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

Neurons are typically sensitive to a small fraction of stimulus space. If the environment changes, making certain stimuli more prevalent, neurons sensitive to those stimuli would respond more often and therefore have a higher average firing rate if the stimulus-response mapping remains fixed. However, sufficiently prolonged exposure to the new environment typically causes such neurons to *adapt* by responding less vigorously. If adaptation consistently returns the average firing rate of neurons, or populations of similarly tuned neurons, to its value prior to environmental shift, it is termed *firing-rate homeostasis*. Another feature of adaptation in sensory cortex is *stimulus specific adaptation*, under which neurons not only adapt their responsiveness, but also reshape their tuning curves away from overrepresented stimuli. Here, we present a normative explanation of firing-rate homeostasis grounded in the efficient coding principle. Unlike previous theories based on efficient coding, we formulate the problem in a computation-agnostic manner, enabling our framework to apply far from the sensory periphery. We show that homeostasis can provide an optimal solution to a trade-off between coding fidelity and the metabolic cost of neural firing. We provide quantitative conditions necessary for the optimality of firing-rate homeostasis, and predict how adaptation should deviate from homeostasis when these conditions are violated. Based on biological estimates of relevant parameters, we show that these conditions do hold in areas of cortex where homeostatic adaptation has been observed. Finally, we apply our framework to distributed distributional codes, a specific computational theory of neural representations serving Bayesian inference. We show that the resultant coding scheme can be accomplished by divisive normalisation with adaptive weights. We further demonstrate how homeostatic coding, coupled with such Bayesian neural representations, explains stimulus-specific adaptation, as observed, *e.g.*, in the primary visual cortex.

1 Introduction

Neurons act as noisy feature detectors. Moreover, when the responsiveness 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 Poisson-like firing, for example, response noise grows as the square root of the mean response. The neuron can therefore increase its signal-to-noise ratio by increasing its responsiveness or response gain. However, this increase in coding fidelity comes at the cost of an elevated firing rate, and thus higher metabolic energy expenditure. From a normative perspective, coding fidelity and metabolic cost are thus two conflicting forces. In this paper, we ask how neurons should adjust their response gains in order to optimally balance these two forces, *irrespective* of the computations they perform or the complexity of the features they detect. In particular, we address the following question: given a neural population with a set of arbitrary tuning curve shapes, how should neurons optimally adjust their gains depending on the stimulus statistics prevailing in the environment?

Suppose an environmental shift makes a feature to which a group of neurons is sensitive more prevalent. Over time, that feature will be presented more often, in turn increasing the average firing rate of these neurons. Typically, the neurons will then adapt by reducing their responsiveness (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 up to the rate prior to the environmental shift. Thus, under firing rate homeostasis, neuronal populations maintain a constant stimulus-averaged firing rate in the face of changes to the environment (Benucci et al., 2013; Hengen et al., 2013; Maffei and Turrigiano, 2008).

There are multiple levels at which homeostatic results 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). 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 preferences remains stable under environmental shifts, but the firing rate of individual neurons within a cluster can change (Benucci et al., 2013). Lastly, homeostasis can occur at the level of individual neurons, in which case the firing rate of each individual neuron is kept constant under shifts in the environment (Marder and Prinz, 2003). Note that these three forms are progressively stronger.

Previous normative explanations for firing rate homeostasis often focus on the necessity of avoiding hypo- or hyper-excited states (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. Our task in this paper is therefore to provide a normative account of homeostasis at levels below that of an entire population with heterogeneous stimulus tuning.

Specifically, our theory provides a normative account of firing rate homeostasis at the level of clusters of similarly tuned neurons. Furthermore, as we will show, when coupled with a specific computational theory of sensory representations (which determines tuning curve shapes), it further predicts the typical finding that such adaptations are stimulus specific and shift tuning curves away from more prevalent stimuli. In *stimulus specific adaptation* a neuron may adapt by a greater reduction (or possibly an increase) in responsiveness to some stimuli than others. In particular, the suppression of firing rate is greater for test stimuli that are closer to an over-represented adaptor stimulus than for test stimuli that are further away (Kohn, 2007; Schwartz et al., 2007). In the case of a one-dimensional stimulus space, this causes a repulsion of tuning curves away from the adaptor.

To address our normative question regarding the optimal adjustment of responsiveness, we work within the framework of efficient coding theory (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 theory begins by asking two questions: What is this neural system attempting to encode, and what features of biology constrain the quality of its encoding? We can then compute (analytically or numerically) the optimal encoding subject to the posited biological constraints and costs. Under the assumption that natural selection has acted to optimise performance, this solution serves as a prediction for encoding properties of real neurons.

Concretely, the *Infomax Principle* (Linsker, 1988; Ganguli and Simoncelli, 2014; Wei and Stocker, 2015; Atick and Redlich, 1990) states that sensory systems optimise the mutual information between a noisy neural representation and an external stimulus, subject to metabolic constraints. In order to do so, the neural system should exploit the statistics of its local environment (Simoncelli and Olshausen, 2001). A non-adaptive system that performs well in one environment may perform poorly in an environment with different stimulus statistics. Efficient coding theories can therefore be used, in particular, to predict how sensory systems should *adapt* to changes in the environment statistics (Wei and Stocker, 2015).

We use the mutual information between the stimulus and the neural response as a measure of coding fidelity, and the expected number of spikes fired as a measure of metabolic cost. We show that when neurons are highly selective and population responses have a high-dimensional geometry, optimal gains yield firing rate homeostasis across environments; in other words, average firing rates are maintained due to optimal gain adjustments, despite changes in the environmental statistics. Our work complements and extends previous theoretical work, in particular that of Ganguli and Simoncelli (2014). A novelty of our framework is that we formulate the problem in a way that is agnostic of the computations that the population is designed to perform, which dictate the shapes and arrangement of its tuning curves. We thus allow the population to have tuning curves of arbitrary configuration and shape, possibly defined on a high-dimensional stimulus, but do not optimise their shapes or placements. Ganguli and Simoncelli, on the other hand, did optimise tuning curve widths and placements, but only for the case of a one-dimensional stimulus and a homogeneous (up to stimulus reparametrisation) population of unimodal or sigmoidal tuning curves. Additionally, our framework allows us to consider more general noise models than simple Poissonian firing, in particular correlated and power-law noise.

Having shown the optimality of homeostatic codes without making assumptions on the nature of the neural representation and computation, we apply our framework to a Bayesian theory of neural computation, namely the *Distributed Distributional Code (DDC)* (Vertes and Sahani, 2018). The theory of DDCs assumes that the brain possesses an internal generative model of sensory input arising from unobserved latent variables. The firing rates of neurons are then assumed to directly represent the Bayesian posterior expectation of a rich set of (fixed) functions of the latent variables. Combining the theory of DDCs with our normative results yields what we call a *homeostatic DDC*. Homeostatic DDCs are able to account for stimulus specific adaptation effects which cannot be fully accounted for in previous efficient coding frameworks (Wei and Stocker, 2015; Ganguli and Simoncelli, 2014; Snow et al., 2016). A special case of a homeostatic DDC is a form of representation 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 the basic mathematical framework we use to address our normative question. This includes formulating the notion of neuron clustering based on stimulus selectivity. We then show that our framework predicts that the firing rate of each cluster should remain constant subject to shifts in environmental statistics, with individual neurons free to shuffle their average rates. We additionally demonstrate that our framework can account for a wide distribution of mean single-neuron firing rates as observed in cortex. We show that the conditions necessary for our theory to apply are expected to hold in cortical areas, in particular in the primary visual cortex (V1). We then extend our analysis to apply to the cases of correlated and power-law noise, and numerically validate the quality of our homeostatic solutions. Lastly, we apply our theory to DDC representational codes, showing how homeostatic DDC's can account for stimulus specific adaptation effects observed experimentally.

2 Results

2.1 Theoretical framework

We consider a population of N neurons, responding to the (possibly high-dimensional) stimulus $\boldsymbol{s}$, with marginal distribution $P(\boldsymbol{s})$. The distribution $P(\boldsymbol{s})$ is determined by the environment; accordingly, if the environment changes (making certain stimuli more or less prevalent), so will P. 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 $\boldsymbol{n} = (n_1, \ldots, n_N)$.

We are interested in how changes in the stimulus distribution, P, affect the responsiveness of neurons.

We therefore adopt a shape-amplitude decomposition of the neural tuning curves. The tuning curve of the i-th neuron, $h_i(\boldsymbol{s})$, is factorised into a *representational curve*, $\Omega_i(\boldsymbol{s})$, and a *gain*, g_i:

$$h_i(\boldsymbol{s}) = g_i \Omega_i(\boldsymbol{s}) = \text{gain} \times \text{representational curve}.$$

Fig. 1 (a) demonstrates the effect of changing the gain while keeping the representational curve constant. Importantly, we do not make any assumptions on the shape of the representational curve: Ω_i can be any complex (*e.g.* multi-modal, discontinuous) function of the possibly high-dimensional stimulus, and can thus represent *any* computation. This makes our treatment 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. In particular, this generality enables our theory to apply to populations located deep in the processing pathway, and not just to primary sensory neurons.

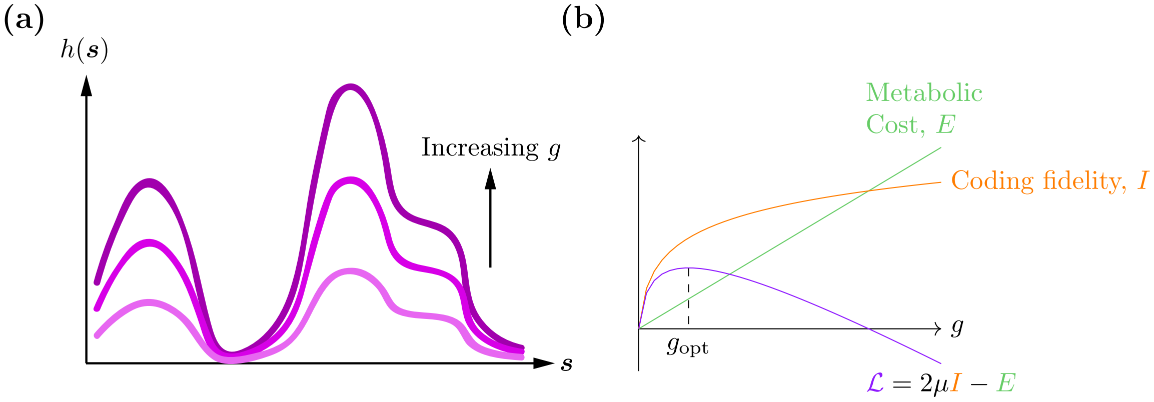

Figure 1: (a) As the gain, g, of a neuron is increased, the shape of its tuning curve remains the same, but all firing rates are scaled upwards. The cartoon shows a one-dimensional example, but our theory applies to general tuning curve shapes and joint configurations of population tuning curves, on high-dimensional stimulus spaces. (b) Cartoon representation of our objective function. Neural gains will be chosen to maximise an objective function, $\mathcal{L}$, which is the weighted difference between mutual information, I, which captures coding fidelity, and metabolic cost, E, given by average population firing rate.

Neural responses $\boldsymbol{n}$ are taken to be noisy encoding of neural tuning curves $\boldsymbol{h}(\boldsymbol{s})$, $\boldsymbol{n}|\boldsymbol{s} \sim P_{\text{noise}}(\boldsymbol{n}|\boldsymbol{h}(\boldsymbol{s}))$ with $\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}] = \boldsymbol{h}(\boldsymbol{s})$. We consider a number of different noise models, including correlated and power-law noise, but begin with the simplest case of uncorrelated Gaussian noise with Poisson-like scaling of variance. Our framework is summarised in Fig. 2.

$$\begin{tikzcd}
\text{Stimulus } \boldsymbol{s}\sim P(\boldsymbol{s}) \arrow[d, swap, "\textit{Computation}"] & \text{Environment} \arrow[l] \\
\text{Representational curves } \Omega_i(\boldsymbol{s}) \arrow[r, "\times g_i"] & \text{Tuning curves } h_i(\boldsymbol{s}) = g_i\Omega_i(\boldsymbol{s}) \arrow[d, swap, "\textit{Noise}"] \\
 & \text{Neural responses } \boldsymbol{n}\sim P_{\text{noise}}(\boldsymbol{n}|\boldsymbol{h}(\boldsymbol{s}))
\end{tikzcd}$$

Figure 2: The basic framework for our analysis. The environment gives rise to a stimulus distribution P from which the stimulus $\boldsymbol{s}$ is drawn. Conceptually, the brain performs some computation on $\boldsymbol{s}$ yielding representational curves $\Omega_i(\boldsymbol{s})$. These are multiplied by adaptive gains, g_i, to yields the tuning curves $h_i(\boldsymbol{s})$. The actual, single-trial neural responses are noisy emissions based on $h_i(\boldsymbol{s})$.

We theorise that neurons adapt their gains to maximise an objective function, $\mathcal{L}^0(\boldsymbol{g})$, that trades off the metabolic cost of neural activity with the information conveyed by the responses (Levy and Baxter, 1996; Ganguli and Simoncelli, 2014):

$$\mathcal{L}^0(\boldsymbol{g}) := 2\mu I(\boldsymbol{n};\boldsymbol{s}) - \sum_{i=1}^{N} \mathbb{E}[n_i] = \text{Information} - \text{Metabolic cost}. \tag{1}$$

Here, $I(\boldsymbol{n};\boldsymbol{s})$ is the mutual information between the stimulus and response. The second term penalises the average population spike count, as a measure of metabolic energy cost, and $\mu > 0$ controls the information-energy trade-off. This trade-off is illustrated in Fig. 1. In the following sections we will optimise an approximation to $\mathcal{L}^0$ to make predictions about the behaviour of the optimal gains. Here we make a few remarks on our theoretical framework and its assumptions.

First, note that, because Ω_i depend deterministically on the stimulus $\boldsymbol{s}$, and in turn fully determine the statistics of n_i, we have the identity $I(\boldsymbol{n};\boldsymbol{s}) = I(\boldsymbol{n};\boldsymbol{\Omega})$. This means that the coding fidelity term in the objective function can equivalently be interpreted, without any reference to the low-level stimulus, as the mutual information between the population's noisy responses, $\boldsymbol{n}$, and its ideal, noise-free outputs (up to scale), $\boldsymbol{\Omega}$, as determined by the computational goals of the circuit.

Next, note that adjusting gains does not restrict the computations which can be performed by downstream populations, as adjustments in gains can be compensated for by reciprocal adjustments in readout weights. Consider a downstream neuron which receives synaptic input from the population with synaptic weights w_i. The response of this neuron is then a nonlinear function of its total input $\sum_j w_j h_j(\boldsymbol{s})$, which, given our decomposition, can be written as $\sum_j w_j g_j \Omega_j(\boldsymbol{s})$. We therefore see that any modification of the gains, g_i, can be compensated for by a modification of the synaptic weights, w_i, by keeping the product of the two the same. This demonstrates that fixing Ω_i and optimising $\mathcal{L}^0$ in the gains provides no additional constraints on what downstream neurons encode in their responses. Thus, we not only do not constrain the representations of the population in consideration, but also do not constrain the representations of downstream read-out populations.

Finally, note that reciprocal adjustment of synaptic weights with the gains (in such a way to keep $w_j g_j$ constant) will keep constant the synaptic currents and therefore also the metabolic cost of synaptic transmission. This justifies our choice not to explicitly include synaptic transmission costs in our objective function, and take the metabolic cost term to be linear in the firing rates of the population.

2.2 Poisson-like noise model and upper bound to mutual information

We start by considering the simplest case of an uncorrelated Gaussian noise model with Poisson-like scaling of variance (*i.e.*, with unit Fano factor). According to this noise models individual neuron responses, n_i, are conditionally independent with conditional distribution

$$n_i | \boldsymbol{s} \sim \mathcal{N}(h_i(\boldsymbol{s}), h_i(\boldsymbol{s})). \tag{2}$$

Later in Sec. 2.7, we show that, to a good approximation, the solutions for this model extend to the case of true Poisson noise.

We consider the problem of calculating the optimal gains $\{g_j\}_{j=1}^N$ by maximising our target functional $\mathcal{L}^0$ (Eq. (1)). Unfortunately, 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 therefore consider optimising a bound on $\mathcal{L}^0$. 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,

$$H[\boldsymbol{n}] \leq \frac{1}{2} \ln \det(2\pi e \operatorname{Cov}(\boldsymbol{n}))). \tag{3}$$

Making this replacement gives us a new objective function

$$\mathcal{L}(\boldsymbol{g}) = 2\mu \left(\frac{1}{2} \ln \det(2\pi e \operatorname{Cov}(\boldsymbol{n}))) - H[\boldsymbol{n}|\boldsymbol{s}] \right) - \sum_{j=1}^{N} \mathbb{E}[n_j] \geq \mathcal{L}^0(\boldsymbol{g}). \tag{4}$$

Note that $\mathcal{L}$ depends tacitly on $P(\boldsymbol{s})$ and the tuning curves $\Omega_i(\boldsymbol{s})$. However, one can show (see App. A.1) that $\mathcal{L}$ depends on P and Ω_i only through the following collection of parameters:

$$\omega_j := \mathbb{E}[\Omega_j(\boldsymbol{s})] = \int \Omega_j(\boldsymbol{s}) P(\boldsymbol{s}) d\boldsymbol{s}, \tag{5}$$

$$\mathrm{CV}_j := \frac{\sqrt{\operatorname{Var}(\Omega_j(\boldsymbol{s}))}}{\mathbb{E}[\Omega_j(\boldsymbol{s})]}, \tag{6}$$

$$\rho_{ij} := \frac{\operatorname{Cov}[\Omega_i(\boldsymbol{s}), \Omega_j(\boldsymbol{s})]}{\sqrt{\operatorname{Var}(\Omega_i(\boldsymbol{s})) \operatorname{Var}(\Omega_j(\boldsymbol{s}))}}. \tag{7}$$

These parameters characterise different aspects of the population response statistics: ω_j is the *pre-modulated* (*i.e.*, without multiplying by g_i) average spike count; CV_j is the coefficient of variation of $\Omega_j(\boldsymbol{s})$, or equivalently of $h_j(\boldsymbol{s})$; and ρ is the signal correlation matrix, *i.e.*, the matrix of Pearson correlation coefficients of the vector $\boldsymbol{\Omega}(\boldsymbol{s})$, or equivalently $\boldsymbol{h}(\boldsymbol{s})$ (note that CV_j and ρ_{ij} are independent of the gains, as the Ω_i's in their definition can be replaced with h_i's with no effect). As we show in App. A.1, in terms of these parameters, $\mathcal{L}(\boldsymbol{g})$ is given by:

$$\mathcal{L}(\boldsymbol{g}) = \mu \ln \det(I_N + P\, \hat{\boldsymbol{\omega}} \hat{\boldsymbol{g}}) - \sum_{j=1}^{N} g_j \omega_j + \text{const.}, \tag{8}$$

where

$$P := \hat{\mathbf{CV}}\, \rho\, \hat{\mathbf{CV}} \tag{9}$$

and I_N is the $N \times N$ identity matrix. We use a hat to denote a vector turned into a diagonal matrix (in other words, $\hat{\boldsymbol{v}}$ denotes the $N \times N$ diagonal matrix with the elements of the vector $\boldsymbol{v}$ on its diagonal). We will now consider the form of ρ corresponding to a population composed of clusters of similarly tuned neurons.

2.3 Clustering neurons based on similarity of response

In cortex, it is found that many neurons located spatially nearby to one another also have very similar response properties (Obermayer and Blasdel, 1993). We will call a collection of neurons which have very similar representational curves (but not necessarily similar gains) a *cluster*. To gain intuition and allow for analytic solutions, we will subsequently adopt a toy model in which the neurons in the population are sorted into K such clusters, each containing k neurons (thus $N = Kk$). In this toy model, neurons within a cluster have very similar stimulus tuning, but neurons in different clusters are tuned differently. This approximates a more realistic regime in which there is a gradual smooth transition from more similar to less similar tuning curves within the entire population.

