations of the numerical simulations we actually perform in Sec. 2.6.

**Uncorrelated power-law noise** In the first special case under consideration, $W = I_K$, $\mathrm{CV}^2$ is constant, and $\rho$ has approximately a $A_1/n$ eigenspectrum, where $A_1$ is chosen to normalise the trace of $\rho$ to be equal to $K$, *i.e.*,

$$A_1 = \frac{K}{\sum_{n=1}^{K} \frac{1}{n}}$$

The spectrum of $\sigma^{-2}W^{-1}P$ is therefore $\lambda_n = b/n$ where

$$b = \sigma^{-2}A_1\mathrm{CV}^2 \tag{120}$$

In this case, (defining $u = n/K$) we can make the following approximation to the right hand side of Eq. (119):

$$\begin{aligned}
1 - \frac{1}{K}\sum_{n=1}^{K}\frac{1}{1+\chi^\beta\lambda_n} &= \frac{1}{K}\sum_{n=1}^{K}\frac{\chi^\beta\lambda_n}{1+\chi^\beta\lambda_n} \\
&= \frac{1}{K}\sum_{n=1}^{K}\frac{\frac{b\chi^\beta}{n}}{1+\frac{b\chi^\beta}{n}} \\
&\approx \int_{1/K}^{1}\frac{\frac{b\chi^\beta}{Ku}}{1+\frac{b\chi^\beta}{Ku}}du \\
&\approx \frac{b}{K}\chi^\beta\int_0^1\frac{1}{u+\frac{b}{K}\chi^\beta}du \\
&= \frac{b}{K}\chi^\beta\ln\left(\frac{1+\frac{b}{K}\chi^\beta}{\frac{b}{K}\chi^\beta}\right) \\
&= \frac{b}{K}\chi^\beta\ln\left(1+\frac{K}{b\chi^\beta}\right).
\end{aligned}$$

This gives the new equation

$$\chi = \beta\mu\frac{b}{K}\chi^\beta\ln\left(1+\frac{K}{b\chi^\beta}\right). \tag{121}$$

Approximating $A_1 \approx K/\ln(K)$ (see App. B.6) and using Eq. (120), we obtain that

$$\frac{K}{b} \approx \frac{\sigma^2\ln(K)}{\mathrm{CV}^2}. \tag{122}$$

Substituting Eq. (122) into Eq. (121) gives us

$$\frac{\chi}{\beta\mu}\left(\frac{\sigma^2\ln(K)}{\chi^\beta\mathrm{CV}^2}\right) = \ln\left(1+\frac{\sigma^2\ln(K)}{\chi^\beta\mathrm{CV}^2}\right). \tag{123}$$

**Aligned noise** In the aligned noise case, we approximate $\rho$ and $W$ as having the same eigenbasis. $\rho$ still has an eigenspectrum of $A_1/n$, and we take $W$ to have an eigenspectrum of $A_\gamma/n^\gamma$ where $A_\gamma$ normalises the trace of $W$ to be equal to $K$,

$$A_\gamma = \frac{K}{\sum_{n=1}^{K}\frac{1}{n^\gamma}}.$$

The matrix $\sigma^2W^{-1}P$ therefore has eigenspectrum $bn^{\gamma-1}$ where $b = \sigma^{-2}A_1\mathrm{CV}^2/A_\gamma$. Inserting this into Eq. (119) gives us

$$\chi = \mu\beta\left(1 - \frac{1}{K}\sum_n\frac{1}{1+b\chi^\beta n^{\gamma-1}}\right). \tag{124}$$

See App. B.6 for estimates of $A_\gamma$.