More concretely, we treat the representational curves of the neurons in cluster a as small perturbation to a cluster-wide representational curve $\tilde{\Omega}_a(\boldsymbol{s})$. Signal correlations between neurons within a cluster are therefore close to 1. Similarly, signal correlations between neurons in distinct clusters a and b are approximately given by the correlation between $\tilde{\Omega}_a(\boldsymbol{s})$ and $\tilde{\Omega}_b(\boldsymbol{s})$, which we denote by $\tilde{\rho}_{ab}$ (which jointly form a $K \times K$ matrix, $\tilde{\rho}$). More precisely, we assume the correlation between two neurons i and j, belonging to clusters a and b, respectively, is given by

$$\rho_{ij} = \tilde{\rho}_{ab} - \epsilon\, T_{ij}, \tag{10}$$

where ϵ is a small parameter controlling the deviation from perfect clusters, and T is an arbitrary symmetric $N \times N$ matrix (subject only to the requirement that the full matrix, ρ, remains a valid correlation matrix). We will analogously denote the mean and coefficient of variation of $\tilde{\Omega}_a(\boldsymbol{s})$ by $\tilde{\omega}_a$ and $\tilde{\mathrm{CV}}_a$ (defined as in Equations (5)–(7), but with Ω's replaced with $\tilde{\Omega}$'s). We will assume that the coefficient of variation of all representational curves in cluster a is approximately $\tilde{\mathrm{CV}}_a$, and their pre-modulated firing rates are approximately $\tilde{\omega}_a$. Finally, let $c(a)$ denote the set of indices for neurons belonging to cluster a.

We show in App. A.2 that

$$\mathcal{L}(\boldsymbol{g}) = \tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}) + \epsilon \mathcal{L}_\epsilon(\boldsymbol{g}) + \mathcal{O}(\epsilon^2), \tag{11}$$

where $\tilde{g}_a \equiv \sum_{i \in c(a)} g_i$ is the sum of g_i over cluster a, and

$$\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}) = \mu \ln \det \left(I_K + \tilde{P} \hat{\tilde{\boldsymbol{\omega}}} \hat{\tilde{\boldsymbol{g}}} \right) - \sum_{a=1}^{K} \tilde{\omega}_a \tilde{g}_a, \tag{12}$$

where

$$\tilde{P} = \hat{\tilde{\mathbf{CV}}}\, \tilde{\rho}\, \hat{\tilde{\mathbf{CV}}}. \tag{13}$$

Note that $\tilde{\mathcal{L}}(\boldsymbol{g})$ is identical in form to the expression (8), with all quantities relating to individual neurons replaced by the corresponding quantities for a cluster. This is because for clusters of identically tuned neurons (corresponding to $\epsilon = 0$) neurons in a cluster can be considered as a single coherent unit whose firing rate is given by the sum of the firing rates of its neurons. Furthermore, in this approximation, the efficient coding objective function is indifferent to the precise distribution of firing rates among the neurons of a cluster, as long as the total rate of the cluster is held fixed.

We therefore approximate the maximisation of the total objective function $\mathcal{L}$ as follows. We first specify the cluster firing rates by maximising $\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}})$. We then specify the firing rates of individual neurons within a cluster by maximising $\mathcal{L}_\epsilon(\boldsymbol{g})$, subject to the constraint that the neuron firing rates in each cluster must sum to that cluster's previously optimised total firing rate. Although this is not the same as maximising the total objective function $\mathcal{L}$, this approximation is justified since terms of order ϵ and above have negligible effect on determining the cluster firing rates, but (as we will see in Sec. 2.5) are crucial for breaking the degeneracy within an individual cluster.

2.4 Optimal cluster gains display homeostasis across environments

To obtain an analytic approximation for the cluster gains, we apply an additional condition, discussed further in Sec. 2.6, that

$$\Delta = (\mu \hat{\tilde{\mathbf{CV}}}\, \tilde{\rho}\, \hat{\tilde{\mathbf{CV}}})^{-1} \ll 1. \tag{14}$$

The matrix Δ^{-1} characterises the structure of signal-to-noise ratio (SNR) in the space of population responses, and so our condition can be interpreted as the requirement that signal-to-noise ratio is strong. In App. A.3 we derive perturbative solutions for the cluster gains that maximise Equation (12). In the first approximation, *i.e.*, to zeroth order in Δ, we find

$$\tilde{g}_a \approx \frac{\mu}{\tilde{\omega}_a} := \tilde{g}_a^{(0)}. \tag{15}$$

To see the significance of this result, consider the stimulus-averaged spike count of cluster a. This is given by

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in c(a)} n_i\right] &= \sum_{i\in c(a)} \mathbb{E}[h_i(\boldsymbol{s})] = \sum_{i\in c(a)} g_i\omega_i \\ &\approx \tilde{g}_a\tilde{\omega}_a \approx \mu.\end{aligned}$$

Thus, in the leading approximation, the stimulus-average spike count of cluster a is given by the constant μ. Importantly, this is constant across *both* clusters and environments. In particular, the zeroth-order solution yields *homeostasis* of cluster firing rates. This is depicted in Fig. 3a.

We also note that, as long as the SNR structure, Δ, remains fixed between two environments, optimal gains imply exact homeostasis of each cluster's average firing rate between environments, irrespective of the size of Δ and how pre-adaptation mean rates, $\tilde{\omega}_a$, change between the two environments. However, unless Δ is small, firing rates are not equalised across clusters in the optimal solution, in general.

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

$$\tilde{g}_a \approx \frac{\mu}{\tilde{\omega}_a}\left(1 - \frac{[\tilde{\rho}^{-1}]_{aa}}{\mu \tilde{\mathrm{CV}}_a^2}\right) = \frac{\mu}{\tilde{\omega}_a}(1-\Delta_{aa}) := \tilde{g}_a^{(1)}. \tag{16}$$

which (assuming Δ_{aa} are small) yields approximate homeostasis, with small variations depending on the correlational structure between clusters and how that structure shifts between environments. Furthermore, inspired by the approximate homeostatic solution, we also numerically optimized Eq. (12), within a one-parameter family of putative solutions which by construction yield homeostasis. These solutions are of the form Eq. (15) with the constant μ replaced by an a priori unknown scalar variable, χ, which we numerically optimised; we denote the resulting solution by $\tilde{g}_a^{hom}$. In Sec. 2.8, we compare the performance of these solutions, the first and zeroth order approximations given above, as well as to numerically optimised gains.

2.5 Distribution of firing rates within neuronal clusters

Having derived an analytic approximation to the cluster firing rates $\tilde{\boldsymbol{g}}$, we next maximise the perturbation term $\mathcal{L}_\epsilon(\boldsymbol{g})$ in Eq. (11), subject to the constraint that the sum of gains within a cluster equals the obtained cluster solution, *i.e.*, $\sum_{i\in c(a)} g_i = \tilde{g}_a$. Since the term $\tilde{\mathcal{L}}$ depends only on the sum of the gains in a cluster, there is a great deal of redundancy in potential optimal solutions of single-neuron gains. The term $\mathcal{L}_\epsilon(\boldsymbol{g})$ acts to break symmetry within the cluster and yield a unique optimal solution. However, since $\mathcal{L}_\epsilon(\boldsymbol{g})$ is small and relatively flat, this symmetry breaking yields a solution with broadly heterogeneous mean firing rates across the neurons in a cluster, despite (approximately) equal total firing rate across clusters.

We show in App. A.4 that, to first order in ϵ and zeroth order in Δ, individual neural gains in the a-th cluster are given by the following constrained maximisation problem, which is a quadratic program in $\boldsymbol{g}$:

$$\min_{\boldsymbol{g}} \sum_{i,j\in c(a)} g_i\rho_{ij}g_j \tag{17}$$

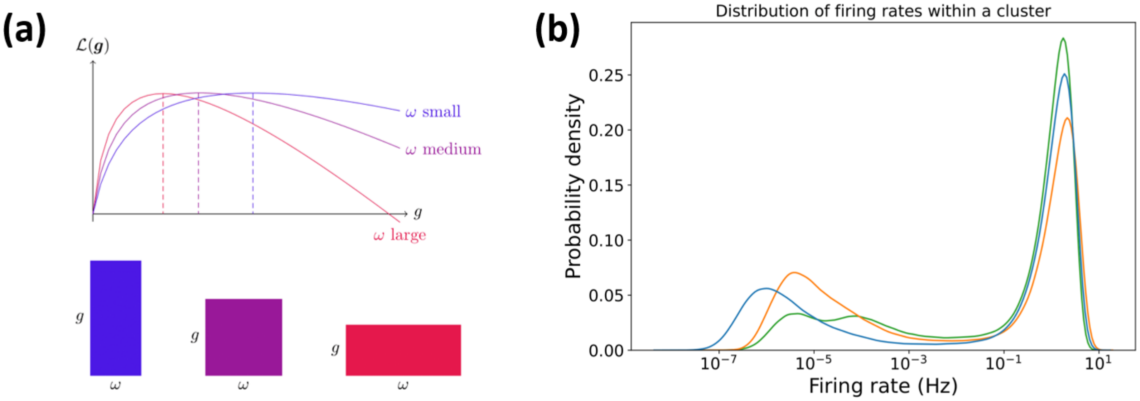

Figure 3: (a) Firing rate homeostasis at the level of clusters of similarly tuned neurons. We consider three environments, indexed by colour. Across these environments, the pre-modulated cluster rate $\tilde{\omega}$ is different, as is the functional $\tilde{\mathcal{L}}(\tilde{g})$. However, the argmax of the target function is so as to keep the product $\tilde{\omega}\tilde{g}$ approximately constant (here depicted as area of the corresponding rectangle). Since this product is the cluster firing rate, this strategy yields homeostasis. (b) The distribution of firing rates within a cluster of neurons. The densities obtained here were obtained by averaging over $10,000$ random draws of the within-cluster correlation matrix ρ using the method described in appendix 3.1. This method uses two parameters, (α, q). The values of (α, q) were as follows: blue $(10, 0.7)$; orange $(10, 0.9)$; green $(20, 0.9)$. We fix $\tilde{g} = 10$ and $k = 100$ throughout.

subject to the constraints

$$\sum_{i \in c(a)} g_i = \tilde{g}_a, \qquad g_i \geq 0. \tag{18}$$

Just as the correlation between clusters was the dominant force in determining the firing rates for each cluster, the dominant force in determining firing rates within a cluster is the correlational structure of the cluster itself, and not correlations between clusters. Note that we can interpret this quadratic program as aligning the rates within a cluster with the direction of minimal signal correlation, thereby maximising the information conveyed by the population.

We solved Eq. (17) numerically for correlation matrices ρ which were generated to have a high level of correlation within clusters; specifically, within each cluster ρ has a large eigenvalue with eigenvector closely aligned to the vector with all components equal to 1 (see Sec. 3.1). The firing rate distributions resulting from the numerically optimised gains are shown in Fig. 3b. The three lines in the plot correspond to different parameter values used to generate the correlation matrix ρ within the cluster. Note importantly that the firing rate is plotted on a logarithmic axis. Firstly, we see that maximising $\mathcal{L}_\epsilon$ leads to a diverse range of firing rates within each cluster. In particular, the distribution of rates spans multiple orders of magnitude, in agreement with cortical observations (*e.g.* (Slomowitz et al., 2015; Hengen et al., 2013; Maffei and Turrigiano, 2008)). Furthermore, the distributions we found are bimodal. The more prominent peak yields an approximately log-normal distribution consistent with empirical findings in cortex (Buzsáki and Mizuseki, 2014). Given that the smaller mode occurs between extremely small firing rates, 10^{-3} to 10^{-7} Hz, it is reasonable to expect neurons in this mode to be essentially silent during experimental observations and remain undetected. Our theory therefore predicts that a significant but variable fraction of cortical neurons are silent at any given sensory environment, and that which neurons are silent will shuffle following shifts in the environment (and subsequently changes in ρ).

2.6 Conditions for the validity of the homeostatic solution hold in cortex

In this section, we consider conditions under which the homeostatic solution, Eq. (15), provides a good approximation to the optimal gains. We obtained this solution, and the first order correction to it, Eq. (16), by expanding in Δ (Eq. (14)). The rigorous condition for the validity of this expansion requires that Δ is small in norm (*i.e.*, that its largest eigenvalue is small). Here we instead consider conditions under which the mean eigenvalue of Δ is small, *i.e.*, that $\mathrm{tr}(\Delta)/K$ is small. Note that $g_a^{(1)} = g_a^{(0)}(1 - \Delta_{aa})$. If the eigenbasis of Δ is unaligned with the standard basis, we expect the diagonal elements Δ_{aa} to be clustered around $\mathrm{tr}(\Delta)/K$. Thus, showing that $\mathrm{tr}(\Delta)/K$ will be small indicates that the first-order correction terms will be small.

By Eq. (14), we have that

$$\frac{1}{K}\mathrm{tr}(\Delta) = \frac{1}{K}\sum_{a=1}^{K} \frac{\tilde{\rho}^{-1}_{aa}}{\mu \tilde{\mathrm{CV}}_a^2}. \tag{19}$$

We will thus obtain estimates for $\mu, \tilde{\mathrm{CV}}$, and $\tilde{\rho}$. Qualitatively, requiring that (19) is small requires that firing rates of each cluster are sufficiently high, each cluster is selective in its response to stimuli, and neural responses correspond to a high dimensional geometry.

- We see from the zeroth order solution (Eq. (15)) that μ is approximately equal to the average spike count of the neuron cluster over the rate coding time-interval. Condition (14) requires that the stimulus-averaged firing rate of all clusters are sufficiently high. A wide range of mean firing rates have been reported in cortex. Here we focus on firing rates in rodent V1 during free behaviour. Reported values tend to range from 0.4 Hz (Greenberg et al., 2008) to 14 Hz (Parker et al., 2022) 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, a firing rate of 5 Hz is reasonable. Assuming a coding interval of 0.1 seconds, and a cluster size of $k = 20$, we obtain $\mu \approx 10$. If firing rates were significantly different, a similar value of μ could be achieved by scaling clusters up or down.

- The coefficient of variation $\tilde{\mathrm{CV}}$ can be seen as a measure of sparseness of responses. Our condition therefore requires that neurons are selective in their responses and only respond to a small fraction of stimuli. To see this, consider a toy-model in which the cluster responds at fixed level to a fraction p_a of stimuli and is silent otherwise. In this case, $\tilde{\mathrm{CV}}_a^2 = (1 - p_a)/p_a \approx 1/p_a$ for small p_a. Reference (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%. This yields an estimate of $\tilde{\mathrm{CV}}^2 \approx 20$. The Bernoulli distribution is particularly sparse, and so we take $\tilde{\mathrm{CV}}^2 \approx 10$ as a more conservative estimate.

- $\tilde{\rho}$ is the signal correlation matrix of neuron clusters. $[\tilde{\rho}^{-1}]_{aa}$ can be seen as a measure of the extent to which cluster a shares its representation with other clusters, *i.e.*, redundancy in representation; $[\tilde{\rho}^{-1}]_{aa} \geq 1$, with equality if and only if cluster a has zero signal correlation with every other cluster. Many traditional efficient coding accounts predict zero signal correlation between neurons (in the low noise limit) (Barlow, 2012; Nadal and Parga, 1994), providing an additional normative justification for low signal correlations. This is known as a factorial code (Nadal and Parga, 1999).

  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 high-dimensional representation of stimuli.

  Stringer *et al.* (Stringer et al., 2019) found that, for high-dimensional (ecological) stimuli, the signal correlation matrix possesses a $1/n$ spectrum. Again, assuming that the eigenbasis of $\tilde{\rho}$ does not align with the standard basis, we expect the values of $\tilde{\rho}^{-1}_{aa}$ to be clustered around $\mathrm{tr}(\tilde{\rho}^{-1})/K$. In this case, for K large, we obtain the estimate $\tilde{\rho}^{-1}_{aa} \approx \mathrm{tr}(\tilde{\rho}^{-1})/K \approx \ln(K)/2$. (see App. A.5)

Given the slow logarithmic increase with K, and the above estimated for μ and $\tilde{\mathrm{CV}}$, we find that our conditions holds even for very large neural populations. For example, suppose we take the entire human V1 as our neural population. This contains roughly 1.5×10^8 neurons (Wandell, 1995), leading to 7.5×10^6 clusters of 20 neurons, and therefore an average value of $[\tilde{\rho}^{-1}]_{aa}$ just under 8.

These estimates show that in V1 and possibly other cortical areas, we can expect the first order correction terms in (16), $\Delta_{aa} = [\tilde{\rho}^{-1}]_{aa}/(\mu \tilde{\mathrm{CV}}_a^2)$ to not exceed 0.1, and equivalently for the average eigenvalue of Δ to be bounded by 0.1. Furthermore, the above analysis makes it clear when we should expect homeostasis in general – when cluster firing rates are not too *low*, responses are highly *selective*, and signal correlation structure corresponds to a *high-dimensional geometry* (Stringer et al., 2019). On the other hand, when these conditions are violated, the optimal gain configuration can deviate strongly from the homeostatic solution.

2.7 Generalisations to alternative noise models

Here we move beyond the simple case of unit Fano factor Gaussian noise to other noise models. In particular, we consider first Poissonian noise, and then correlated and power-law noise.

In Sec. 2.4 we chose the cluster gains $\tilde{\boldsymbol{g}}$ to maximise $\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}})$. Therefore, to extend our analysis, the definition of the functional $\tilde{\mathcal{L}}$ will need to be generalised. To do so, we recall that in Sec. 2.3, we demonstrated that (for Gaussian noise) $\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}})$ arises from applying the Gaussian upper bound to the information-energy trade off applied to the cluster responses $\tilde{\boldsymbol{n}}$. Therefore, for other noise models, we take this as the definition of $\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}})$.

Poisson noise model The toy model of Gaussian noise is unrealistic for two main reasons: spike counts are discrete, while Gaussian variables are continuous; and spike counts are non-negative, while Gaussian variables can be negative. A more realistic noise model is given by Poisson spike counts,

$$n_j|\boldsymbol{s} \sim \mathrm{Poisson}\left(h_j(\boldsymbol{s})\right). \tag{20}$$

Since the sum of independent Poisson random variables is also Poisson, we arrive at the following distribution for cluster spike counts:

$$\tilde{n}_a|\boldsymbol{s} \sim \mathrm{Poisson}\left(\tilde{g}_a \tilde{\Omega}_a(\boldsymbol{s})\right). \tag{21}$$

To find $\tilde{\mathcal{L}}$, we consider the Gaussian upper bound to the information-energy trade-off applied to the cluster responses. A slight subtlety arises from the fact that the entropy for a Poisson random variable is discrete and the Gaussian upper bound is for a continuous random variable. However, we show in App. A.7 that we have the upper bound

$$H[\tilde{\boldsymbol{n}}] \leq \frac{1}{2}\ln\left((2\pi e)^K \det\left(\frac{1}{12}I_K + \mathrm{Cov}(\tilde{\boldsymbol{n}})\right)\right). \tag{22}$$

We will make an approximation by neglecting the $I_K/12$ term, which under the relevant conditions is negligible compared the covariance.

We make one additional assumption on the representational curves $\tilde{\Omega}_a$. Specifically, we assume that these representational curves have a baseline, and never fall down to zero. To be more precise, we assume that $\tilde{\Omega}_a \gg 5\tilde{\omega}_a/\mu$ everywhere. We show in App. A.7 that under this condition, the functional $\tilde{\mathcal{L}}(\tilde{g})$ becomes identical to that obtained in the Gaussian noise case (see Sec. 2.4, Eq. (8)). Therefore we have the same optimal gains and approximations to first and zeroth order in Δ.

This analysis reveals why the Gaussian noise model assumed in earlier sections is far less restrictive than it originally appears. Poisson random variables can be approximated as sums of independent

Bernoulli random variables. According to the Central Limit Theorem, for large enough rates, these sums behave like a normal random variable, giving rise to approximately Gaussian noise.

Likewise, the spike count of a cluster of neurons can be considered as the sum of a large collection of (mostly) independent random variables (*i.e.* the spike count of individual neurons). Therefore, whatever the distribution of individual neuronal spike counts, we expect the distribution of a cluster spike count to be approximately Gaussian, at least in the case of low enough noise correlations and large enough clusters. This demonstrates that the Gaussian noise case for clusters covers all well-behaved individual neuron distributions, at least approximately.

Power-law and correlated noise Here we extend our analysis to the case of correlated noise with a general, power-law scaling of response variance with trial-average response. We now adopt the following noise model for cluster responses

$$\tilde{\boldsymbol{n}}|\boldsymbol{s} \sim \mathcal{N}\left(\tilde{\boldsymbol{h}}(\boldsymbol{s}), \sigma^2 \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \, \Sigma(\boldsymbol{s}) \, \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha\right), \tag{23}$$

where $\Sigma(\boldsymbol{s})$ is the stimulus-dependent noise correlation matrix, and $0 < \alpha < 1$ is a scaling parameter. We define $\beta = 2(1-\alpha)$. Note that $\sigma^2 = 1$, $\alpha = \frac{1}{2}$ yields unit Fano factor noise scaling, and $\Sigma(\boldsymbol{s}) = I$ yields uncorrelated noise. In the case $\alpha = 1$, the noise scales at the same rate as the signal, and the information term becomes degenerate and independent of neural gains. The energetic term therefore collapses the firing rate of all neurons down to zero. We consider this case unrealistic, as empirical data is more consistent with sublinear scaling of noise strength with signal strength even in super-Poisson conditions (Goris et al., 2014).

We once again will expand to zeroth and first order in a parameter that scales with inverse signal-to-noise ratio, which we require is small. In this case, we have the condition

$$\Delta := \sigma^2 \left((\beta\mu)^\beta \, \hat{\tilde{\mathbf{CV}}}(\alpha) \, \tilde{\rho} \, \hat{\tilde{\mathbf{CV}}}(\alpha)\right)^{-1} W \ll 1, \tag{24}$$

where

$$W_{ab} = \frac{\mathbb{E}\left[\tilde{\Omega}_a(\boldsymbol{s})^\alpha \Sigma_{ab}(\boldsymbol{s}) \tilde{\Omega}_b(\boldsymbol{s})^\alpha\right]}{\sqrt{\mathbb{E}[\tilde{\Omega}_a(\boldsymbol{s})^{2\alpha}]\mathbb{E}[\tilde{\Omega}_b(\boldsymbol{s})^{2\alpha}]}}, \tag{25}$$

is an effective stimulus-average noise correlation matrix, and $\tilde{\mathrm{CV}}_a(\alpha)$ is the "α-coefficient of variation" defined by

$$\tilde{\mathrm{CV}}_a(\alpha) = \tilde{\mathrm{CV}}_a \times \frac{\tilde{\omega}_a^\alpha}{\sqrt{\mathbb{E}[\tilde{\Omega}_a(\boldsymbol{s})^{2\alpha}]}}. \tag{26}$$

Note that W is truly a correlation matrix, in the sense of being positive definite and having a unit diagonal. Note also, that $\tilde{\mathrm{CV}}(\alpha)$ and W are independent of neural gains, and can be defined in terms of full tuning curves, $h_i(\boldsymbol{s})$. Moreover, $\tilde{\mathrm{CV}}(\alpha) > \tilde{\mathrm{CV}}$ for $\alpha > 1/2$ and $\tilde{\mathrm{CV}}(\alpha) < \tilde{\mathrm{CV}}$ for $\alpha < 1/2$. Much as before, (24) is a requirement that we are in a high signal-to-noise condition. Note that we recover the previous expression for Δ when $\beta = 1$, $\sigma^2 = 1$ and $W = I_K$.

Expanding in Δ as before (see appendix A.8) we arrive at the following zeroth and first order expressions:

$$\tilde{g}_a^{(0)} = \frac{\beta\mu}{\tilde{\omega}_a}, \tag{27}$$

and

$$\tilde{g}_a^{(1)} = \frac{\beta\mu}{\tilde{\omega}_a}\left(1 - \frac{\sigma^2}{(\beta\mu)^\beta}\sum_{b=1}^{K}\frac{[\tilde{\rho}^{-1}]_{ab}W_{ba}}{\tilde{\mathrm{CV}}_a(\alpha)\tilde{\mathrm{CV}}_b(\alpha)}\right). \tag{28}$$

Note that, since $\tilde{g}_a^{(0)}\tilde{\omega}_a = \beta\mu$, we once again have homeostasis of firing rates at zeroth-order. This provides a considerable generalisation of our previous result. In particular, we have demonstrated that

that constant Fano factor scaling of noise and uncorrelated noise is not necessary to yield homeostasis as an optimal solution; correlated, power-law scaled noise also suffices, provided conditions similar to those discussed in Sec. 2.6 hold.

2.8 Gains derived by enforcing cluster level homeostasis

Examining Eq. (27) and Eq. (28), we note that $\tilde{g}_a^{(1)} < \tilde{g}_a^{(0)}$, since the first order expression is equal to the zeroth order expression multiplied by a *suppressive factor*, $\tilde{g}_a^{(1)} = \tilde{g}_a^{(0)} s_a$, where

$$s_a := \left(1 - \frac{\sigma^2}{(\beta\mu)^\beta} \sum_{b=1}^{K} \frac{[\tilde{\rho}^{-1}]_{ab} W_{ba}}{\tilde{\mathrm{CV}}_a(\alpha)\tilde{\mathrm{CV}}_b(\alpha)} \right) < 1. \tag{29}$$

There are two basic possibilities for the suppressive factor s_a. Either this factor displays significant and important variation between neurons, or it does not. In the latter case, the suppressive factor would be reasonably constant (at least when compared with the scale of variation in $\tilde{\omega}_a$). We therefore consider a set of gains with a *shared* suppressive factor s^{hom} which is fixed both between clusters and environments.

$$\tilde{g}_a^{hom} := \tilde{g}_a^{(0)} s^{hom}. \tag{30}$$

Note that these gains (like $\tilde{\boldsymbol{g}}^{(0)}$) lead to homeostasis both between clusters and environments. If $\tilde{\boldsymbol{g}}^{hom}$ performs comparably to $\tilde{\boldsymbol{g}}^{(1)}$ it would indicate that variation in the suppressive factor is unimportant, and an exactly homeostatic strategy can lead to near optimal performance.

We define $\chi := \beta\mu s^{hom}$, so that $\tilde{g}_a^{hom} = \chi/\tilde{\omega}_a$. χ can be interpreted as the (stimulus-averaged) spike count of each cluster during the coding interval under the gains $\tilde{\boldsymbol{g}}^{hom}$. We can choose χ by maximising the expected value of $\tilde{\boldsymbol{\mathcal{L}}}$ across environments. We show in App. A.9 that this is given by

$$\mathbb{E}[\tilde{\boldsymbol{\mathcal{L}}}] = \mu\mathbb{E}[\log\det(I_K + \chi^\beta Q)] - K\chi, \tag{31}$$

where

$$Q := \sigma^{-2} W^{-1} \hat{\tilde{\mathbf{CV}}}(\alpha)\, \tilde{\rho}\, \hat{\tilde{\mathbf{CV}}}(\alpha). \tag{32}$$

Note that $\tilde{\boldsymbol{\mathcal{L}}}$ depends only on the spectrum of Q, and not upon its eigenbasis. Thus, if Q has constant spectrum λ_n, then our optimised value χ must satisfy

$$\chi = \beta\mu s^{hom} = \beta\mu \left(1 - \frac{1}{K} \sum_n \frac{1}{1 + \chi^\beta \lambda_n} \right). \tag{33}$$

This expression shows a clear similarity to the suppression factors s_a (Eq. (29)). We can make the connection between the two clearer by noting that the average of the suppressive factors s_a across neurons is given by

$$\frac{1}{K} \sum_{a=1}^{K} s_a = 1 - \frac{1}{(\beta\mu)^\beta} \frac{\mathrm{tr}(Q^{-1})}{K}. \tag{34}$$

In the limit as $(\beta\mu)^\beta \lambda_n \gg 1$, (which corresponds to Eq. (24)), Eq. (33) yields

$$s^{hom} \approx \left(1 - \frac{1}{(\beta\mu)^\beta} \frac{\mathrm{tr}(Q^{-1})}{K} \right) = \frac{1}{K} \sum_{a=1}^{K} s_a. \tag{35}$$

Thus, in the high SNR limit, the shared suppressive factor in $\tilde{\boldsymbol{g}}^{hom}$ is equal to the average suppressive factor in $\tilde{\boldsymbol{g}}^{(1)}$

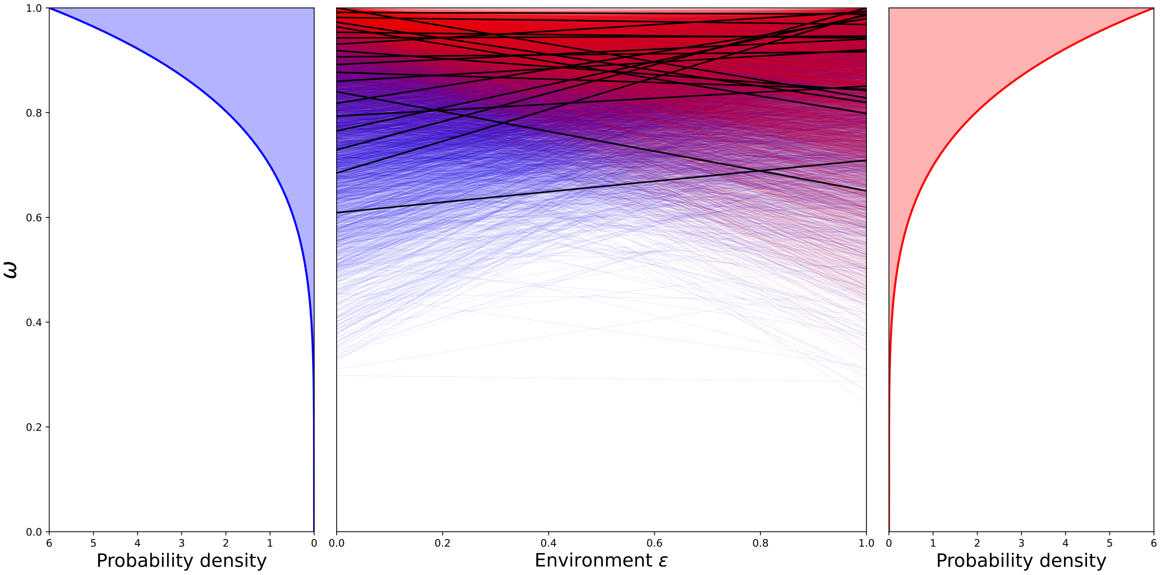

Figure 4: Diagram showing how $\tilde{\omega}_a(\epsilon)$ was generated. The end points $\tilde{\omega}_a(0)$ and $\tilde{\omega}_a(1)$ for each cluster index a is sampled independently from a $\mathrm{Beta}(6, 1)$ distribution whose densities are shown on the left and right. These values are then linearly interpolated (middle plot) to obtain $\tilde{\omega}_a(\epsilon)$. The lines in the middle plot are colour coded from blue to red in terms of the rank order of $\tilde{\omega}_a(\epsilon)$. Every 500th line is coloured black.

2.9 Numerical comparison of approximate gains to optimal gains

Here, we test the performance of $\tilde{\boldsymbol{g}}^{(0)}$, $\tilde{\boldsymbol{g}}^{(1)}$ and $\tilde{\boldsymbol{g}}^{hom}$ using numerical simulations. An environment is specified by $\tilde{\rho}$, $\tilde{\omega}_a$, and $\tilde{\mathrm{CV}}_a(\alpha)$ (via the stimulus density $P(\boldsymbol{s})$). To capture the notion of adaptation under environmental shift, we consider a sequence of environments parameterised by $\epsilon \in [0, 1]$. We will take the coefficient of variation to be constant between both clusters and environments, *i.e.* fix $\tilde{\mathrm{CV}}_a^2(\alpha) = 10$. We also fix $\mu = 10$ and $\sigma^2 = 1$. Thus the only randomness occurs in our choice of $\tilde{\rho}$, W and $\tilde{\boldsymbol{\omega}}$.

We obtain $\tilde{\omega}_a$ and $\tilde{\rho}$ by interpolating between the endpoints at $\epsilon = 0$ and $\epsilon = 1$. For each $a = 1, \ldots, K$, $\tilde{\omega}_a(\epsilon = 0)$ and $\tilde{\omega}_a(\epsilon = 1)$ are drawn independently from a $\mathrm{Beta}(6, 1)$ distribution. The value of $\tilde{\omega}_a(\epsilon)$ was then obtained by linear interpolation. This method for sampling environments is illustrated in Fig. 4. In line with the results of (Stringer et al., 2019), we obtain $\tilde{\rho}(\epsilon)$ by normalising a positive-definite covariance matrix $\Sigma(\epsilon)$ which has a $1/n$ eigen-spectrum (see Sec. 3.2 for further details)

We numerically compare the performance of the zeroth-order homeostatic code $\tilde{g}_a^{(0)} = (\beta\mu)/\tilde{\omega}_a$ (Eq. (27)), the first-order correction $\tilde{g}_a^{(1)}$ (Eq. (28)), and the homeostatic gains $\tilde{g}_a^{hom} = \chi/\tilde{\omega}_a$ (where χ maximises Eq. (31)) against gains $\tilde{g}_a^{\mathrm{opt}}$ which have been numerically optimised by performing gradient ascent on the objective $\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}})$. To compare the performance of each approximate solution, $\tilde{\boldsymbol{g}}^{\mathrm{app}}(\epsilon)$, we use the *relative improvement* measure

$$C^{\mathrm{app}}(\epsilon) = \frac{\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}^{\mathrm{app}}(\epsilon); \epsilon) - \tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}^{\mathrm{app}}(0); \epsilon)}{\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}^{\mathrm{opt}}(\epsilon); \epsilon) - \tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}^{\mathrm{app}}(0); \epsilon)}, \tag{36}$$

which can be interpreted as the improvement in $\tilde{\mathcal{L}}(\cdot; \epsilon)$ (the objective in environment ϵ, and according to its statistics) achieved by the adaptive gains $\boldsymbol{g}^{\mathrm{app}}(\epsilon)$ over the unadapted gains from the original environment $\boldsymbol{g}^{\mathrm{app}}(0)$, relative to the improvement obtained by the optimally adaptive gains. We took

$K = 10,000$ clusters. We additionally consider the mean relative errors

$$\frac{1}{K}\sum_{a=1}^{K}\frac{|g_a^{\text{opt}}(\epsilon) - g_a^{\text{app}}(\epsilon)|}{g_a^{\text{opt}}(\epsilon)}, \tag{37}$$

as another quantification of our approximations to the optimal gains. To give a sense of the scale of variation of the optimal gains, we additionally plot the relative error of the optimal gains between the end points,

$$\frac{1}{K}\sum_{a=1}^{K}\frac{|g_a^{\text{opt}}(0) - g_a^{\text{opt}}(1)|}{g_a^{\text{opt}}(0)}, \tag{38}$$

Uncorrelated power-law noise Consider first the case that the noise correlation matrix W is equal to the identity. We examine this scenario for a range of values for the parameter α (or equivalently $\beta = 2(1-\alpha)$) which controls the scaling of the noise with the signal. If we approximate the spectrum of $\tilde{\rho}(\epsilon)$ as fixed at $1/n$, then in App. A.9 we show that this implies that χ (approximately) satisfies the equation

$$\frac{\chi}{\beta\mu}\left(\frac{\sigma^2\ln(K)}{\chi^\beta \tilde{\text{CV}}(\alpha)^2}\right) = \ln\left(1 + \frac{\sigma^2\ln(K)}{\chi^\beta \tilde{\text{CV}}(\alpha)^2}\right). \tag{39}$$

We can solve this equation numerically to find χ. Note that in our numerical simulations, $\tilde{\rho}(\epsilon)$ does not actually possess an exact $1/n$ spectrum. Therefore, the value found by solving Eq. (39) is sub-optimal for our simulated correlations. However, this implies that the performance of $\tilde{\boldsymbol{g}}^{hom}$ in Fig. 5 represents a lower-bound on performance.

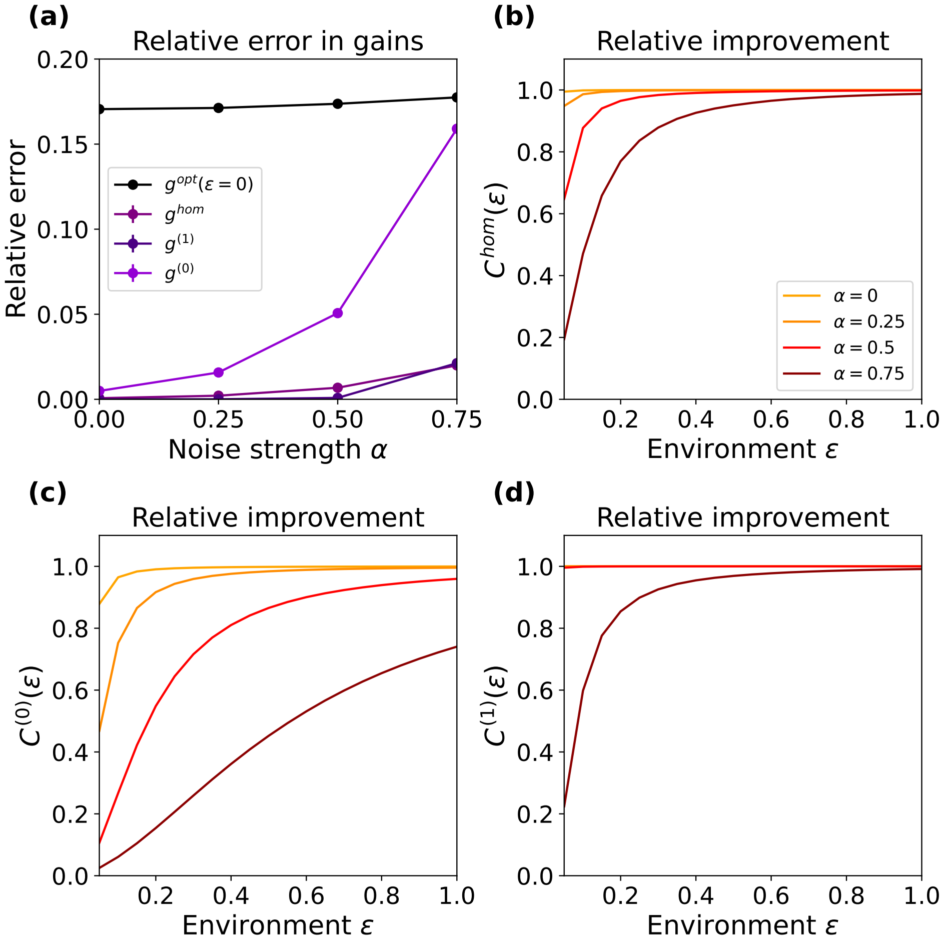

Figure 5: (a) Relative errors of gains. The mean (across ϵ) relative error (see Eq. (37)) of different approximate gains to the optimal gains, with standard deviation error bars, are plotted in purple, as a function of the noise strength parameter α. In black, we plot the deviation of the optimal gains between the end points (see Eq. (38)). (b) Relative improvement of $\tilde{g}^{hom}$. We used the measure given by Eq. (36) to quantify performance of the gains $\tilde{g}^{hom}$. Different colours give the curve of C^{hom} for different noise strengths, with darker colours indicating stronger noise. (c-d) Same as (b), but for $\tilde{g}^{(0)}$ and $\tilde{g}^{(1)}$ respectively.

In Fig. 5a, we plot the average relative errors of the approximate gains as a function of the noise strength (see Eq. (37)). We see that $\tilde{g}^{(1)}$ and $\tilde{g}^{hom}$ have low ($< 3\%$) relative errors for all noise strengths, while the relative error in $\tilde{g}^{(0)}$ becomes large for significantly supralinear noise growth (*i.e.*, $\alpha = 0.75$). This is predicted by our framework, as when α becomes larger the suppressive factors s_a (see Eq. (29)) become smaller, and therefore play a more significant role. This further explains why the relative improvement $C^{(0)}$ (shown in Fig. 5c) is significantly worse for $\alpha = 0.75$. The fact that the relative errors in $\tilde{g}^{(1)}$ and $\tilde{g}^{hom}$ are comparable, and C^{hom} displays similar performance to $C^{(1)}$ (although slightly worse) indicates that a shared suppressive factor performs almost as well as local suppressive factors, even for supralinear noise. Note that for all approximate gains, the performance of the gains (as quantified by C) decreases with increasing noise strength. This is to be expected, since increasing noise increases Δ and hence the influence of second order terms and higher.