**Constant correlation noise** The next special case occurs when $\beta = 1$, and $W = (1-p)I_K + p\mathbf{1}\mathbf{1}^{\mathrm{T}}$. Using the Sherman-Morrison formula, we obtain

$$W^{-1} = \frac{1}{1-p}\left(I_K - \frac{p}{1+p(K-1)}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right).$$

Since $\frac{p}{1+p(K-1)} \leq \frac{1}{K-1}$ we neglect this term and approximate $W^{-1} \approx \frac{1}{1-p}I_K$. This therefore has the same effect as making the replacement $\sigma^2 \mapsto (1-p)\sigma^2$. Substituting this into Eq. (123), and using $\beta = 1$, we obtain

$$\frac{\sigma^2(1-p)\ln(K)}{\mu \mathrm{CV}^2} = \ln\left(1 + \frac{\sigma^2(1-p)\ln(K)}{\chi \mathrm{CV}^2}\right). \tag{125}$$

We define

$$q = \frac{\sigma^2(1-p)}{\mu \mathrm{CV}^2},$$

and rearrange to get

$$\begin{aligned} q\ln(K) &= \ln\left(1 + \frac{\mu q \ln(K)}{\chi}\right) \\ K^q - 1 &= \frac{\mu q \ln(K)}{\chi} \\ \chi &= \mu \frac{q\ln(K)}{K^q - 1}. \end{aligned}$$

## B.8 Homeostatic propagation

In this appendix, we derive how the weights between neural populations must change in order for homeostasis to be propagated between them. We work in a linear rate model.

We start by considering a downstream population of tuning curves $h_m^{(2)}(\boldsymbol{s})$ for $m = 1, \ldots, M$, with $W_{ma}$ the synaptic weight from neuron $a$ to neuron $m$ in the downstream population. The tuning curve of neuron $m$ is

$$h_m^{(2)}(\boldsymbol{s}) = \sum_{a=1}^{K} W_{ma} h_a^{(1)}(\boldsymbol{s}) \tag{126}$$

We assume the downstream population has representational curves $\Omega_m^{(2)}(\boldsymbol{s})$ for $m = 1, \ldots, M$ given by

$$\Omega_m^{(2)}(\boldsymbol{s}) = \sum_{a=1}^{K} w_{ma} \Omega_a^{(1)}(\boldsymbol{s}). \tag{127}$$

If the downstream population is also implementing homeostatic coding, we know that

$$h_m^{(2)}(\boldsymbol{s}) = \frac{\chi \Omega_m^{(2)}(\boldsymbol{s})}{\omega_m^{(2)}},$$

where $\omega_m^{(2)} = \mathbb{E}\left[\Omega_m^{(2)}(\boldsymbol{s})\right]$. But then

$$\begin{aligned} \omega_m^{(2)} &= \mathbb{E}\left[\Omega_m^{(2)}(\boldsymbol{s})\right] \\ &= \sum_{b=1}^{K} w_{mb} \mathbb{E}\left[\Omega_b^{(1)}(\boldsymbol{s})\right] \\ &= \sum_{b=1}^{K} w_{mb} \omega_b^{(1)} \\ \chi\Omega_m^{(2)}(\boldsymbol{s}) &= \sum_{a=1}^{K} w_{ma} \chi \Omega_a^{(1)}(\boldsymbol{s}) \\ &= \sum_{a=1}^{K} w_{ma} \omega_a^{(1)} h_a^{(1)}(\boldsymbol{s}). \end{aligned}$$

Substituting this in, we get

$$h_m^{(2)}(\boldsymbol{s}) = \frac{\sum_{a=1}^K w_{ma}\omega_a^{(1)} h_a^{(1)}(\boldsymbol{s})}{\sum_{b=1}^K w_{mb}\omega_b^{(1)}}. \tag{128}$$

Comparing coefficients, we can see that

$$W_{ma} = \frac{w_{ma}\omega_a^{(1)}}{\sum_{b=1}^K w_{mb}\omega_b^{(1)}},$$

as claimed.