Fig. 6 shows the rates of neuron cluster pre- and post-adaptation for the case of $\alpha = 1/2$ (Poisson noise). We use the optimised gains $\tilde{g}^{opt}$. We also show the density of pre-adaptation firing rates. Note that if we used $\tilde{g}^{(0)}$ instead of $\tilde{g}^{\text{opt}}$, then the pre-adaptation firing rates are $\tilde{g}^{(0)}(0)\tilde{\omega}(1) = \mu\frac{\tilde{\omega}(1)}{\tilde{\omega}(0)}$ which is the ratio of two independent beta random variables. Thus the density on the left will approximately be that of a (scaled) ratio of independent betas. Fig. 6 also gives us a sense of the scale of variation between the simulated environments $\epsilon = 0$ and $\epsilon = 1$.

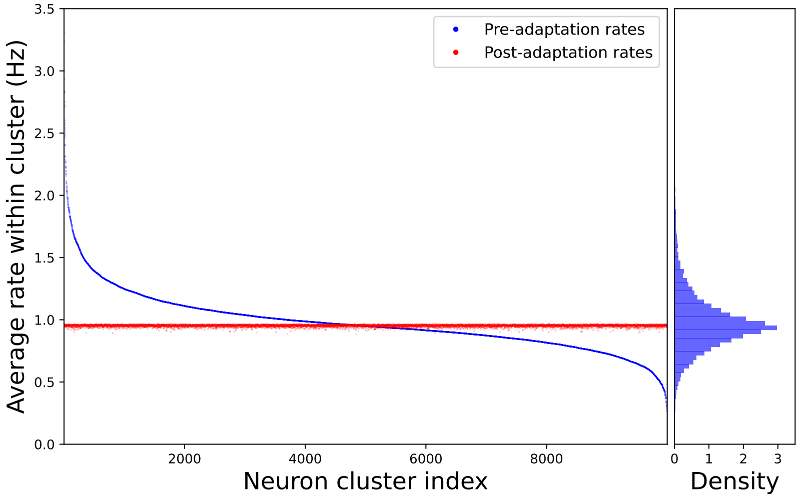

Figure 6: Cluster average firing rates pre- and post-adaptation. In red we plot the pre-adaptation firing rate $\propto \tilde{g}_a^{\text{opt}}(1)\tilde{\omega}_a(1)$ and in blue we plot the post-adaptation rates, $\propto \tilde{g}_a^{\text{opt}}(0)\tilde{\omega}_a(1)$. Clusters are ordered by their pre-adaptation firing rate. On the left we show the histogram of the pre-adaptation firing rates.

Aligned noise We now fix the noise to have Poisson-like scaling (*i.e.*, $\alpha = 1/2$), and take the noise correlation matrix W to have approximately the same eigenbasis as $\tilde{\rho}$, but a different spectrum. In particular, $\tilde{\rho}(\epsilon)$ is obtained by normalising a covariance with $1/n$ spectrum, and $W(\epsilon)$ is obtained by normalising a covariance with the same eigenbasis but a $1/n^{\gamma}$ spectrum. See Sec. 3.2 for further details. If we approximate $\tilde{\rho}$ and W as having exactly aligned bases with exact power-law eigenspectrum decay, we show in App. A.9 that this implies that χ satisfies the equation:

$$\chi = \mu\left(1 - \frac{1}{K}\sum_n \frac{1}{1 + \chi b n^{\gamma-1}}\right), \tag{40}$$

where b is a constant. Once again, since the ideal case leading to Eq. (40) differs from our numerical simulations, the value of χ found using this equation gives a lower bound on the performance of $\tilde{\boldsymbol{g}}^{hom}$.

Discuss Figure 7, and refer to $\gamma = 1$ case being so good, and that this corresponds to the case of Information Limiting Noise, for which there are several lines of evidence in V1. Yashar can add refs and add to thte discussion.

The case $\gamma = 1$ corresponds to perfect alignment between signal and noise correlations, $\tilde{\rho}^{-1}W = I$. The suppressive factors s_a become uniform in this case (see Eq. (29)), and $g^{(1)}$ also displays homeostasis between clusters and environments. Furthermore, the optimal gains $\tilde{g}^{\text{opt}}$ likewise display homeostasis both between clusters and environments. In fact, $\tilde{g}^{\text{opt}} = \tilde{g}^{hom}$, leading to zero relative error (Fig. 7a) and perfect relative improvement C^{hom} (Fig. 7b). This case is that of Information Limiting Noise. Our simulations suggest that a homeostatic strategy is particularly appropriate in this case. For other cases, we can see for higher rates of noise fall off (*i.e.*, γ increasing) the homeostatic gains $\tilde{g}^{(0)}$ and $\tilde{g}^{hom}$ tend to perform better, while $\tilde{g}^{(1)}$ tends to perform worse.

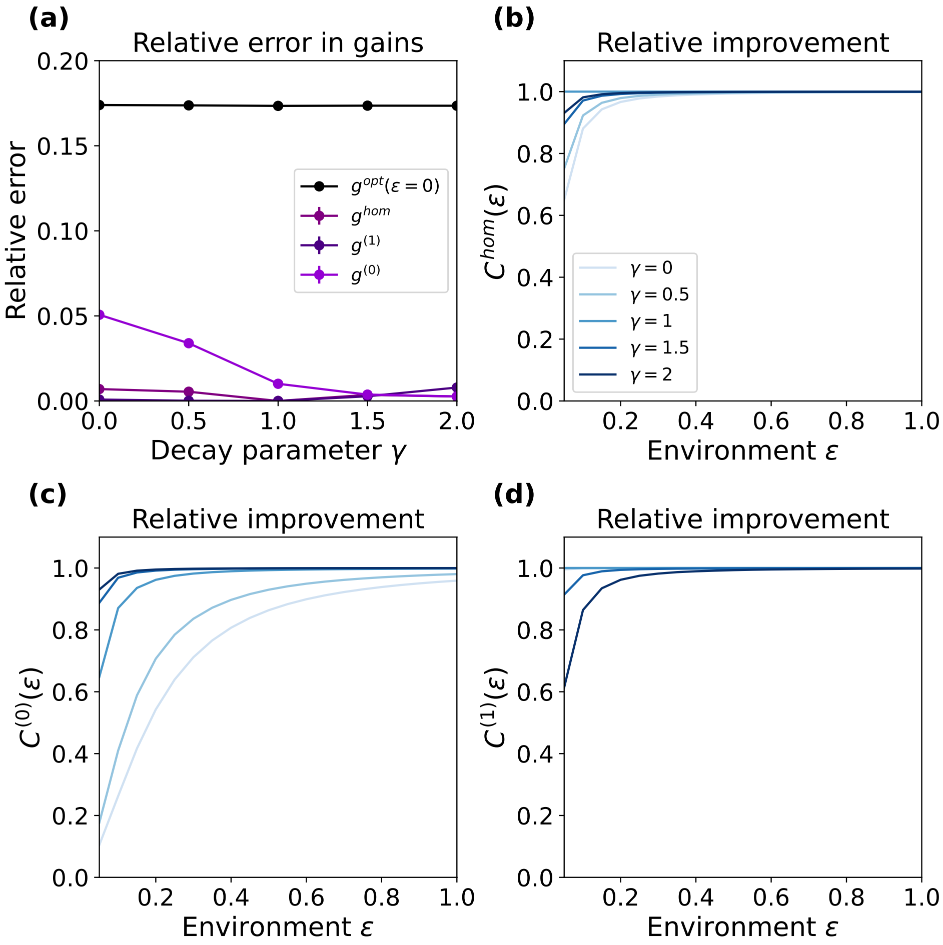

Figure 7: (a) Relative errors of gains. The mean (across ϵ) relative error (see Eq. (37)) of different approximate gains to the optimal gains, with standard deviation error bars, are plotted in purple, as a function of the decay parameter γ. In black, we plot the deviation of the optimal gains between the end points (see Eq. (38)). (b) Relative improvement of $\tilde{g}^{hom}$. We used the measure given by Eq. (36) to quantify performance of the gains $\tilde{g}^{hom}$. Different colours give the curve of C^{hom} for different noise correlation decay, with darker colours indicating a faster eigenvalue decay for noise correlations. (c-d) Same as (b), but for $\tilde{g}^{(0)}$ and $\tilde{g}^{(1)}$ respectively.

Constant correlation noise The last setting we consider is that of constant correlation noise. We once again fix $\alpha = 1/2$. We now take $W = (1-p)I_K + p\mathbf{1}\mathbf{1}^T$, where p is constant, *i.e.*, the noise correlation between any two distinct clusters is p. In this case, $W^{-1} = \frac{1}{1-p}\left(I_K - \frac{p}{1+p(K-1)}\mathbf{1}\mathbf{1}^T\right)$, which we approximate as $I_K/(1-p)$. Under this approximation, adding constant (positive) correlations across neurons has the same effect as scaling down the base level of noise $\sigma^2 \mapsto (1-p)\sigma^2$. We can obtain the following analytic approximation for χ (see App. A.9):

$$\chi = \mu \frac{\ln(K^q)}{K^q - 1}, \; q = \frac{\sigma^2(1-p)}{\mu \mathrm{CV}^2}. \tag{41}$$

Note that in the limit as $q \to 0$, this reduces to $\chi = \mu$. Because of our approximations, the value of χ found this way once again gives only a lower bound on the performance of $\tilde{\boldsymbol{g}}^{hom}$.

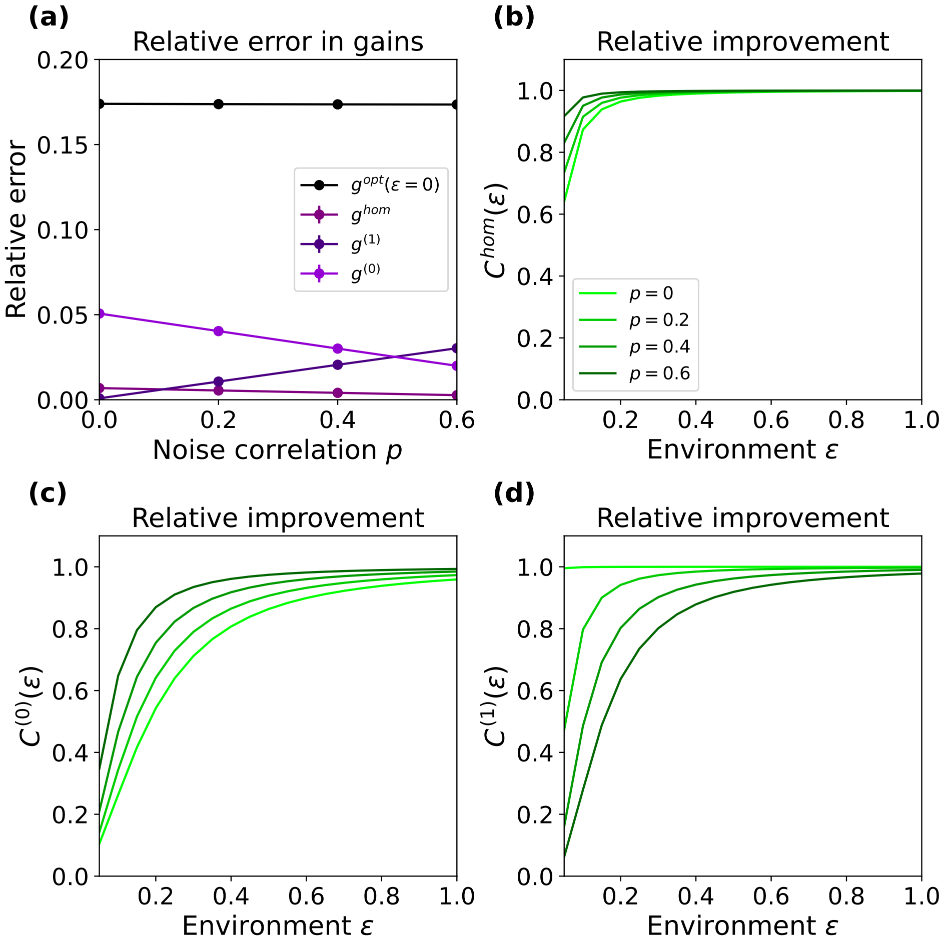

Figure 8: (a) Relative errors of gains. The mean (across ϵ) relative error (see Eq. (37)) of different approximate gains to the optimal gains, with standard deviation error bars, are plotted in purple, as a function of the strength of noise correlation p. In black, we plot the deviation of the optimal gains between the end points (see Eq. (38)). (b) Relative improvement of $\tilde{g}^{hom}$. We used the measure given by Eq. (36) to quantify performance of the gains $\tilde{g}^{hom}$. Different colours give the curve of C^{hom} for different noise correlations p, with darker colours indicating higher levels of noise correlation. (c-d) Same as (b), but for $\tilde{g}^{(0)}$ and $\tilde{g}^{(1)}$ respectively.

In Fig. 8a we examine the relative error in the approximate gains as a function of the noise correlation p. Note that our results display a linear relationship between the relative error and noise strength, with

$\tilde{g}^{(0)}$ and $\tilde{g}^{hom}$ decreasing and $\tilde{g}^{(1)}$ increasing. In a similar fashion, we see (Fig. 8b-d) that $C^{(0)}$ and C^{hom} increase as a function of p, while $C^{(1)}$ decreases. At a noise correlation of $p \approx 0.5$, the relative errors display an intersection between $\tilde{g}^{(0)}$ and $\tilde{g}^{(1)}$, suggesting that for noise correlations stronger than 0.5 the zeroth order approximation outperforms the first order approximation. This is reinforced by the fact that, for $p = 0.6$, $C^{(0)}$ lies above $C^{(1)}$. Furthermore, for all noise correlations $p \neq 0$, C^{hom} lies above $C^{(1)}$. Taken together, this indicates that constant positive noise correlations between neurons strongly favours a homeostatic strategy.

2.10 Synaptic Normalisation allows for propagation of homeostasis between populations

In this section, we drop corrections to the zeroth order and work only with the homeostatic coding regime for individual neurons, that is $g_i = \chi/\omega_i$. Given that we have demonstrated the approximate optimality of homeostatic coding, we now consider how a homeostatic code can be propagated between populations.

Consider a downstream population with tuning curves $H_m(\boldsymbol{s})$ for $m = 1, \ldots, M$. Let W_{mi} be the synaptic weight[1] from neuron i to neuron m in the downstream population. Working in a linear rate model, this gives us that the tuning curve of neuron m is

$$H_m(\boldsymbol{s}) = \sum_{i=1}^{N} W_{mi} h_i(\boldsymbol{s}). \tag{42}$$

Suppose that the downstream population has representational curves $\Gamma_m(\boldsymbol{s})$ for $m = 1, \ldots, M$. We can write each representational curve as a linear combination of the upstream cluster representational curves,

$$\Gamma_m(\boldsymbol{s}) = \sum_{j=1}^{N} w_{mj} \Omega_j(\boldsymbol{s}). \tag{43}$$

We now address the question of the relationship between the representational coefficients w_{mj} and the synaptic weights W_{mj}. Assuming the downstream population is also implementing homestatic coding, we obtain (see App. A.6) that

$$W_{mj} = \frac{w_{mj}\omega_j}{\sum_{i=1}^{N} w_{mi}\omega_i}. \tag{44}$$

Note that this scheme normalises the total synaptic input mass onto neuron m to be 1. Different normalisation factors can be achieved by different values of χ for different populations (arising from *e.g.*, different noise correlation statistics), or different rate coding intervals. Thus, we have demonstrated that homeostatic coding, applied sequentially to multiple populations, provides an additional normative interpretation to the learning rule of synaptic normalisation, in which synapses onto a neuron are jointly scaled to keep total input mass constant. This may be the computational reason for the "synaptic scaling" observed in certain studies of homeostasis (Turrigiano et al., 1998; Turrigiano, 2008).

2.11 Homeostatic 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. (14). We now apply our framework to a specific theory of neural representation, namely the distributive distributional code (DDC) (Vertes and Sahani, 2018). A DDC is based around a Bayesian encoding model, in which the stimulus $\boldsymbol{s}$ is drawn from a conditional distribution $f(\boldsymbol{s}|\boldsymbol{z})$. $\boldsymbol{z}$ is called a latent variable, and has prior distribution $\pi(\boldsymbol{z})$. The task of the brain

[1] Synaptic weights here are to be understood as the effect of a pre-synaptic neuron on post-synaptic firing rate, and not as *e.g.* the magnitude of post-synaptic mEPSP. Thus, changes in the intrinsic excitability of a post-synaptic neuron are incorporated into changes in the synaptic weights.

is to invert this generative model, and calculate the posterior distribution over latent variables given a stimulus $\boldsymbol{s}$. This is shown in Fig. 9.

$$\begin{tikzcd}
\text{Latent variables } \boldsymbol{z} \sim \pi \arrow[rr, bend left, "\text{Likelihood } f(\boldsymbol{s}|\boldsymbol{z})"] & & \text{Stimulus } \boldsymbol{s} \sim P \arrow[ll, bend left, "\text{Posterior } \Pi(\boldsymbol{z}|\boldsymbol{s})"]
\end{tikzcd}$$

Figure 9: In this section, the stimulus $\boldsymbol{s}$ is drawn from a distribution conditioned on the value of a latent variable $\boldsymbol{z}$ drawn from a prior π. The task of the brain is to compute the posterior Π over latent variables given the observed stimulus.

In a DDC, each neuron is equipped with a kernel function of the latent variables $\phi_i = \phi_i(\boldsymbol{z})$. The representational curves are then given by the posterior expectation of ϕ_i given the stimulus $\boldsymbol{s}$. For homeostatic gains $g_i = \chi/\omega_i$, this gives tuning curves

$$h_i(\boldsymbol{s}) = \chi \frac{\mathbb{E}[\phi_i(\boldsymbol{z})|\boldsymbol{s}]}{\mathbb{E}[\phi_i(\boldsymbol{z})]}. \tag{45}$$

where we have used the tower property to obtain $\mathbb{E}[\mathbb{E}[\phi_i(\boldsymbol{z})|\boldsymbol{s}]] = \mathbb{E}[\phi_i(\boldsymbol{z})]$. Thus our tuning curves are the ratio of a posterior to prior expectation. Here we have assumed that the internal prior possessed by the brain agrees with the external prior according to which $\boldsymbol{z}$ is drawn. In Sec. 2.13 we will relax this assumption.

2.12 Bayes-ratio coding via divisive normalisation

A special case of the homeostatic DDC arises in the case where the kernel functions are delta functions, $\phi_i(\boldsymbol{z}) = \delta(\boldsymbol{z} - \boldsymbol{z}_i)$ where $\boldsymbol{z}_i$ is a point in latent variable space corresponding to that neuron. In this case, the homeostatic DDC becomes

$$h_i(\boldsymbol{s}) = \chi \frac{\Pi(\boldsymbol{z}_i|\boldsymbol{s})}{\pi(\boldsymbol{z}_i)} = \chi \frac{f(\boldsymbol{s}|\boldsymbol{z}_i)}{P(\boldsymbol{s})}. \tag{46}$$

We call this coding scheme *Bayes-ratio coding.*

Bayes-ratio coding can be achieved via divisive normalisation with adaptive weights, a canonical cortical operation (Carandini and Heeger, 2012; Westrick et al., 2016). Given a collection of feed-forward inputs, $F_i(\boldsymbol{s})$, weighted divisive normalisation computes the response (and thus the tuning curve) of neuron i as

$$h_i(\boldsymbol{s}) = \chi \frac{F_i(\boldsymbol{s})^n}{\sigma^n + \sum_j w_j F_j(\boldsymbol{s})^n}, \tag{47}$$

where w_j are a collection of *normalisation weights*, and $\mu, \sigma \geq 0$ and $n \geq 1$ are constants. Bayes-ratio coding can be achieved naturally by a divisive normalisation model in which $n = 1$ and the feed-forward inputs are given by the generative model's likelihood function $f(\boldsymbol{s}|\boldsymbol{z}^i)$. We then choose the adaptive normalisation weights w_i to encode prior probabilities, $w_i = \pi(\boldsymbol{z}^i)\delta\boldsymbol{z}^i$, where the volume element $\delta\boldsymbol{z}^i$ is chosen such that the latent variable space is the disjoint union of volumes of size $\delta\boldsymbol{z}^i$ each containing their corresponding sample point $\boldsymbol{z}^i$. Note that these normalisation weights are adaptive, and vary between environments as π varies. Then we obtain:

$$\begin{aligned}
\sum_j w_j F_j(\boldsymbol{s}) &\approx \int f(\boldsymbol{s}|\boldsymbol{z})\pi(\boldsymbol{z})d\boldsymbol{z} \\
&= P(\boldsymbol{s}),
\end{aligned}$$

and hence

$$h_i(\boldsymbol{s}) \approx \chi \frac{f(\boldsymbol{s}|\boldsymbol{z}^i)}{\sigma + P(\boldsymbol{s})}. \tag{48}$$

Taking the limit as $\sigma \to 0$, we obtain Bayes-ratio coding Eq. (46). Therefore, provided σ is small compared to $P(\boldsymbol{s})$, divisive normalisation can be used to approximate Bayes-ratio coding. Not only does this show that implementing Bayes-ratio coding is biologically plausible, this framework gives a normative interpretation to both the feedforward inputs (as the generative model likelihoods) and the normalisation weights (as the prior probabilities).