## B.9 Hierarchical Bayes-ratio coding

In this appendix, we calculate synaptic weights for propagation of Bayes-ratio coding between populations. We start with a generative model $\boldsymbol{z}^{(2)} \to \boldsymbol{z}^{(1)} \to \boldsymbol{s}$. The downstream representational curves are

$$\Omega_m^{(2)}(\boldsymbol{s}) = \Pi\left(\boldsymbol{z}^{(2)} = \boldsymbol{z}_m^{(2)}|\boldsymbol{s}\right).$$

We recall that the synaptic weights $W_{ma}$ are given by the Eq. (25), repeated here:

$$W_{ma} = \frac{w_{ma}\omega_a^{(1)}}{\sum_{b=1}^K w_{mb}\omega_b^{(1)}},$$

where $w_{ma}$ are the coefficients of the expansion $\Omega_m^{(2)}(\boldsymbol{s}) = \sum_a w_{ma}\Omega_a^{(1)}(\boldsymbol{s})$. We start by calculating $w_{ma}$.

$$\begin{aligned}
\Omega_m^{(2)}(\boldsymbol{s}) &= \Pi\left(\boldsymbol{z}^{(2)} = \boldsymbol{z}_m^{(2)}|\boldsymbol{s}\right) \\
&= \int g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}^{(1)}\right)\Pi\left(\boldsymbol{z}^{(1)}|\boldsymbol{s}\right)d\boldsymbol{z}^{(1)} \\
&\approx \sum_{a=1}^K g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_a^{(1)}\right)\Pi\left(\boldsymbol{z}^{(1)} = \boldsymbol{z}_a^{(1)}|\boldsymbol{s}\right)\delta\boldsymbol{z}_a^{(1)} \\
&= \sum_{a=1}^K g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_a^{(1)}\right)\delta\boldsymbol{z}_a^{(1)}\Omega_a^{(1)}(\boldsymbol{s}) \\
w_{ma} &= g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_a^{(1)}\right)\delta\boldsymbol{z}_a^{(1)}.
\end{aligned}$$

We plug this into our formula, and use additionally that $\omega_b^{(1)} = \pi\left(\boldsymbol{z}_b^{(1)}\right)$. This gives us

$$\begin{aligned}
\sum_{b=1}^K w_{mb}\omega_b^{(1)} &= \sum_{b=1}^K g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_b^{(1)}\right)\delta\boldsymbol{z}_b^{(1)}\pi\left(\boldsymbol{z}_b^{(1)}\right) \\
&\approx \int g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}^{(1)}\right)\pi\left(\boldsymbol{z}^{(1)}\right)d\boldsymbol{z}^{(1)} \\
&= \pi\left(\boldsymbol{z}_m^{(2)}\right) \\
W_{ma} &= \frac{w_{ma}\omega_a^{(1)}}{\sum_{b=1}^K w_{mb}\omega_b^{(1)}} \\
&\approx \frac{g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_a^{(1)}\right)\delta\boldsymbol{z}_a^{(1)}\pi\left(\boldsymbol{z}_a^{(1)}\right)}{\pi\left(\boldsymbol{z}_m^{(2)}\right)} \\
&= g\left(\boldsymbol{z}_a^{(1)}|\boldsymbol{z}_m^{(2)}\right)\delta\boldsymbol{z}_a^{(1)},
\end{aligned}$$

which gives the result as required.

## B.10 Stimulus specific adaptation for non-ideal-observer models

In this appendix, we show how discrepancies between the internal model and the external environment lead to simultaneous stimulus specific and neuron specific adaptation effects in a homeostatic DDC code. We begin with the special case of a Bayes-ratio code, which is a DDC in which the kernel functions are delta functions (as discussed in Sec. 2.9). In a Bayes-ratio code, representational curves are given by

$$\Omega_a(\boldsymbol{s}) = \Pi(\boldsymbol{z}_a|\boldsymbol{s}), \tag{129}$$

where $\Pi$ is the posterior distribution under the internal generative model. We consider the case of a non-ideal-observer generative model, in which case there will be a discrepancy between the marginal stimulus distribution predicted by the model, $P^I(\boldsymbol{s})$ and the external environment stimulus marginal $P^E(\boldsymbol{s})$. We reason as follows:

$$\begin{aligned}\Omega_a(\boldsymbol{s}) &= \Pi^I(\boldsymbol{z}_a|\boldsymbol{s}) \\ &= \frac{f(\boldsymbol{s}|\boldsymbol{z}_a)\pi^I(\boldsymbol{z}_a)}{P^I(\boldsymbol{s})}\end{aligned} \tag{130}$$

where

$$\begin{aligned}P^I(\boldsymbol{s}) &\equiv \int f(\boldsymbol{s}|\boldsymbol{z})\pi^I(\boldsymbol{z})d\boldsymbol{z} \\ &\approx \sum_a f(\boldsymbol{s}|\boldsymbol{z}_a)\pi^I(\boldsymbol{z}_a)\Delta\boldsymbol{z}\end{aligned} \tag{131}$$

According to homeostatic coding, the tuning curve is given by

$$h_a(\boldsymbol{s}) = \chi\frac{\Omega_a(\boldsymbol{s})}{\omega_a}, \tag{132}$$

where $\omega_a$ is the temporal average of the representational curve, or equivalently, the expectation of $\Omega_a(\boldsymbol{s})$ under the *true* environmental stimulus distribution. Thus

$$\begin{aligned}\omega_a &= \mathbb{E}^E[\Omega_a(\boldsymbol{s})] \\ &= \int \frac{f(\boldsymbol{s}|\boldsymbol{z}_a)\pi^I(\boldsymbol{z}_a)}{P^I(\boldsymbol{s})}P^E(\boldsymbol{s})d\boldsymbol{s} \\ &= \pi^I(\boldsymbol{z}_a)F_a\end{aligned} \tag{133}$$

where we defined

$$F_a := \int f(\boldsymbol{s}|\boldsymbol{z}_a)\frac{P^E(\boldsymbol{s})}{P^I(\boldsymbol{s})}d\boldsymbol{s}. \tag{134}$$

Notice that when $P^E = P^I$, as would be the case for an ideal-observer, we will have $F_a = 1$, due to the normalization of $f(\boldsymbol{s}|\boldsymbol{z})$. Substituting Eq. (130) and Eq. (133) into Eq. (132) we obtain

$$h_a(\boldsymbol{s}) = \chi\frac{f(\boldsymbol{s}|\boldsymbol{z}_a)}{P^I(\boldsymbol{s})F_a} \tag{135}$$

as claimed in Eq. (32).

We now consider the more general case of a DDC code. We define

$$F(\boldsymbol{z}) = \int f(\boldsymbol{s}|\boldsymbol{z})\frac{P^E(\boldsymbol{s})}{P^I(\boldsymbol{s})}d\boldsymbol{s}, \tag{136}$$

so that $F_a$ in the above is equal to $F(\boldsymbol{z}_a)$. An analogous derivation yields:

$$h_a(\boldsymbol{s}) = \chi\frac{\int \phi_a(\boldsymbol{z})\pi^I(\boldsymbol{z})f(\boldsymbol{s}|\boldsymbol{z})d\boldsymbol{z}}{P^I(\boldsymbol{s})\int \phi_a(\boldsymbol{z})\pi^I(\boldsymbol{z})F(\boldsymbol{z})d\boldsymbol{z}}. \tag{137}$$

We can see that in this case there is no clean separation of neuron specific and stimulus specific factors. In particular, the generalization of Eq. (34) takes the form:

$$\frac{h_a(\boldsymbol{s};v=1)}{h_a(\boldsymbol{s};v=0)} = \frac{P^I(\boldsymbol{s};v=0)}{P^I(\boldsymbol{s};v=1)} \times \frac{\int \phi_a(\boldsymbol{z})\pi^I(\boldsymbol{z};v=1)f(\boldsymbol{s}|\boldsymbol{z})d\boldsymbol{z}}{\int \phi_a(\boldsymbol{z})\pi^I(\boldsymbol{z};v=0)f(\boldsymbol{s}|\boldsymbol{z})d\boldsymbol{z}} \times \frac{\int \phi_a(\boldsymbol{z})\pi^I(\boldsymbol{z};v=0)F(\boldsymbol{z};v=0)d\boldsymbol{z}}{\int \phi_a(\boldsymbol{z})\pi^I(\boldsymbol{z};v=1)F(\boldsymbol{z};v=1)d\boldsymbol{z}}. \tag{138}$$

In the limit as the width of the kernels becomes small, the integrals involving kernels collapse to sampling at a single point, as occurs with the Bayes-ratio code. When this happens, we get full separation to stimulus and neuron specific effects; otherwise these effects are mixed together, with more mixing occurring the larger the width of the kernels. However, when the kernels are unimodal and sufficiently narrow, we would expect an approximate factorization of the effect of adaptation into a stimulus-specific and a neuron-specific suppression factor.

## References


Atick, J. J. and Redlich, A. N. (1990). Towards a Theory of Early Visual Processing. *Neural Computation*, 2(3):308–320.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193. Place: US Publisher: American Psychological Association.

Barlow, H. B. (2012). Possible Principles Underlying the Transformations of Sensory Messages. In Rosenblith, W. A., editor, *Sensory Communication*, pages 216–234. The MIT Press.

Beck, J., Ma, W. J., Latham, P. E., and Pouget, A. (2007). Probabilistic population codes and the exponential family of distributions. In Cisek, P., Drew, T., and Kalaska, J. F., editors, *Progress in Brain Research*, volume 165 of *Computational Neuroscience: Theoretical Insights into Brain Function*, pages 509–519. Elsevier.

Benucci, A., Saleem, A. B., and Carandini, M. (2013). Adaptation maintains population homeostasis in primary visual cortex. *Nature Neuroscience*, 16(6):724–729.

Bethge, M., Rotermund, D., and Pawelzik, K. (2002). Optimal short-term population coding: when Fisher information fails. *Neural Computation*, 14(10):2317–2351.

Brunel, N. and Nadal, J.-P. (1998). Mutual Information, Fisher Information, and Population Coding. *Neural Computation*, 10(7):1731–1757.

Buzsáki, G. and Mizuseki, K. (2014). The log-dynamic brain: how skewed distributions affect network operations. *Nature Reviews. Neuroscience*, 15(4):264–278.

Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51–62.

Clifford, C. W. G., Webster, M. A., Stanley, G. B., Stocker, A. A., Kohn, A., Sharpee, T. O., and Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research*, 47(25):3125–3131.

Desai, N. (2003). Homeostatic plasticity in the CNS: Synaptic and intrinsic forms. *Journal of physiology, Paris*, 97:391–402.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5.

Dragoi, V., Sharma, J., Miller, E. K., and Sur, M. (2002). Dynamics of neuronal sensitivity in visual cortex and local feature discrimination. *Nature Neuroscience*, 5(9):883–891.

Dragoi, V., Sharma, J., and Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron*, 28(1):287–298.

Ganguli, D. and Simoncelli, E. P. (2014). Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations. *Neural Computation*, 26(10):2103–2134.

Gershon, E. D., Wiener, M. C., Latham, P. E., and Richmond, B. J. (1998). Coding Strategies in Monkey V1 and Inferior Temporal Cortices. *Journal of Neurophysiology*, 79(3):1135–1144. Publisher: American Physiological Society.

Geurts, L. S., Cooke, J. R. H., van Bergen, R. S., and Jehee, J. F. M. (2022). Subjective confidence reflects representation of Bayesian probability in cortex. *Nature Human Behaviour*, 6(2):294–305. Number: 2 Publisher: Nature Publishing Group.

Ghisovan, N., Nemri, A., Shumikhina, S., and Molotchnikoff, S. (2009). Long adaptation reveals mostly attractive shifts of orientation tuning in cat primary visual cortex. *Neuroscience*, 164(3):1274–1283. Publisher: Elsevier.