Additionally, consider a hierarchical generative model $\boldsymbol{z}^{(2)} \to \boldsymbol{z}^{(1)} \to \boldsymbol{s}$. Suppose a downstream population represents the posterior probabilities of $\boldsymbol{z}^{(2)}$ conditional on $\boldsymbol{s}$ by means of Bayes-ratio coding. Working in the notation and framework of Sec. 2.10, this means taking downstream representational curves

$$\Gamma_m(\boldsymbol{s}) = \Pi\left(\boldsymbol{z}^{(2)} = \boldsymbol{z}_m^{(2)}|\boldsymbol{s}\right). \tag{49}$$

In App. A.10, we show that the corresponding synaptic weights are

$$W_{mj} = g\left(\boldsymbol{z}_j^{(1)}|\boldsymbol{z}_m^{(2)}\right)\delta\boldsymbol{z}_j^{(1)}, \tag{50}$$

where $\delta\boldsymbol{z}_j^{(1)}$ is the size of the volume element to which $\boldsymbol{z}_j^{(1)}$ belongs, and g is the conditional distribution of $\boldsymbol{z}^{(1)}$ given $\boldsymbol{z}^{(2)}$.

This result is significant for two reasons. Firstly, the synaptic weights make no reference to the prior distribution over $\boldsymbol{z}^{(2)}$. There is therefore no need to adapt the weights when the environment changes. Secondly, the downstream population represents a posterior distribution $\Pi\left(\boldsymbol{z}^{(2)} = \boldsymbol{z}_m^{(2)}|\boldsymbol{s}\right)$, and therefore acts as part of a *recognition* model. However, the synaptic weights are proportional to $g\left(\boldsymbol{z}_j^{(1)}|\boldsymbol{z}_m^{(2)}\right)$, and therefore require only knowledge of the *generative* probabilities. Inductively, we can see that this scheme can be propagated backwards through many layers, each representing posterior probabilities of latent variables further up a generative hierarchy.

2.13 Homeostatic DDCs account for stimulus specific adaptation in V1

Homeostatic DDCs can be used to explain certain stimulus-specific adaptation effects. It is typical of stimulus specific adaptation in V1 (Benucci et al., 2013) that orientation tuning curves display a repulsion and suppression around the over-represented orientation (adaptor stimulus).

The experiments performed by Benucci et al. 2013 (Benucci et al., 2013) examined the effects of adaptation on orientation tuned neurons in primary visual cortex of anaesthetised cats. Orientation tuned cells in V1 have tuning curves which are decreasing functions of the difference between the orientation of a grating placed in their visual field and a preferred orientation. Anaesthetised cats were shown a distribution of such gratings 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 2 seconds (or approximately 50 stimulus presentations), the tuning curves of the cats had adapted. In particular, both suppressive and repulsive effects were seen.

To model the findings of (Benucci et al., 2013) with homeostatic DDCs, we took stimulus and latent variable spaces to be the orientation space $[-90, 90)$, and a translation invariant likelihood $f(s|z) = f(s-z)$ proportional to the normal density with standard deviation σ_f, normalised over the circle S^1. Likewise, we take the kernel functions $\phi_i(z) = \phi(z - z^{(i)})$ proportional to the normal density with standard deviation σ_ϕ.

The distribution over stimuli orientations used in the experiment is highly unusual. The internal prior density possessed by the brain is more likely to be smooth, reflecting an inductive bias. To model this, we constrained the internal prior to a uniform-Gaussian mixture, with the Gaussian component having standard deviation σ_π. The mixing proportion was chosen to enforce that the density of the

Gaussian component, plus the density from the uniform component over the adaptor stimulus region, is equal to the adaptor probability. Our model therefore has only three free parameters, σ_f, σ_ϕ and σ_π.

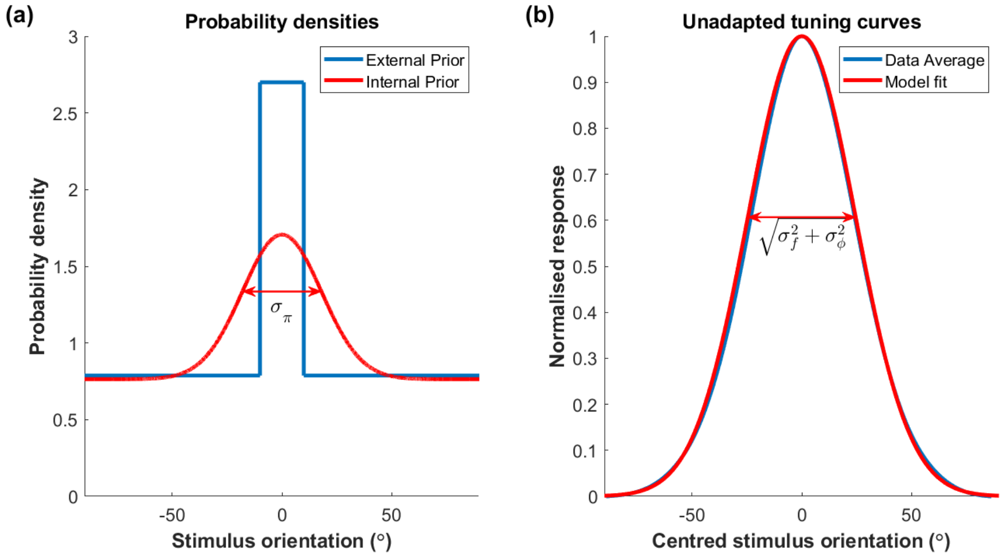

Figure 10: (a) The density of orientations used in the experiments (Benucci et al., 2013), for an adaptor probability of 30% and 8 stimulus orientations is shown in red. In blue is the Gaussian-uniform mixture distribution used in our model. (b) Unadapted tuning curves, normalised to have response of 1 at the preferred orientation and 0 at orthogonal orientations. The blue curve is obtained from neural recordings, the red curve is our model fit.

Using the dataset from Benucci et al., we found the tuning curves of the unadapted population by assuming that neural response is a function only of the difference between the preferred orientation and stimulus orientation (see Sec. 3.3). We then calculate the changes in preferred orientation of neurons for each of the experimental conditions. We then fit our model to both the unadapted tuning curves (Fig. 10b) and the changes in preferred orientation (see Sec. 3.3 for further details). The fit parameters for our model were $\sigma_\pi = 18°$, $\sigma_f = 14.2°$, and $\sigma_\phi = 20°$. Our model is compared to their experimental data in Fig. 11. Panels (a,d) show the tuning curves for the unadapted (blue) and adapted (red) populations, averaged across experiments. We can see that our model captures the suppression and repulsion of tuning curves around the origin. In panels (b,e) we examine the firing rate of the two populations averaged over the stimulus ensemble. The data and our model both show increased firing rate of the unadapted population around the adaptor stimulus (set to 0°). Additionally, we see homeostasis of firing rates, demonstrated by the uniformity of the red curves. Lastly, our model recapitulates the repulsion of preferred orientation found experimentally in panels (c,f). Repulsion here means that the change in preferred orientation has the same sign as the pre-adaptation preferred orientation. Repulsion is stronger nearer the adaptor, with the largest magnitude of repulsion occurring 30% from the adaptor in both cases.

3 Methods

3.1 Randomly generating high-correlation matrices

In our numerical simulations of firing rates within neuronal clusters, we randomly generated correlation matrices. Since we are focusing on behaviour within a cluster, such correlation matrices must have large and positive off-diagonal elements.

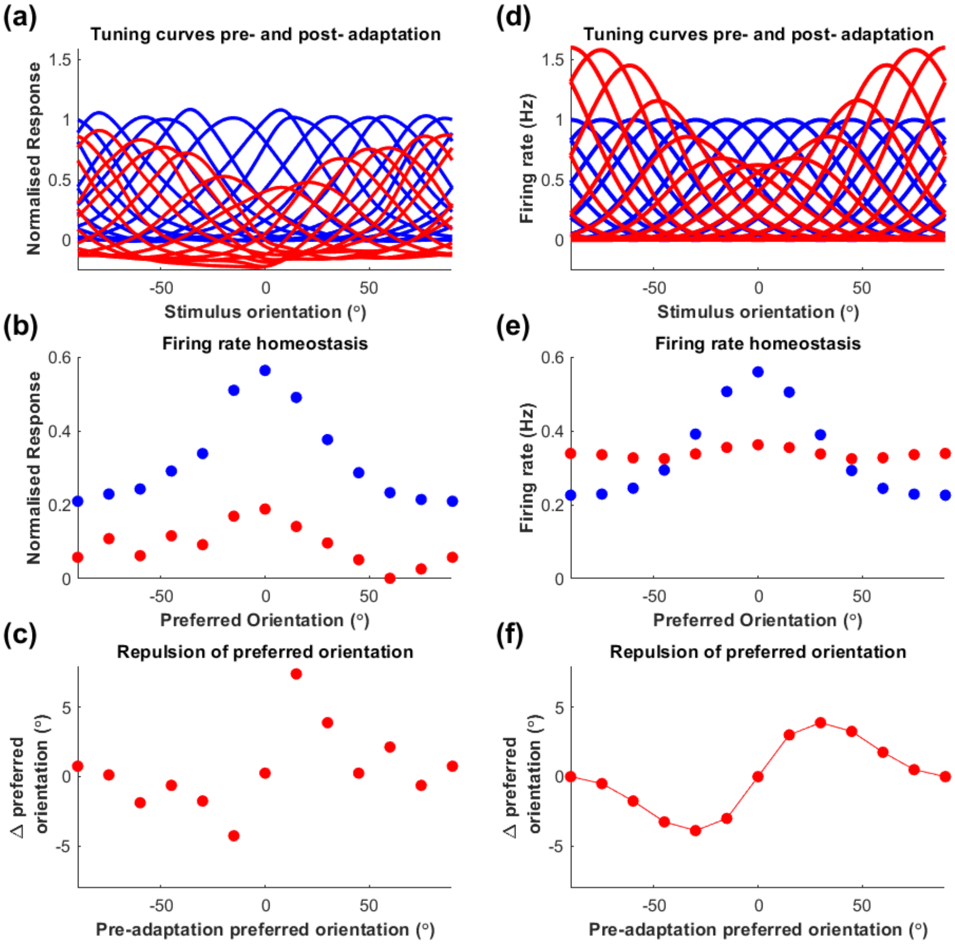

Figure 11: A comparison of the results obtained by Benucci et al. 2013 and our model's fit. (a) The tuning curves for the adapted (red) and unadapted (blue) populations, averaged across experimental conditions. (b) The firing rates of the adapted (red) and unadapted (blue) populations averaged across exposure to the non-uniform stimulus ensamble. (c) The repulsion of preferred orientations, obtained from the average tuning curves in panel a. (e-f) The same as (a-c), but for our model.

Our method is parameterised by α and q. q can be approximately interpreted as the average off-diagonal element (*i.e.* the correlation coefficient between different neurons within a cluster), while α controls the spread (*i.e.* variance) of correlations coefficients. The matrix ρ within each cluster is derived by normalising a covariance matrix Σ, given by

$$\Sigma = q\mathbf{1}_k\mathbf{1}_k^T + (1-q)U\hat{\boldsymbol{\xi}}U^T$$

U is an orthogonal matrix drawn uniformly (*i.e.* according to the Haar measure), and $\boldsymbol{\xi}$ is a vector of independent $\Gamma(\alpha, 1/\alpha)$ random variables.

3.2 Specification of environments for simulations

To generate $\tilde{\rho}(\epsilon)$ in our numerical simulations (Sec. 2.9) we first generate a covariance matrix $\Sigma(\epsilon)$, and let $\tilde{\rho}(\epsilon)$ be the corresponding correlation matrix. The procedure for generating $\Sigma(\epsilon)$ is as follows.

We randomly and independently sample two $K \times K$ random Gaussian matrices $R_0, R_1 \sim \mathcal{N}_{K\times K}(0, I_{K\times K})$ and obtain symmetric matrices $S_0 = R_0 + R_0^T$ and $S_1 = R_1 + R_1^T$. 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 $O(K)$.

Now for $\epsilon \in (0, 1)$, set $S(\epsilon) = (1-\epsilon)S_0 + \epsilon S_1$. Almost surely these matrices have a non-degenerate spectrum, and hence a unique representation as $S(\epsilon) = U(\epsilon)D(\epsilon)U(\epsilon)^T$ where $D(\epsilon)$ is diagonal with strictly decreasing eigenvalues and $U(\epsilon)$ is orthogonal. Moreover, $U(\epsilon)$ depends continuously on ϵ. Finally, we define $\Sigma(\epsilon) = U(\epsilon)\Lambda_1 U(\epsilon)^T$, where $\Lambda_1 = \text{diag}(1, 1/2, 1/3, \ldots, 1/K)$.

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

3.3 Fitting a homeostatic DDC to Benucci *et al.* 2013

The data we obtained from the Benucci *et al.* 2013 paper (Benucci et al., 2013) comprised 11 data sets. In each data set, the neural population was clustered into 12 groups based on preferred orientation. These neural populations were then exposed (on different occasions) to a distribution of 6-12 oriented gratings. This distribution was either uniform or biased, with one particular grating (arbitrarily set to 0°) having higher prevalence, either $30\%, 35\%, 40\%$, or 50%. We discarded all data sets with a 50% prevalence, since (Benucci et al., 2013) report that homeostasis was not obtained in this case. After exposure to the distribution of gratings, the responses (*i.e.*, spike count) of each cluster were then measured to a test set of 20 oriented gratings. These responses were then normalised via an affine transform so that the after exposure to the uniform stimulus ensemble, the normalised response ranged from 0 to 1. See their paper (Benucci et al., 2013) for further details.

To obtain the data average tuning curve after exposure to the uniform ensemble (see Fig. 10, (b), blue) we used a cubic spline to interpolate all the tuning curves of the populations after exposure to the uniform ensemble in each data set. These were then translated to center the preferred orientation at 0° and averaged. The width of the resultant curve at height e^{-2} was found, and divided by 4 to obtain the standard deviation of a Gaussian fit to the unadapted tuning curve (see Fig. 10, (b), red). In subsequent fitting, $\sqrt{\sigma_f^2 + \sigma_\phi^2}$ was constrained to equal this value.

This reduces the number of free parameters to 2, which we choose to be σ_π and σ_f. We fit these parameters to the change in preferred orientation of each data set. For each data set, cluster (*i.e.*, group of neurons with a similar preferred orientation) and condition (either exposure to uniform or biased ensemble) a smoothing kernel was applied before using cubic interpolation to generate tuning curves. The smoothing kernel was applied to ensure that none of the tuning curves were multimodal. 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 data sets with the same adaptor probability to give change in preferred orientation curves for each adaptor probability.

We then performed a grid search across values of σ_π and σ_f to compare these to the change in preferred orientation curves generated by our model. We chose the pair of values which minimised the sum of the absolute value of the difference between the curves obtained by our model and those in their data.

4 Discussion

4.1 Adaptation

The argument pursued in this paper is exclusively at the normative level; we make no specific claims about how (mechanistically) homeostasis may be achieved. We also do not commit ourselves to a specific time-scale upon which homeostasis is implemented.

The strategy of homeostatic adaptation can be easily implemented by biological neurons, since it only requires that each neuron keep track of its recent activity and, *e.g.*, adjusts its input weights or excitability accordingly. Thus, firing rate homeostasis is a biologically plausible and nearly optimal strategy for navigating the trade-off between coding fidelity and metabolic cost.

Moreover, homeostasis only requires first order statistics regarding firing rates. Other strategies for maximising information content - which may necessitate the use of more complicated statistics, such as correlations or entropy - would require the neural population to keep track of more features of its own responses and wait longer to obtain sufficient samples for reliable estimation of the relevant statistics. As such, homeostasis may be a primitive but effective initial adaptation mechanism, which can be implemented quickly and locally.

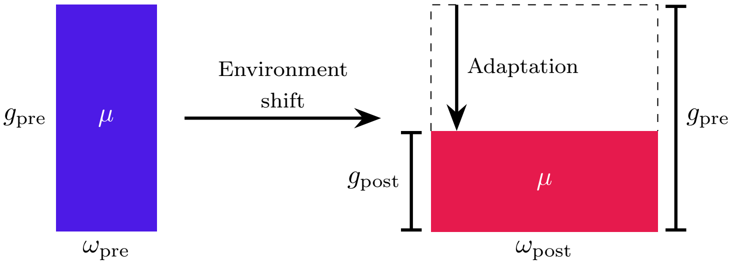

Figure 12: Adaptation can achieve homeostasis swiftly following an environmental shift, and thereby approximately maximise the information-energy trade off.

4.2 Fisher Information and Mutual Information

In our preceding arguments, we have used an upper bound on mutual information based upon approximating a non-Gaussian random variable as Gaussian. Another commonly used bound for mutual information is the Fisher Information lower bound (FILB), first derived by Bruner and Nadal (Brunel and Nadal, 1998). This states that

$$I(\boldsymbol{n};\boldsymbol{S}) \geq I_{\text{Fisher}} = H[\boldsymbol{s}] + \frac{1}{2}\int \ln\left(\frac{\det(J(\boldsymbol{s}))}{(2\pi e)^N}\right) P(\boldsymbol{s})d\boldsymbol{s} \tag{51}$$

where J is the Fisher Information of the neural response as a function of the stimulus.

For a Poissonian noise model, the Fisher Information is given by

$$J(\boldsymbol{s}) = \sum_{i=1}^{N} \frac{\nabla h_j \nabla^T h_j}{h_j} = \sum_{i=1}^{N} g_j \frac{\nabla \Omega_j \nabla^T \Omega_j}{\Omega_j} \tag{52}$$

In the framework we have been working in, neuron tuning curves are well approximated by their cluster tuning curves. This gives

$$J(\boldsymbol{s}) \approx \sum_{a=1}^{K} \left(\sum_{i \in c(a)} g_i \right) \frac{\nabla \tilde{\Omega}_a \nabla^T \tilde{\Omega}_a}{\tilde{\Omega}_a} = \sum_{a=1}^{K} \tilde{g}_a \frac{\nabla \tilde{\Omega}_a \nabla^T \tilde{\Omega}_a}{\tilde{\Omega}_a} \tag{53}$$

Therefore, we see that (neglecting corrections to the cluster tuning curves) both the Fisher information lower bound and the Gaussian upper bound depend on the cluster firing rates.

Moreover, both bounds display the same asymptotic behaviour in the gains, namely that $I(\tilde{\boldsymbol{n}}; \boldsymbol{I}) \sim \frac{1}{2}\ln(\tilde{g})$ as $\tilde{g} \to \infty$. This asymptotic behaviour is crucial to understanding the core of the results obtained above. We now examine where this behaviour arises from.