Goris, R. L. T., Movshon, J. A., and Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, 17(6):858–865.

Greenberg, D. S., Houweling, A. R., and Kerr, J. N. D. (2008). Population imaging of ongoing neuronal activity in the visual cortex of awake rats. *Nature Neuroscience*, 11(7):749–751.

Hengen, K. B., Lambo, M. E., Van Hooser, S. D., Katz, D. B., and Turrigiano, G. G. (2013). Firing Rate Homeostasis in Visual Cortex of Freely Behaving Rodents. *Neuron*, 80(2):335–342.

Kanitscheider, I., Coen-Cagli, R., and Pouget, A. (2015). Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences of the United States of America*, 112(50):E6973–6982.

Keck, T., Keller, G. B., Jacobsen, R. I., Eysel, U. T., Bonhoeffer, T., and Hübener, M. (2013). Synaptic Scaling and Homeostatic Plasticity in the Mouse Visual Cortex In Vivo. *Neuron*, 80(2):327–334.

Kersten, D., Mamassian, P., and Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, 55:271–304.

Kohn, A. (2007). Visual adaptation: physiology, mechanisms, and functional benefits. *Journal of Neurophysiology*, 97(5):3155–3164.

Koyama, S. (2015). On the Spike Train Variability Characterized by Variance-to-Mean Power Relationship. *Neural Computation*, 27(7):1530–1548.

Laughlin, S. (1981). A Simple Coding Procedure Enhances a Neuron's Information Capacity. *Zeitschrift für Naturforschung C*, 36(9-10):910–912. Publisher: De Gruyter.

Lennie, P. (2003). The Cost of Cortical Computation. *Current Biology*, 13(6):493–497.

Levy, W. and Baxter, R. (1996). Energy Efficient Neural Codes. *Neural computation*, 8:531–43.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3):105–117. Conference Name: Computer.

Maffei, A. and Turrigiano, G. G. (2008). Multiple Modes of Network Homeostasis in Visual Cortical Layer 2/3. *The Journal of Neuroscience*, 28(17):4377–4384.

Marder, E. and Prinz, A. A. (2003). Current compensation in neuronal homeostasis. *Neuron*, 37(1):2–4.

Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014). Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417.

Moshitch, D. and Nelken, I. (2014). Using Tweedie distributions for fitting spike count data. *Journal of Neuroscience Methods*, 225:13–28.

Movshon, J. A. and Lennie, P. (1979). Pattern-selective adaptation in visual cortical neurones. *Nature*, 278(5707):850–852. Number: 5707 Publisher: Nature Publishing Group.

Müller, J. R., Metha, A. B., Krauskopf, J., and Lennie, P. (1999). Rapid adaptation in visual cortex to the structure of images. *Science*, 285(5432):1405–1408. Place: US Publisher: American Assn for the Advancement of Science.

Nadal, J.-P. and Parga, N. (1994). Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5(4):565–581. Publisher: Taylor & Francis _eprint: https://doi.org/10.1088/0954-898X_5_4_008.

Nadal, J.-P. and Parga, N. (1999). Sensory coding: information maximization and redundancy reduction. In *Neuronal Information Processing*, volume Volume 7 of *Series in Mathematical Biology and Medicine*, pages 164–171. WORLD SCIENTIFIC.

Parker, P. R. L., Abe, E. T. T., Leonard, E. S. P., Martins, D. M., and Niell, C. M. (2022). Joint coding of visual input and eye/head position in V1 of freely moving mice. *Neuron*, 110(23):3897–3906.e5.

Rumyantsev, O. I., Lecoq, J. A., Hernandez, O., Zhang, Y., Savall, J., Chrapkiewicz, R., Li, J., Zeng, H., Ganguli, S., and Schnitzer, M. J. (2020). Fundamental bounds on the fidelity of sensory cortical coding. *Nature*, 580(7801):100–105. Number: 7801 Publisher: Nature Publishing Group.