We first claim that mutual information scales like 1/2 log of the signal to noise ratio squared of the neural responses. The Gaussian upper bound approximates neural information flow as a Gaussian channel. In general, if signal $\boldsymbol{X} \sim N(0, \Sigma_s)$ is perturbed with Gaussian noise $\boldsymbol{Z} \sim N(0, \Sigma_n)$, then the resultant mutual information is

$$I(\boldsymbol{X}, \boldsymbol{X} + \boldsymbol{Z}) = \frac{1}{2} \ln \left(\det \left(I + \Sigma_n^{-1} \Sigma_s \right) \right) = \frac{1}{2} \ln(\det(I + \mathrm{SNR}^2)) \tag{54}$$

Up to an additive constant, this is the same expression we derive for the information term in $\tilde{\boldsymbol{\mathcal{L}}}$. In the Fisher Information Lower Bound (Eq. (53)) interpreting J as the signal-to-noise ratio gives the same qualitative behaviour. For Poissonian neurons, with Fisher Information given by Eq. (52), this interpretation comes from interpreting the numerator $\nabla \tilde{h}_a \nabla^T \tilde{h}_a$ as a squared signal term (quantifying the sensitivity of cluster a to changes in stimuli) and the denominator $\tilde{h}_a$ as a squared noise term (a baseline level of spike firing which obscures changes in firing rate).

Next, we claim that the signal scales linearly with the gain, and the noise scales sublinearly, specifically as $\sqrt{\tilde{g}}$. Putting these together, we obtain $\mathrm{SNR}^2 \sim \tilde{g}$. In the Gaussian channel case, the squared signal is the firing rate covariance $\mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}])$ and the squared noise is the mean spike count variance $\mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})]$. The squared signal term scales quadratically with the firing rate, and therefore the gain. For constant Fano factor neurons, the squared noise term scales linearly with the mean, and therefore linearly with the gain, giving the required result. Taking the inverse of the squared noise gives us the expression

$$\frac{1}{2} \ln \left(\det \left(I_K + \tilde{P} \hat{\tilde{\boldsymbol{\omega}}} \hat{\tilde{\boldsymbol{g}}} \right) \right) \tag{55}$$

which should be compared to the Gaussian Channel result Eq. (54) above. For the Fisher Information expression (Eq. (53)), the squared "signal" term $\nabla \tilde{h}_a \nabla^T \tilde{h}_a = \tilde{g}_a^2 \nabla \tilde{\Omega}_a \nabla^T \tilde{\Omega}_a$ scales quadratically, and the squared noise $\tilde{h}_a = \tilde{g}_a \tilde{\Omega}_a$ scales linearly.

We have therefore shown that for both the upper and lower bounds, the mutual information behaves approximately as 1/2 log of the gains. This requires that we are in a high signal-to-noise regime. For the Fisher Information, this requires a large number of neurons and a long coding interval, with the signal vectors $\nabla \tilde{\Omega}_a$ not too cluster together in stimulus space. Our condition on the matrix Δ is an analogous high signal-to-noise condition.

Since metabolic cost scales linearly with the gains, our target function $\tilde{\boldsymbol{\mathcal{L}}}$ has the asymptotic behaviour

$$\tilde{\boldsymbol{\mathcal{L}}} \sim \mu \ln(\tilde{g}) - \tilde{g}\tilde{\omega} \tag{56}$$

This function is maximised when $\tilde{g} = \mu/\tilde{\omega}$, which yields homeostasis.

Although the arguments that we have pursued in earlier sections revolve around use of one particular upper bound, this gives us reason to believe that our results are robust to modelling assumptions. We have demonstrated in this section that the mutual information is upper and lower bounded by terms that scale as 1/2 log gain, and therefore must also display the same scaling. Moreover, we have shown that this scaling, combined with a linear metabolic cost term, yields homeostasis as an optimal solution. There is therefore reason to believe that other approximations to mutual information will give similar results, provided we are in a high signal-to-noise ratio regime.

4.3 Comparison with Ganguli and Simoncelli 2014

The previous work in efficient coding theory which our framework most naturally builds off is that of Ganguli and Simoncelli (Ganguli and Simoncelli, 2014). They consider a collection of tuning curves h_j acting on a one-dimensional stimulus $\boldsymbol{s} \in [\boldsymbol{s}_{\min}, \boldsymbol{s}_{\max}]$. These curves arise as a warped and scaled convolution population; specifically, the tuning curve of neuron j is given by

$$h_j(\boldsymbol{s}) = g(D^{-1}(j))h(D(\boldsymbol{s}) - j) \tag{57}$$

where h is a unimodal tuning curve with peak at 0, g is a gain function, and $D(\boldsymbol{s})$ is an increasing function satisfying $D(\boldsymbol{s}_{\min}) = 0$, $D(\boldsymbol{s}_{\max}) = N$. The problem then becomes optimising the Fisher Information lower bound (51) in the parameters g and D, subject to a constraint on the total activity of the network which enforces population level homeostasis.

Under various smoothness assumptions, Ganguli and Simoncelli showed that the optimal solution to this problem is $D(\boldsymbol{s}) = \int_{\boldsymbol{s}_{\min}}^{\boldsymbol{s}} P(\boldsymbol{s})d\boldsymbol{s}$ and $g(\boldsymbol{s})$ is constant. The firing rate of neuron j is then given by

$$\begin{aligned}\int h_j(\boldsymbol{s})P(\boldsymbol{s})d\boldsymbol{s} &= \int gh(D(\boldsymbol{s}) - j)D'(\boldsymbol{s})d\boldsymbol{s} \\ &= g\int h(x-j)dx\end{aligned}$$

which (assuming circular boundary conditions) is independent of neuron identity j. This therefore gives identical firing rates and, provided adaptation attunes D to changes in P, homeostasis across environments.

Our results are naturally considered as an extension and generalisation of the results obtained by Ganguli and Simoncelli. Their framework places tight constraints on the form that neural representational curves possess, through the parameters g and D. In contrast, we make very few assumptions about the representational curves Ω, other than clustering and a high signal-to-noise ratio. Our framework can therefore be applied to a much larger variety of tuning curves which are *i.e.* multimodal, inhomogenous between neurons, and acting on a high-dimensional stimulus space. Ganguli and Simoncelli also only consider the case of Poissonian (uncorrelated, unit Fano factor) noise; in section 2.7 we show that our framework can also handle correlated and power-law noise. Additionally, we have computed the first-order correction to homeostasis that arises from the sharing of representations between neurons.

Moreover, since the density of preferred orientations D' is proportional to stimulus density, this framework predicts that adaptation should cluster tuning curves together around an adaptor stimulus. As we demonstrated in Sec. 2.13, our more general explanation of homeostatic coding coupled with Bayesian theories of representation can account for stimulus specific adaptation effects involving the repulsion of preferred orientation from an adaptor.

4.4 Other models of V1 Stimulus Specific Adaptation

Westrick et al. 2016 (Westrick et al., 2016) also give a model of the results of the Benucci et al. 2013 experiment (Benucci et al., 2013). Their model uses divisive normalisation (Carandini and Heeger, 2012) with adaptive weights to achieve homeostasis and stimulus specific repulsion. As discussed above in Sec. 2.12, Bayes-ratio coding (a special case of homeostatic DDCs) can be accomplished by such a divisive normalisation scheme. Our framework and modelling work therefore builds off their earlier results, giving a normative interpretation to their descriptive model.

Snow, Coen-Calli, and Schwartz 2016 (Snow et al., 2016) give two alternative models of stimulus specific adaptation in V1. In their models, divisive normalisation also plays a key role, but is given a different normative interpretation. They consider generative models of visual scenes in the class of *mixtures of Gaussian scale mixture* models (MGSM). In each of these, the response of a collection of linear filters applied to an image patch are assumed to be independently Gaussian, multiplied by a scaling variable across space and time. The task of V1 is then to infer which filters share a common scaling variable, and then "factor out" this scaling variable via divisive normalisation to compute the latent Gaussian variables.

Their models are primarily designed to capture adaptation effects over a very short time-scale (*i.e.* the last 9 frames of a video). Although they are able to capture effects such classical receptive field orientation adaptation effects well, their model not designed to account for longer time-scale homeostatic adaptation. Accordingly, their models did not successfully reproduce the results of Benucci *et al.* 2013. In particular, each of the two models could only account for either stimulus specific adaptation or neuron specific adaptation, but not both simultaneously. This opens up the possibility that Bayesian homeostatic coding coupled with a more sophisticated generative model could account for both short time-scale CRF effects and longer time-scale homeostatic effects. We believe that this would be a fruitful direction for future research.

4.5 Limitations

In order to obtain the analytic results throughout the paper, we have been forced to make a number of assumptions and simplifications. These necessarily limit our results. However, our hope is that the underlying logic of our argument (outlined in section 4.2) holds up despite our simplifications.

Firstly, instead of a direct maximisation of mutual information, we maximise a proxy - in this case an upper bound obtained by replacing the marginal entropy with the entropy of a corresponding Gaussian (see section 2.2). The practice of maximising bounds on mutual information for analytic tractability is standard within efficient coding theory (Ganguli and Simoncelli, 2014; Brunel and Nadal, 1998). In order for this substitution to not significantly affect our underlying argument, we require that the mutual information and our upper bound display the same behaviour in the relevant portion of our parameter regime. Note that this is different from the bound being tight. Given the argument in section 4.2, there is a good reason to believe that this is so.

Secondly, our analysis requires a semi-artificial imposition of a cluster structure onto the neural population (section 2.3). This involved grouping neurons by similar response properties. Our argument then proceeded by utilising the fact that within a cluster, signal correlations are very high and that between different clusters, signal correlations are generally quite low. In reality there is not such a sharp distinction between neurons with low correlations and neurons with high correlations. As this distinction weakens, the validity neglecting terms of order $\mathcal{O}(\epsilon^2)$ in the expansion $\mathcal{L} = \tilde{\mathcal{L}} + \epsilon\mathcal{L}_\epsilon + \mathcal{O}(\epsilon^2)$ begins to break down.

Lastly, instead of treating the full problem at once, we broke the problem down into two successive stages: first solving the between cluster problem by maximising $\tilde{\mathcal{L}}$ in the cluster gains $\tilde{\boldsymbol{g}}$ and then fixing these and maximising $\mathcal{L}_\epsilon$ in the individual neuron gains. The fundamental idea here is that there are directions in which the target functional $\mathcal{L}$ vary relatively slowly and directions in which variation is relatively fast. Instead of maximising the function over the entire space, we first move along the dimensions of fast variation, finding the maximum. We then move along the directions of slow variation, finding the maximum along these dimensions. Provided the extent of variation is sufficiently different, our hope is that such a procedure will not land us too far away from the global optimum.

4.6 Directions of future research

The results obtained here for the cluster gains correspond to the high signal-to-noise ratio limit. We do not give analytic results for the gains in the low signal-to-noise limit. This is for two reasons: firstly, there is good reason to believe that the nervous system is operating in a high signal-to-noise regime, since otherwise robustness of function would be difficult to explain; secondly, our analytically tractable approximation to the mutual information is only valid in this regime, and it is not clear what insight could be gained using it when signal-to-noise is low. Indeed, the Fisher Information lower bound for mutual information (see section 4.2) is also only valid for high signal-to-noise: when signal-to-noise is low, the bound can be violated (Yarrow et al., 2012; Bethge et al., 2002). There is thus presently no way to predict analytically features of neural coding when signal-to-noise is low using the Infomax principle. This presents an important opportunity for future work.

Our analysis here only considers a limited class of noise models. However, the key ideas used in our framework could potentially be applied to a wider class of noise models, such as exponential families (Beck et al., 2007).

Up to section 2.11, we make no assumptions about what the nervous system is attempting to represent via the curves $\Omega_j(\boldsymbol{s})$. We then apply our framework to a popular 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 representation, for example Probabilistic Population Codes (PPCs) (Beck et al., 2007).

4.7 Conclusion

We developed a theory of optimal gain modulation for combating noise in neural representations. We demonstrated that, when mean neural firing rates are not too small, selectivities are sufficiently sparse, and 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 within neural clusters. We further demonstrated that our framework could account for the optimality of a diversity of firing rates within neural clusters. By examining parameter values obtained from experiments, we confirmed that the conditions necessary for our analytic approximation to be valid did indeed hold. We further validated our approximation by demonstrating numerically that it performs well compared to the optimal gains. We then demonstrated that our results could be extended to both Poissonian noise and correlated, power-law noise, showing the full breadth of our framework. Having developed a normative theory of neural homeostasis, we show how homeostasis can lead to stimulus specific adaptation when coupled with Bayesian theories of representation. In particular, we focused on Bayesian DDCs, and their low-width kernel limit, the Bayes-ratio code.

5 Acknowledgements

We thank Máté Lengyel for helpful discussions and Guillaume Hennequin for comments on the manuscript. EY was supported by a studentship from the Engineering and Physical Sciences Research Council. YA was supported by UKRI Biotechnology and Biological Sciences Research Council research grant BB/X013235/1.

A Appendix

A.1 Analytic expression for $\mathcal{L}$

The functional $\mathcal{L}^0(\boldsymbol{g}) = 2\mu I(\boldsymbol{n};\boldsymbol{s}) - \sum_{j=1}^{N} \mathbb{E}[n_j]$ can be upper bounded by $\mathcal{L}$ using

$$H[\boldsymbol{n}] \leq H[\mathcal{N}(\boldsymbol{h}(\boldsymbol{s}), \mathrm{Cov}(\boldsymbol{n}))] = \frac{1}{2}\ln\left((2\pi e)^N \det\left(\mathrm{Cov}(\boldsymbol{n})\right)\right)$$

giving

$$\mathcal{L}^0(\boldsymbol{g}) \leq \mathcal{L}(\boldsymbol{g}) = 2\mu\left(\frac{1}{2}\ln\left((2\pi e)^N \det(\mathrm{Cov}(\boldsymbol{n}))\right) - H[\boldsymbol{n}|\boldsymbol{s}]\right) - \sum_{j=1}^{N} \mathbb{E}[h_j(\boldsymbol{s})]$$

In this section of the appendix we demonstrate that, up to an additive constant,

$$\mathcal{L}(\boldsymbol{g}) = \mu\ln\left(\det\left(I_N + \hat{\mathbf{CV}}\rho\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)\right) - \sum_{j=1}^{N} g_j\omega_j$$

Firstly, we demonstrate that $\mathbb{E}[n_j] = g_j \omega_j$.

$$\begin{aligned}
\mathbb{E}[n_j] &= \mathbb{E}[\mathbb{E}[n_j|\boldsymbol{s}]] \\
&= \mathbb{E}[h_j(\boldsymbol{s})] \\
&= \mathbb{E}[g_j \Omega_j(\boldsymbol{s})] \\
&= g_j \mathbb{E}[\Omega_j(\boldsymbol{s})] \\
&= g_j \omega_j
\end{aligned}$$

Next, we derive an expression for $H[\boldsymbol{n}|\boldsymbol{s}]$. Recall that $n_i|\boldsymbol{s} \sim N(h_i(\boldsymbol{s}), h_i(\boldsymbol{s}))$ independently. In this case,

$$\begin{aligned}
H[\boldsymbol{n}|\boldsymbol{s}] &= \sum_{j=1}^{N} H[n_j|\boldsymbol{s}] \\
H[n_j|\boldsymbol{s}] &= \int H[N(h_j(\boldsymbol{s}), h_j(\boldsymbol{s}))] P(\boldsymbol{s}) d\boldsymbol{s} \\
&= \int \frac{1}{2} \ln(2\pi e h_j(\boldsymbol{s})) P(\boldsymbol{s}) d\boldsymbol{s} \\
&= \frac{1}{2} \ln(2\pi e) + \frac{1}{2} \ln(g_j) + \frac{1}{2} \mathbb{E}[\ln(\Omega_j(\boldsymbol{s}))] \\
H[\boldsymbol{n}|\boldsymbol{s}] &= \frac{N}{2} \ln(2\pi e) + \frac{1}{2} \sum_{j=1}^{N} \ln(g_i) + \sum_{j=1}^{N} \frac{1}{2} \mathbb{E}\left[\ln(\Omega_j(\boldsymbol{s}))\right]
\end{aligned}$$

Lastly, we find an expression for $\text{Cov}(\boldsymbol{n})$. We use the decomposition $\text{Cov}(\boldsymbol{n}) = \mathbb{E}[\text{Cov}(\boldsymbol{n}|\boldsymbol{s})] + \text{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}])$.

$$\begin{aligned}
\text{Cov}(\boldsymbol{n}|\boldsymbol{s}) &= \text{Var}(\hat{\boldsymbol{n}}|\boldsymbol{s}) \\
&= \hat{\boldsymbol{h}}(\boldsymbol{s}) \\
\mathbb{E}[\text{Cov}(\boldsymbol{n}|\boldsymbol{s})] &= \mathbb{E}[\hat{\boldsymbol{h}}(\boldsymbol{s})] \\
&= \hat{\boldsymbol{g}} \hat{\boldsymbol{\omega}} \\
\text{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) &= \text{Cov}(\boldsymbol{h}(\boldsymbol{s})) \\
&= \hat{\boldsymbol{g}} \text{Cov}(\boldsymbol{\Omega}(\boldsymbol{s})) \hat{\boldsymbol{g}} \\
\text{Cov}(\boldsymbol{\Omega}(\boldsymbol{s}))_{ij} &= \\
&= \sqrt{\text{Var}(\Omega_i(\boldsymbol{s}))} \frac{\text{Cov}(\Omega_i(\boldsymbol{s}), \Omega_j(\boldsymbol{s}))}{\sqrt{\text{Var}(\Omega_i(\boldsymbol{s}) \ \text{Var}(\Omega_j(\boldsymbol{s})))}} \sqrt{\text{Var}(\Omega_j(\boldsymbol{s}))} \\
&= \omega_i \frac{\sqrt{\text{Var}(\Omega_i(\boldsymbol{s}))}}{\omega_i} \rho_{ij} \frac{\sqrt{\text{Var}(\Omega_j(\boldsymbol{s}))}}{\omega_j} \omega_j \\
&= \omega_i \text{CV}_i \rho_{ij} \text{CV}_j \omega_j \\
\text{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) &= \hat{\boldsymbol{g}} \hat{\boldsymbol{\omega}} \hat{\mathbf{CV}} \rho \hat{\mathbf{CV}} \hat{\boldsymbol{\omega}} \hat{\boldsymbol{g}}
\end{aligned}$$

Substituting these into the covariance, we obtain

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{n}) &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) \\
&= \hat{\boldsymbol{g}}\hat{\boldsymbol{\omega}} + \hat{\boldsymbol{g}}\hat{\boldsymbol{\omega}}\hat{\mathbf{CV}}\rho\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}} \\
&= \hat{\boldsymbol{g}}\hat{\boldsymbol{\omega}}\left(I_N + \hat{\mathbf{CV}}\rho\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right) \\
\ln\left(\det\left(\mathrm{Cov}(\boldsymbol{n})\right)\right) &= \sum_{j=1}^{N}\ln(g_j) + \sum_{j=1}^{N}\ln(\omega_j) + \ln\left(\det\left(I_N + \hat{\mathbf{CV}}\rho\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)\right)
\end{aligned}$$

Putting all these parts together gives us

$$\begin{aligned}
\ln\left((2\pi e)^N \det(\mathrm{Cov}(\boldsymbol{n}))\right) - 2H[\boldsymbol{n}|\boldsymbol{s}] &= N\ln(2\pi e) + \ln(\det(\mathrm{Cov}(\boldsymbol{n}))) \\
&- N\ln(2\pi e) - \sum_{j=1}^{N}\ln(g_i) - \sum_{j=1}^{N}\mathbb{E}\left[\ln(\Omega_j(\boldsymbol{s}))\right] \\
&= \ln(\det(\mathrm{Cov}(\boldsymbol{n}))) - \sum_{j=1}^{N}\ln(g_j) - \sum_{j=1}^{N}\mathbb{E}\left[\ln(\Omega_j(\boldsymbol{s}))\right] \\
&= \ln\left(\det\left(I_N + \hat{\mathbf{CV}}\rho\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)\right) \\
&- \sum_{j=1}^{N}\mathbb{E}\left[\ln\left(\frac{\Omega_j(\boldsymbol{s})}{\omega_j}\right)\right]
\end{aligned}$$