Sanzeni, A., Palmigiano, A., Nguyen, T. H., Luo, J., Nassi, J. J., Reynolds, J. H., Histed, M. H., Miller, K. D., and Brunel, N. (2023). Mechanisms underlying reshuffling of visual responses by optogenetic stimulation in mice and monkeys. *Neuron*, 111(24):4102–4115.e9.

Schwartz, O., Hsu, A., and Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, 8(7):522–535. Number: 7 Publisher: Nature Publishing Group.

Shadlen, M. N. and Newsome, W. T. (1998). The Variable Discharge of Cortical Neurons: Implications for Connectivity, Computation, and Information Coding. *The Journal of Neuroscience*, 18(10):3870–3896.

Simoncelli, E. P. and Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, 24(1):1193–1216.

Slomowitz, E., Styr, B., Vertkin, I., Milshtein-Parush, H., Nelken, I., Slutsky, M., and Slutsky, I. (2015). Interplay between population firing stability and single neuron dynamics in hippocampal networks. *eLife*, 4:e04378.

Snow, M., Coen-Cagli, R., and Schwartz, O. (2016). Specificity and timescales of cortical adaptation as inferences about natural movie statistics. *Journal of Vision*, 16(13):1.

Solomon, S. G. and Kohn, A. (2014). Moving sensory adaptation beyond suppressive effects in single neurons. *Current biology: CB*, 24(20):R1012–1022.

Stocker, A. A. and Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585. Number: 4 Publisher: Nature Publishing Group.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. (2019). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365. Number: 7765 Publisher: Nature Publishing Group.

Szuts, T. A., Fadeyev, V., Kachiguine, S., Sher, A., Grivich, M. V., Agrochão, M., Hottowy, P., Dabrowski, W., Lubenov, E. V., Siapas, A. G., Uchida, N., Litke, A. M., and Meister, M. (2011). A wireless multi-channel neural amplifier for freely moving animals. *Nature Neuroscience*, 14(2):263–269.

Torrado Pacheco, A., Tilden, E. I., Grutzner, S. M., Lane, B. J., Wu, Y., Hengen, K. B., Gjorgjieva, J., and Turrigiano, G. G. (2019). Rapid and active stabilization of visual cortical firing rates across light–dark transitions. *Proceedings of the National Academy of Sciences*, 116(36):18068–18077.

Turrigiano, G. G. (2008). The self-tuning neuron: synaptic scaling of excitatory synapses. *Cell*, 135(3):422–435.

Turrigiano, G. G., Leslie, K. R., Desai, N. S., Rutherford, L. C., and Nelson, S. B. (1998). Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391(6670):892–896.

Turrigiano, G. G. and Nelson, S. B. (2004). Homeostatic plasticity in the developing nervous system. *Nature Reviews Neuroscience*, 5(2):97–107. Number: 2 Publisher: Nature Publishing Group.

van Bergen, R. S., Ji Ma, W., Pratte, M. S., and Jehee, J. F. M. (2015). Sensory uncertainty decoded from visual cortex predicts behavior. *Nature Neuroscience*, 18(12):1728–1730. Number: 12 Publisher: Nature Publishing Group.

Vertes, E. and Sahani, M. (2018). Flexible and accurate inference and learning for deep generative models. arXiv:1805.11051 [cs, stat].

Wandell, B. A. (1995). *Foundations of vision*. Foundations of vision. Sinauer Associates, Sunderland, MA, US. Pages: xvi, 476.

Wei, X.-X. and Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, 18(10):1509–1517.

Wei, X.-X. and Stocker, A. A. (2016). Mutual Information, Fisher Information, and Efficient Coding. *Neural Computation*, 28(2):305–326.

Westrick, Z. M., Heeger, D. J., and Landy, M. S. (2016). Pattern Adaptation and Normalization Reweighting. *Journal of Neuroscience*, 36(38):9805–9816.

Yarrow, S., Challis, E., and Seriès, P. (2012). Fisher and Shannon Information in Finite Neural Populations. *Neural Computation*, 24(7):1740–1780.