Plugging this in, we obtain

$$\mathcal{L}(\boldsymbol{g}) = \mu\ln\left(\det\left(I_N + \hat{\mathbf{CV}}\rho\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)\right) - \sum_{j=1}^{N} g_j\omega_j - \mu\sum_{j=1}^{N}\mathbb{E}\left[\ln\left(\frac{\Omega_j(\boldsymbol{s})}{\omega_j}\right)\right]$$

A.2 Expanding $\tilde{\mathcal{L}}$ in ϵ

We now take $\rho = \tilde{\rho} \otimes \mathbf{1}_k\mathbf{1}_k^T - \epsilon T$. Analogously to Eq. (9), we define

$$\tilde{P} = \hat{\tilde{\mathbf{CV}}}\,\tilde{\rho}\,\hat{\tilde{\mathbf{CV}}}\hat{\tilde{\boldsymbol{\omega}}} \tag{58}$$

Substituting ρ into Eq. (8), $\mathcal{L}$ becomes (up to an additive constant)

$$\begin{aligned}
\mathcal{L}(\boldsymbol{g}) &= \mu \ln \det \Big(I_N + \hat{\mathbf{CV}}[\tilde{\rho} \otimes \mathbf{1}_k \mathbf{1}_k^T - \epsilon T]\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}})\Big) - \sum_{j=1}^{N} g_j \omega_j \\
&= \mu \ln \det \Big(I_N + \hat{\mathbf{CV}}[\tilde{\rho} \otimes \mathbf{1}_k \mathbf{1}_k^T]\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}} - \epsilon \hat{\mathbf{CV}} T \hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}})\Big) - \sum_{j=1}^{N} g_j \omega_j \\
&= \mu \ln \det \Big(I_N + [\tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T]\hat{\boldsymbol{g}} - \epsilon \hat{\mathbf{CV}} T \hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}})\Big) - \sum_{j=1}^{N} g_j \omega_j \\
&= \mu \ln \Big(\det \Big(I_N + \tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \; \hat{\boldsymbol{g}}\Big)\Big) - \sum_{j=1}^{N} g_j \omega_j \\
&\quad - \epsilon \operatorname{Tr}\bigg(\Big(I_N + \tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \; \hat{\boldsymbol{g}}\Big)^{-1} \hat{\mathbf{CV}} \, T \, \hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}})\bigg) \\
&\quad + \mathcal{O}(\epsilon^2)
\end{aligned}$$

We call the zeroth and first order term $\tilde{\mathcal{L}}$ and $\mathcal{L}_\epsilon$ respectively, thereby obtaining

$$\mathcal{L} = \tilde{\mathcal{L}} + \epsilon \mathcal{L}_\epsilon + \mathcal{O}(\epsilon^2)$$

We now show that the zeroth order term depends only on the cluster rates $\tilde{g}_a = \sum_{i \in c(a)} g_i$. We start by showing that

$$\ln \Big(\det \Big(I_N + \tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \hat{\boldsymbol{g}}\Big)\Big)$$

can be re-written as

$$\ln \Big(\det \Big(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}}\Big)\Big)$$

Consider the following basis for $\mathbb{R}^N$. For each cluster $a = 1, \ldots, K$ define $\boldsymbol{v}^a$ given by

$$\boldsymbol{v}^a = \boldsymbol{e}_a \otimes \mathbf{1}_k$$

where $\boldsymbol{e}_a$ is the ath standard basis element on $\mathbb{R}^K$. Then

$$\tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \hat{\boldsymbol{g}} \boldsymbol{v}^a = (\tilde{P}\hat{\tilde{\boldsymbol{g}}}\boldsymbol{e}_a) \otimes 1_k = \sum_{b=1}^{K} \Big[\tilde{P}\hat{\tilde{\boldsymbol{g}}}\Big]_{ba} \boldsymbol{v}^b$$

We then extend this to a basis of $\mathbb{R}^N$ as follows. For each cluster a, we add an additional $k-1$ vectors which have support only within cluster a. Specifically, for each index $i \neq (a-1)k+1$, we add the vector which is equal to $(-1/g_{(a-1)k+1}, 0, \ldots, 0, 1/g_i, 0, \ldots, 0)$ on cluster a and zero elsewhere. Note that all such vectors are in the kernel of $\tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \hat{\boldsymbol{g}}$. In this basis of $\mathbb{R}^N$, $\tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \hat{\boldsymbol{g}}$ is represented by the block matrix

$$\begin{pmatrix} \tilde{P}\hat{\tilde{\boldsymbol{g}}} & 0 \\ 0 & 0 \end{pmatrix}$$

Accordingly,

$$\begin{aligned}
\ln(\det(I_N + \tilde{P} \otimes \mathbf{1}_k \mathbf{1}_k^T \hat{\boldsymbol{g}}))) &= \ln \left(\det \begin{pmatrix} I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}} & 0 \\ 0 & I_{N-K} \end{pmatrix}\right) \\
&= \ln \Big(\det(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}}) \det(I_{N-K})\Big) \\
&= \ln \Big(\det(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}})\Big)
\end{aligned}$$

Next, note that

$$\begin{aligned}\sum_{i=1}^{N} g_i \omega_i &\approx \sum_{a=1}^{K} \sum_{i \in c(a)} g_i \tilde{\omega}_a \\ &= \sum_{a=1}^{K} \tilde{g}_a \tilde{\omega}_a\end{aligned}$$

This therefore allows us to write that

$$\tilde{\mathcal{L}} = \ln \Big(\det(I_K + \tilde{P} \hat{\tilde{\boldsymbol{g}}}) \Big) - \sum_{a=1}^{K} \tilde{g}_a \tilde{\omega}_a$$

as claimed in Sec. 2.3.

A.3 First-order maximisation of $\tilde{\mathcal{L}}$

We now consider optimising

$$\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}) = \mu \ln \Big(\det \Big(I_K + \tilde{P} \hat{\tilde{\boldsymbol{g}}} \Big) \Big) - \sum_{a=1}^{K} \tilde{g}_a \tilde{\omega}_a$$

Taking the derivative with respect to $\tilde{g}_a$, and setting to zero, we obtain the condition

$$\mu \Big[(I_K + \tilde{P} \hat{\tilde{\boldsymbol{g}}})^{-1} \tilde{P} \Big]_{aa} = \tilde{\omega}_a$$

This can be re-written as

$$\frac{\mu}{\tilde{g}_a} (I_K + \hat{\tilde{\boldsymbol{g}}}^{-1} \tilde{P}^{-1})^{-1}_{aa} = \tilde{\omega}_a$$

We then note that

$$\begin{aligned}\tilde{P} &= \hat{\tilde{\mathbf{CV}}} \tilde{\rho} \hat{\tilde{\mathbf{CV}}} \hat{\tilde{\boldsymbol{\omega}}} \\ \tilde{P}^{-1} &= \hat{\tilde{\boldsymbol{\omega}}}^{-1} \hat{\tilde{\mathbf{CV}}}^{-1} \tilde{\rho}^{-1} \hat{\tilde{\mathbf{CV}}}^{-1} \\ &= \mu \hat{\tilde{\boldsymbol{\omega}}}^{-1} \Delta \\ \tilde{\omega}_a &= \frac{\mu}{\tilde{g}_a} \Big(I_K + \mu \hat{\tilde{\boldsymbol{g}}}^{-1} \hat{\tilde{\boldsymbol{\omega}}}^{-1} \Delta \Big)^{-1}_{aa}\end{aligned}$$

Since Δ is small, we can use a first-order Neumann expansion for $\Big(I_K + \mu \hat{\tilde{\boldsymbol{g}}}^{-1} \hat{\tilde{\boldsymbol{\omega}}}^{-1} \Delta \Big)^{-1}$. This expansion gives:

$$\Big(I_K + \mu \hat{\tilde{\boldsymbol{g}}}^{-1} \hat{\tilde{\boldsymbol{\omega}}}^{-1} \Delta \Big)^{-1} \approx I_K - \mu \hat{\tilde{\boldsymbol{g}}}^{-1} \hat{\tilde{\boldsymbol{\omega}}}^{-1} \Delta$$

Plugging this in, we obtain

$$\tilde{\omega}_a \approx \frac{\mu}{\tilde{g}_a} \left(1 - \frac{\mu}{\tilde{g}_a \tilde{\omega}_a} \Delta_{aa} \right) \tag{59}$$

which can be rearranged to the quadratic equation

$$\left(\frac{\tilde{g}_a \tilde{\omega}_a}{\mu} \right)^2 = \left(\frac{\tilde{g}_a \tilde{\omega}_a}{\mu} \right) - \Delta_{aa} \tag{60}$$

To zeroth-order in Δ, this has solution

$$\frac{\tilde{g}_a\tilde{\omega}_a}{\mu} = 1$$

and to first-order, the solution is

$$\frac{\tilde{g}_a\tilde{\omega}_a}{\mu} = 1 - \Delta_{aa}$$

A.4 Manipulating $\mathcal{L}_\epsilon$

Recall that

$$\mathcal{L}(\boldsymbol{g}) = \tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}) + \epsilon\mathcal{L}_\epsilon(\boldsymbol{g}) + \mathcal{O}(\epsilon^2)$$

In this section, we derive a simplified form of $\mathcal{L}_\epsilon$. We begin with the expression given in App. A.2. This is

$$\mathcal{L}_\epsilon(\boldsymbol{g}) = -\mathrm{tr}\left(\left(I_N + \tilde{P} \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}\right)^{-1} \hat{\mathbf{CV}}T\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)$$

We will write $\tilde{P} \otimes \mathbf{1}_k\mathbf{1}_k^T = (I_K \otimes \mathbf{1}_k)(\tilde{P} \otimes \mathbf{1}_k^T)$ and apply the Woodbury matrix identity to obtain that

$$\begin{aligned}
\left(I_N + \tilde{P} \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}\right)^{-1} &= I_N - (I_K \otimes \mathbf{1}_k)\left(I_K + (\tilde{P} \otimes \mathbf{1}_k^T)\hat{\boldsymbol{g}}(I_K \otimes \mathbf{1}_k)\right)^{-1}(\tilde{P} \otimes \mathbf{1}_k^T)\hat{\boldsymbol{g}} \\
(\tilde{P} \otimes \mathbf{1}_k^T)\hat{\boldsymbol{g}}(I_K \otimes \mathbf{1}_k) &= \tilde{P}\hat{\tilde{\boldsymbol{g}}} \\
\left(I_N + \tilde{P} \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}\right)^{-1} &= I_N - \left(\left(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}}\right)^{-1}\tilde{P}\right) \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}
\end{aligned}$$

Substituting in and using linearity of the trace, we obtain

$$\begin{aligned}
\mathcal{L}_\epsilon = &-\mathrm{tr}\left(\hat{\mathbf{CV}}T\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right) \\
&+ \mathrm{tr}\left(\left(\left(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}}\right)^{-1}\tilde{P}\right) \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}\hat{\mathbf{CV}}T\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)
\end{aligned}$$

The first term is equal to

$$-\sum_{i=1}^{N} \mathrm{CV}_i^2 g_i\omega_i T_{ii}$$

But recall that $\rho = \tilde{\rho} \otimes \mathbf{1}_k\mathbf{1}_k^T - \epsilon T$. Since $\tilde{\rho}$ and ρ are both correlation matrices, their diagonal elements are both equal to 1, and therefore the diagonal elements of T are all zero. So the first term vanishes. This leaves us with the expression

$$\mathcal{L}_\epsilon = \mathrm{tr}\left(\left(\left(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}}\right)^{-1}\tilde{P}\right) \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}\hat{\mathbf{CV}}T\hat{\mathbf{CV}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)$$

Note that this is exact, and does not require taking any truncations. We now simplify this functional by truncating to zeroth-order in an expansion in

$$\Delta = \left(\mu\hat{\tilde{\mathbf{CV}}}\tilde{\rho}\hat{\tilde{\mathbf{CV}}}\right)^{-1}$$

In A.3, we demonstrate that to zeroth-order, $\tilde{g}_a = \mu/\tilde{\omega}_a$, and therefore $\tilde{P}\hat{\tilde{\boldsymbol{g}}} = \Delta^{-1}$. Therefore

$$\begin{aligned}\left(I_K + \tilde{P}\hat{\tilde{\boldsymbol{g}}}\right)^{-1}\tilde{P} &= \left(I + \Delta^{-1}\right)^{-1}\tilde{P}\\ &= (\Delta + I_K)^{-1}\Delta\tilde{P}\\ &= (\Delta + I_K)^{-1}\mu\hat{\boldsymbol{\omega}}^{-1}\\ &\approx \mu\hat{\boldsymbol{\omega}}^{-1}\end{aligned}$$

where we have taken a zeroth-order expansion in Δ. We substitute this into our functional. We will use $a(i)$ to denote the index of the cluster to which neuron i belongs.

$$\begin{aligned}\mathcal{L}_\epsilon &= \mathrm{tr}\left(\mu\hat{\boldsymbol{\omega}}^{-1} \otimes \mathbf{1}_k\mathbf{1}_k^T\hat{\boldsymbol{g}}\mathbf{C}\hat{\mathbf{V}}T\mathbf{C}\hat{\mathbf{V}}\hat{\boldsymbol{\omega}}\hat{\boldsymbol{g}}\right)\\ &= \sum_{i=1}^{N}\frac{\mu}{\tilde{\omega}_{a(i)}}\sum_{j:a(j)=a(i)} g_j\mathrm{CV}_j T_{ji}\mathrm{CV}_i\omega_i g_i\\ &= \sum_{i=1}^{N}\frac{\mu}{\tilde{\omega}_{a(i)}}\sum_{j:a(j)=a(i)} g_j\mathrm{C\tilde{V}}_{a(j)} T_{ji}\mathrm{C\tilde{V}}_{a(i)}\tilde{\omega}_{a(i)} g_i\\ &= \sum_{i=1}^{N}\mu\mathrm{C\tilde{V}}_{a(i)}^2\sum_{j:a(j)=a(i)} g_j T_{ji} g_i\end{aligned}$$

Note that this has decoupled the problem between clusters. In particular, the problem now can be written as a collection the following K problems:

$$\{g_i\}_{i\in c(a)} \in \arg\max\left\{\sum_{i,j\in c(a)} g_i T_{ij} g_j\right\} \text{ s.t. } \sum_{i\in c(a)} g_i = \tilde{g}_a,\ g_i \geq 0$$

However, recall that for $i, j \in c(a)$, $T_{ij} = (\tilde{\rho}_{aa} - \rho_{ij})/\epsilon$. Substituting this in gives the problems

$$\{g_i\}_{i\in c(a)} \in \arg\max\left\{\tilde{\rho}_{aa}\sum_{i,j\in c(a)} g_i g_j - \sum_{i,j\in c(a)} g_i\rho_{ij}g_j\right\} \text{ s.t. } \sum_{i\in c(a)} g_i = \tilde{g}_a,\ g_i \geq 0$$

We can use that $\sum_{i\in c(a)} g_i = \tilde{g}_a$ to rewrite this as

$$\{g_i\}_{i\in c(a)} \in \arg\min\left\{\sum_{i,j\in c(a)} g_i\rho_{ij}g_j\right\} \text{ s.t. } \sum_{i\in c(a)} g_i = \tilde{g}_a,\ g_i \geq 0$$

A.5 Estimates for $\mathrm{tr}(\tilde{\rho}^{-1})$

In this appendix, we give estimates for $[\tilde{\rho}^{-1}]_{aa}$ through the average value $\mathrm{tr}\left(\tilde{\rho}^{-1}\right)/K$. Stringer *et al.* (Stringer et al., 2019) found that in response to a ecological visual stimuli, signal correlations of neural responses display a $1/n$ power-law decay. The eigenvalues of $\tilde{\rho}$ are therefore $\lambda_n = A_1/n$ where A_1

normalises the trace to be K. In particular,

$$\begin{aligned}
A_1 &= \frac{K}{\sum_{n=1}^{K} \frac{1}{n}} \\
\frac{K}{A_1} &= \sum_{n=1}^{K} \frac{1}{n} \\
&\approx \int_1^K \frac{1}{x} dx \\
&= \ln(K) \\
A_1 &= \frac{K}{\ln(K)}
\end{aligned}$$

We then calculate

$$\begin{aligned}
\operatorname{tr}\left(\tilde{\rho}^{-1}\right) &= \sum_{n=1}^{K} \frac{1}{\lambda_n} \\
&= \frac{1}{A_1} \sum_{n=1}^{K} n \\
&= \frac{1}{A_1} \frac{K(K+1)}{2} \\
\frac{\operatorname{tr}\left(\tilde{\rho}^{-1}\right)}{K} &= \frac{1}{A_1} \frac{(K+1)}{2} \\
&\approx \frac{\ln(K)}{2}
\end{aligned}$$

A.6 Homeostatic propagation

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

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

$$H_m(\boldsymbol{s}) = \sum_{i=1}^{N} W_{mi} h_i(\boldsymbol{s}) \tag{61}$$

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

$$\Gamma_m(\boldsymbol{s}) = \sum_{j=1}^{N} w_{mj} \Omega_j(\boldsymbol{s}) \tag{62}$$

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

$$H_m(\boldsymbol{s}) = \frac{\chi \Gamma_m(\boldsymbol{s})}{\gamma_m}$$

where $\gamma_m = \mathbb{E}[\Gamma_m(\boldsymbol{s})]$. But then

$$\begin{aligned}
\gamma_m &= \mathbb{E}[\Gamma_m(\boldsymbol{s})] \\
&= \sum_{j=1}^{N} w_{mj} \mathbb{E}[\Omega_j(\boldsymbol{s})] \\
&= \sum_{j=1}^{N} w_{mj} \omega_j \\
\chi \Gamma_m(\boldsymbol{s}) &= \sum_{j=1}^{N} w_{mj} \chi \Omega_j(\boldsymbol{s}) \\
&= \sum_{j=1}^{N} w_{mj} \omega_j h_j(\boldsymbol{s})
\end{aligned}$$

Substituting this in, we get

$$H_m(\boldsymbol{s}) = \frac{\sum_{j=1}^{N} w_{mj} \omega_j h_j(\boldsymbol{s})}{\sum_{i=1}^{N} w_{mi} \omega_i} \tag{63}$$

Comparing coefficients, we can see that

$$W_{mj} = \frac{w_{mj} \omega_j}{\sum_{i=1}^{N} w_{mi} \omega_i}$$

As claimed.

A.7 Upper bound for Poissonian noise

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

$$\tilde{n}_a | \boldsymbol{s} \sim \text{Poisson}\left(\tilde{g}_a \tilde{\Omega}_a\right) \tag{64}$$

Here we derive a Gaussian upper bound to the mutual information, and show that an approximation to it leads to the same objective $\tilde{\boldsymbol{\mathcal{L}}}$ derived in the uncorrelated, unit Fano-factor Gaussian noise case. We start with the objective

$$\tilde{\mathcal{L}}^0(\tilde{\boldsymbol{g}}) = 2\mu I(\tilde{\boldsymbol{n}}; \boldsymbol{s}) - \sum_{a=1}^{K} \tilde{g}_a \tilde{\omega}_a \tag{65}$$

We decompose $I(\tilde{\boldsymbol{n}}; \boldsymbol{s}) = H[\tilde{\boldsymbol{n}}] - H[\tilde{\boldsymbol{n}}|\boldsymbol{s}]$. We will once again upper bound the marginal entropy $H[\tilde{\boldsymbol{n}}]$. 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 $\tilde{\boldsymbol{n}}$. Then $\tilde{\boldsymbol{n}} + \boldsymbol{U}$ is a continuous random variable. We apply the Gaussian bound to this. Let p be the p.d.f. of $\tilde{\boldsymbol{n}} + \boldsymbol{U}$,

$\mathcal{P}$ the p.m.f. of $\tilde{\boldsymbol{n}}$, and u the p.d.f. of $\boldsymbol{U}$.

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

Putting this together gives us the upper bound

$$H[\tilde{\boldsymbol{n}}] \leq \frac{1}{2}\ln\left((2\pi e)^K\det\left(\frac{1}{12}I_K+\mathrm{Cov}(\tilde{\boldsymbol{n}})\right)\right) \tag{66}$$

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

$$H[\tilde{\boldsymbol{n}}|\boldsymbol{s}] = \sum_{a=1}^{K}\int H[\mathrm{Poisson}(\tilde{g}_a\tilde{\Omega}_a(\boldsymbol{s}))]P(\boldsymbol{s})d\boldsymbol{s} \tag{67}$$

We now make use of the assumption discussed in Sec. 2.7 that the representational curves $\tilde{\Omega}_a$ have a baseline, and in particular $\tilde{\Omega}_a \gg 5\tilde{\omega}_a/\mu$ everywhere. Under this condition we can use the fact that for fixed $\boldsymbol{s}$,

$$H[\mathrm{Poisson}(\tilde{g}_a\tilde{\Omega}_a(\boldsymbol{s}))] \approx H[\mathcal{N}(\tilde{g}_a\tilde{\Omega}_a(\boldsymbol{s}),\tilde{g}_a\tilde{\Omega}_a(\boldsymbol{s}))] \tag{68}$$

This means we can obtain the approximate upper bound:

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

where $\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}})$ is the same functional that we defined earlier for Gaussian random variables, Eq. (12).

A.8 Power-law and correlated noise

In this appendix, we consider the case of power-law, correlated noise for the clusters. We will take cluster spike counts to be given by

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

where $\Sigma(\boldsymbol{s})$ is the stimulus-dependent noise correlation matrix, and $0 < \alpha < 2$ is a scaling parameter.

The gains will once again be taken to maximise

$$\tilde{\mathcal{L}}(\tilde{\boldsymbol{g}}) = \mu \left(\ln \left((2\pi e)^K \det \left(\mathrm{Cov}(\tilde{\boldsymbol{n}}) \right) \right) - 2H[\tilde{\boldsymbol{n}}|\boldsymbol{s}] \right) - \sum_{a=1}^{K} \tilde{g}_a \tilde{\omega}_a$$

Much as in appendix A.1, we will simplify this expression. To keep our derivation clean, we will use the expression $\tilde{r}_a = \tilde{g}_a \tilde{\omega}_a$.

We start by deriving an expression for $2H[\boldsymbol{n}|\boldsymbol{s}]$.

$$\begin{aligned}
2H[\boldsymbol{n}|\boldsymbol{s}] &= \int 2H \left[\mathcal{N} \left(\tilde{\boldsymbol{h}}(\boldsymbol{s}), \sigma^2 \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \right) \right] P(\boldsymbol{s}) d\boldsymbol{s} \\
&= \int \ln \left((2\pi e)^K \det \left(\sigma^2 \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \right) \right) P(\boldsymbol{s}) d\boldsymbol{s} \\
&= \ln \left(\det \left(\hat{\tilde{\boldsymbol{r}}}^{2\alpha} \right) \right) + K \ln(2\pi e \sigma^2) \\
&+ \int \ln \left(\det \left(\hat{\tilde{\boldsymbol{\omega}}}^{-\alpha} \hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) \hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^\alpha \hat{\tilde{\boldsymbol{\omega}}}^{-\alpha} \right) \right) P(\boldsymbol{s}) d\boldsymbol{s} \\
&= \ln \left(\det \left(\hat{\tilde{\boldsymbol{r}}}^{2\alpha} \right) \right) + \text{ const.} \\
&= 2\alpha \sum_{a=1}^{K} \ln(\tilde{r}_a) + \text{ const.}
\end{aligned}$$

Next, we find an expression for $\mathrm{Cov}(\tilde{\boldsymbol{n}})$. We use the decomposition $\mathrm{Cov}(\tilde{\boldsymbol{n}}) = \mathbb{E}[\mathrm{Cov}(\tilde{\boldsymbol{n}}|\boldsymbol{s})] + \mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{n}}|\boldsymbol{s}])$.

$$\begin{aligned}
\mathrm{Cov}(\tilde{\boldsymbol{n}}|\boldsymbol{s}) &= \sigma^2 \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \Sigma(\boldsymbol{s}) \hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha \\
&= \sigma^2 \sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]} \times \frac{\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha}{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]}} \Sigma(\boldsymbol{s}) \frac{\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha}{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]}} \times \sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]} \\
\mathbb{E}[\mathrm{Cov}(\tilde{\boldsymbol{n}}|\boldsymbol{s})] &= \sigma^2 \sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]} W \sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]}
\end{aligned}$$

Where we have defined

$$W := \mathbb{E} \left[\frac{\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha}{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]}} \Sigma(\boldsymbol{s}) \frac{\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^\alpha}{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]}} \right] = \mathbb{E} \left[\frac{\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^\alpha}{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}} \Sigma(\boldsymbol{s}) \frac{\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^\alpha}{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}} \right]$$

which is the stimulus-averaged noise correlation matrix. Next, we find an expression for $\mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{n}}|\boldsymbol{s}])$

$$\begin{aligned}
\mathbb{E}[\tilde{\boldsymbol{n}}|\boldsymbol{s}] &= \tilde{\boldsymbol{h}}(\boldsymbol{s}) \\
\mathrm{Cov}(\mathbb{E}[\tilde{\boldsymbol{n}}|\boldsymbol{s}]) &= \mathrm{Cov} \left(\tilde{\boldsymbol{h}}(\boldsymbol{s}) \right) \\
&= \hat{\tilde{\boldsymbol{r}}} \mathbf{C} \hat{\tilde{\mathbf{V}}} \tilde{\rho} \mathbf{C} \hat{\tilde{\mathbf{V}}} \hat{\tilde{\boldsymbol{r}}} \\
\mathrm{Cov}(\tilde{\boldsymbol{n}}) &= \mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) \\
&= \sigma^2 \sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]} W \sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{h}}}(\boldsymbol{s})^{2\alpha}]} + \hat{\tilde{\boldsymbol{r}}} \mathbf{C} \hat{\tilde{\mathbf{V}}} \tilde{\rho} \mathbf{C} \hat{\tilde{\mathbf{V}}} \hat{\tilde{\boldsymbol{r}}} \\
&= \sigma^2 \hat{\tilde{\boldsymbol{r}}}^\alpha \frac{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}}{\hat{\tilde{\boldsymbol{\omega}}}^\alpha} W \frac{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}}{\hat{\tilde{\boldsymbol{\omega}}}^\alpha} \hat{\tilde{\boldsymbol{r}}}^\alpha + \hat{\tilde{\boldsymbol{r}}} \mathbf{C} \hat{\tilde{\mathbf{V}}} \tilde{\rho} \mathbf{C} \hat{\tilde{\mathbf{V}}} \hat{\tilde{\boldsymbol{r}}}
\end{aligned}$$

We now make use of the α-coefficient of variation,

$$\tilde{\mathrm{CV}}_a(\alpha) := \tilde{\mathrm{CV}}_a \times \frac{\tilde{\omega}_a^\alpha}{\sqrt{\mathbb{E}[\tilde{\Omega}_a(\boldsymbol{s})^{2\alpha}]}}.$$

Substituting this in, we obtain

$$\begin{aligned}\mathrm{Cov}(\tilde{\boldsymbol{n}}) &= \sigma^2 \hat{\tilde{\boldsymbol{r}}}^\alpha \frac{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}}{\hat{\tilde{\boldsymbol{\omega}}}^\alpha} W \frac{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}}{\hat{\tilde{\boldsymbol{\omega}}}^\alpha} \hat{\tilde{\boldsymbol{r}}}^\alpha + \hat{\tilde{\boldsymbol{r}}} \hat{\tilde{\mathbf{CV}}} \tilde{\rho} \hat{\tilde{\mathbf{CV}}} \hat{\tilde{\boldsymbol{r}}} \\ &= \hat{\tilde{\boldsymbol{r}}} \frac{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}}{\hat{\tilde{\boldsymbol{\omega}}}^\alpha} [\sigma^2 \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} + \hat{\tilde{\mathbf{CV}}}(\alpha) \tilde{\rho} \hat{\tilde{\mathbf{CV}}}(\alpha)] \frac{\sqrt{\mathbb{E}[\hat{\tilde{\boldsymbol{\Omega}}}(\boldsymbol{s})^{2\alpha}]}}{\hat{\tilde{\boldsymbol{\omega}}}^\alpha} \hat{\tilde{\boldsymbol{r}}}\end{aligned}$$

Therefore, up to an additive constant, we obtain the expression

$$\begin{aligned}\ln\left(\det\left(\mathrm{Cov}(\tilde{\boldsymbol{n}})\right)\right) &= 2\sum_{a=1}^{K} \ln(\tilde{r}_a) + \ln\left(\det\left(\sigma^2 \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} + \hat{\tilde{\mathbf{CV}}}(\alpha) \tilde{\rho} \hat{\tilde{\mathbf{CV}}}(\alpha)\right)\right) \\ &= 2\sum_{a=1}^{K} \ln(\tilde{r}_a) + \ln(\det(\hat{\tilde{\mathbf{CV}}}(\alpha) \tilde{\rho} \hat{\tilde{\mathbf{CV}}}(\alpha))) \\ &\quad + \ln\left(\det\left(\sigma^2 \hat{\tilde{\mathbf{CV}}}(\alpha)^{-1} \tilde{\rho}^{-1} \hat{\tilde{\mathbf{CV}}}(\alpha)^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} + I_K\right)\right) \\ &= 2\sum_{a=1}^{K} \ln(\tilde{r}_a) + \ln\left(\det\left(I_K + MW^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)\right) + \text{ const.}\end{aligned}$$

where we have used the substitution

$$M = \sigma^2 \hat{\tilde{\mathbf{CV}}}(\alpha)^{-1} \tilde{\rho}^{-1} \hat{\tilde{\mathbf{CV}}}(\alpha)^{-1} W$$

Putting these results together gives (up to an additive constant)

$$\begin{aligned}\tilde{\mathcal{L}} &= \mu\beta \sum_{a=1}^{K} \ln(\tilde{r}_a) - \sum_{a=1}^{K} \tilde{r}_a \\ &\quad + \mu \ln\left(\det\left(I_K + MW^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)\right)\end{aligned}$$

Taking the derivative of this with respect to $\tilde{r}_a$ gives the expression

$$\begin{aligned}0 &= \frac{\beta\mu}{\tilde{r}_a} - 1 \\ &\quad + \mu \mathrm{tr}\left(\left(I_K + MW^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)^{-1} \frac{\partial}{\partial \tilde{r}_a}\left(MW^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)\right)\end{aligned}$$

Notice that the trace is order M. Truncating to zeroth-order in M, we have $\tilde{r}_a = \beta\mu$. We will also find a first-order approximation. Write $\tilde{r}_a = \beta\mu(1 - \gamma_a)$ where γ_a is order M. Then to first order in M, we have

$$0 = \gamma_a + \mu \mathrm{tr}\left(\left(I_K + MW^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)^{-1} \frac{\partial}{\partial \tilde{r}_a}\left(MW^{-1} \hat{\tilde{\boldsymbol{r}}}^{\alpha-1} W \hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)\right)$$

We take the zeroth-order Neumann expansion $\left(I_K + MW^{-1}\hat{\tilde{\boldsymbol{r}}}^{\alpha-1}W\hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)^{-1} \approx I_K$, which gives

$$\begin{aligned}
-\gamma_a &= \mu \mathrm{tr}\left(\frac{\partial}{\partial \tilde{r}_a}\left(MW^{-1}\hat{\tilde{\boldsymbol{r}}}^{\alpha-1}W\hat{\tilde{\boldsymbol{r}}}^{\alpha-1}\right)\right) \\
&= \mu(\alpha-1)\tilde{r}_a^{\alpha-2}\left[W\hat{\tilde{\boldsymbol{r}}}^{\alpha-1}MW^{-1} + MW^{-1}\hat{\tilde{\boldsymbol{r}}}^{\alpha-1}W\right]_{aa} \\
&\approx \mu(\alpha-1)[2(1-\alpha)\mu]^{\alpha-2}\left[W[2(1-\alpha)\mu]^{\alpha-1}MW^{-1} + MW^{-1}[2(1-\alpha)\mu]^{\alpha-1}W\right]_{aa} \\
&= -\frac{1}{2}[\beta\mu]^{-\beta}\left[M^T + M\right]_{aa} \\
&= -[\beta\mu]^{-\beta}M_{aa} \\
\gamma_a &= [\beta\mu]^{-\beta}M_{aa}
\end{aligned}$$

This gives the following first order expression

$$\tilde{g}_a^{(1)} = \frac{\beta\mu}{\tilde{\omega}_a}\left(1 - \frac{\sigma^2}{(\beta\mu)^\beta}\sum_{b=1}^{K}\frac{[\tilde{\rho}^{-1}]_{ab}W_{ba}}{\tilde{\mathrm{CV}}_a(\alpha)\tilde{\mathrm{CV}}_b(\alpha)}\right).$$

and corresponding zeroth-order expansion

$$\tilde{g}_a^{(0)} = \frac{\beta\mu}{\tilde{\omega}_a}$$

A.9 Optimal homeostatic gains

We now consider the case where we enforce homeostasis on the gains. Our objective, considered as a function of $\tilde{r}_a = \tilde{g}_a\tilde{\omega}_a$ is (up to additive constants)

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

Enforcing homeostasis at the cluster level means setting $\tilde{r}_a = \chi$ for all clusters. Substituting this in, we obtain the function

$$\begin{aligned}
\tilde{\mathcal{L}}(\hat{\boldsymbol{r}} = \chi I_K) &= -K\chi + \mu\log\det\Big(I_K + \chi^{2(1-\alpha)}Q\Big) && (69) \\
&= -K\chi + \mu\sum_n \log\left(1 + \chi^\beta\lambda_n\right) && (70)
\end{aligned}$$

where the matrix Q is defined by

$$Q = \sigma^{-2}W^{-1}\hat{\tilde{\mathbf{CV}}}(\alpha)\tilde{\rho}\hat{\tilde{\mathbf{CV}}}(\alpha),$$

with λ_n its eigenvalues. The optimal χ, within this family of approximate solutions, obeys

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

$$\chi = \mu\beta\left(1 - \frac{1}{K}\sum_{n}\frac{1}{1+\chi^{\beta}\lambda_n}\right) \tag{71}$$

We now consider the special cases discussed in Sec. 2.7 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. 2.9.

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

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

The spectrum of Q is therefore $\lambda_n = b/n$ where

$$b = \sigma^{-2}A_1\mathrm{CV}(\alpha)^2 \tag{72}$$

In this case, (defining $u = n/K$) we can make the following approximation to the right hand side of Eq. (71):

$$\begin{aligned}
1 - \frac{1}{K}\sum_{n=1}^{K}\frac{1}{1+\chi^{\beta}\lambda_n} &= \frac{1}{K}\sum_{n=1}^{K}\frac{\chi^{\beta}\lambda_n}{1+\chi^{\beta}\lambda_n}\\
&= \frac{1}{K}\sum_{n=1}^{K}\frac{\frac{b\chi^{\beta}}{n}}{1+\frac{b\chi^{\beta}}{n}}\\
&\approx \int_{1/K}^{1}\frac{\frac{b\chi^{\beta}}{Ku}}{1+\frac{b\chi^{\beta}}{Ku}}du\\
&\approx \frac{b}{K}\chi^{\beta}\int_{0}^{1}\frac{1}{u+\frac{b}{K}\chi^{\beta}}du\\
&= \frac{b}{K}\chi^{\beta}\ln\left(\frac{1+\frac{b}{K}\chi^{\beta}}{\frac{b}{K}\chi^{\beta}}\right)\\
&= \frac{b}{K}\chi^{\beta}\ln\left(1+\frac{K}{b\chi^{\beta}}\right)
\end{aligned}$$

This gives the new equation

$$\chi = \beta\mu\frac{b}{K}\chi^{\beta}\ln\left(1+\frac{K}{b\chi^{\beta}}\right) \tag{73}$$

Approximating $A_1 \approx K/\ln(K)$ (see App. A.5) and using Eq. (72), we obtain that

$$\frac{K}{b} \approx \frac{\sigma^2\ln(K)}{\tilde{\mathrm{CV}}(\alpha)^2} \tag{74}$$

Substituting Eq. (74) into Eq. (73) gives us

$$\frac{\chi}{\beta\mu}\left(\frac{\sigma^2\ln(K)}{\chi^{\beta}\tilde{\mathrm{CV}}(\alpha)^2}\right) = \ln\left(1+\frac{\sigma^2\ln(K)}{\chi^{\beta}\tilde{\mathrm{CV}}(\alpha)^2}\right) \tag{75}$$

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

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

The matrix Q therefore has eigenspectrum $bn^{\gamma-1}$ where $b = \sigma^{-2} A_1 \tilde{\mathrm{CV}}^2 / A_\gamma$. Inserting this into Eq. (71) gives us

$$\chi = \mu\beta \left(1 - \frac{1}{K} \sum_n \frac{1}{1 + b\chi^\beta n^{\gamma-1}} \right) \tag{76}$$

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

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

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. (75), and using $\beta = 1$, we obtain

$$\frac{\sigma^2 (1-p) \ln(K)}{\mu \tilde{\mathrm{CV}}^2} = \ln \left(1 + \frac{\sigma^2 (1-p) \ln(K)}{\chi \tilde{\mathrm{CV}}^2} \right) \tag{77}$$

We define

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

and rearrange to get

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

A.10 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

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

We recall that the synaptic weights W_{mj} are given by the formula

$$W_{mj} = \frac{w_{mj} \omega_j}{\sum_{i=1}^{N} w_{mi} \omega_i}$$

where w_{mi} are the coefficients of the expansion $\Gamma_m(\boldsymbol{s}) = \sum_i w_{mi} \Omega_i(\boldsymbol{s})$.

We start by calculating w_{mi}.

$$\begin{aligned}
\Gamma_m(\boldsymbol{s}) &= \Pi\left(\boldsymbol{z}^{(2)} = \boldsymbol{z}_m^{(2)}|\boldsymbol{s}\right) \\
&= \int g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}^{(1)}\right)\Pi\left(\boldsymbol{z}^{(1)}|\boldsymbol{s}\right)d\boldsymbol{z}^{(1)} \\
&\approx \sum_{j=1}^{N} g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_j^{(1)}\right)\Pi\left(\boldsymbol{z}^{(1)} = \boldsymbol{z}_j^{(1)}|\boldsymbol{s}\right)\delta\boldsymbol{z}_j^{(1)} \\
&= \sum_{j=1}^{N} g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_j^{(1)}\right)\delta\boldsymbol{z}_j^{(1)}\Omega_j(\boldsymbol{s}) \\
w_{mj} &= g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_j^{(1)}\right)\delta\boldsymbol{z}_j^{(1)}
\end{aligned}$$

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

$$\begin{aligned}
\sum_{j=1}^{N} w_{mj}\omega_j &= \sum_{j=1}^{N} g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_j^{(1)}\right)\delta\boldsymbol{z}_j^{(1)}\pi\left(\boldsymbol{z}_j^{(1)}\right) \\
&= \int g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_j^{(1)}\right)\pi\left(\boldsymbol{z}_j^{(1)}\right)d\boldsymbol{z}^{(1)} \\
&= \pi\left(\boldsymbol{z}_m^{(2)}\right) \\
W_{mi} &= \frac{w_{mi}\omega_i}{\sum_{j=1}^{N} w_{mj}\omega_j} \\
&= \frac{g\left(\boldsymbol{z}_m^{(2)}|\boldsymbol{z}_i^{(1)}\right)\delta\boldsymbol{z}_i^{(1)}\pi\left(\boldsymbol{z}_i^{(1)}\right)}{\pi\left(\boldsymbol{z}_m^{(2)}\right)} \\
&= g\left(\boldsymbol{z}_i^{(1)}|\boldsymbol{z}_m^{(2)}\right)\delta\boldsymbol{z}_i^{(1)}
\end{aligned}$$

which gives the result as required.

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.

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–None.

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

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.

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

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.

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.

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

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: CB*, 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.

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.

Obermayer, K. and Blasdel, G. G. (1993). Geometry of orientation and ocular dominance columns in monkey striate cortex. *Journal of Neuroscience*, 13(10):4114–4129. Publisher: Society for Neuroscience Section: Articles.

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.

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

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.

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.

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.

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

Yarrow, S., Challis, E., and Seriès, P. (2012). Fisher and Shannon information in finite neural populations. *Neural Computation*, 24(7):1740–1780.