# 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

Across changes in their sensory environment or input statistics, cortical neurons display a homeostasis of firing rates [2, 10, 8, 7]. We present a normative explanation of such firing-rate homeostasis grounded in the infomax principle [6]. We further demonstrate how homeostatic coding, coupled with Bayesian theories of neural representation [11] can explain stimulus-specific adaptation effects widely observed in the nervous system (*e.g.*, in V1 [2]). This can be achieved by divisive normalisation with adaptive weights [3, 12].

Consider a neuron acting as a feature detector. If the feature becomes more prevalent, the neuron responds more often and has a higher average firing rate. Prolonged exposure to such an environment causes the neuron to *adapt* by responding less vigorously [2, 10, 8]. If such adaptation returns the average firing rate to its value prior to environmental shift, it is termed *firing rate homeostasis*. This paper uses efficient coding theory to give a general normative account of firing rate homeostasis. We show that homeostasis optimises a trade-off between the information conveyed by the neural response and the metabolic cost of the response. We apply this result to Bayesian theories of representation, showing that this can lead to stimulus specific adaptation effects observed in V1 [2].

## 1 Problem statement and framework

We consider a population of $N$ neurons, responding to the (possibly high-dimensional) stimulus $\boldsymbol{s}$, with marginal distribution $Z(\boldsymbol{s})$ that can vary across environments. Assuming rate coding in time bins of a fixed duration, we denote the population spike counts in a coding interval by $\boldsymbol{n} = (n_1, \ldots, n_N)$. Neural tuning curves, $h_i(\boldsymbol{s})$, are factorised into a *representational curve*, $\Omega_i(\boldsymbol{s})$, and a *gain*, $g_i$:

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

Approximating the spike counts as continuous, we adopt a Gaussian noise model with unit Fano factor: conditional on $\boldsymbol{s}$, $n_i$ are independent with $n_i | \boldsymbol{s} \sim \mathcal{N}(h_i(\boldsymbol{s}), h_i(\boldsymbol{s}))$.

We suggest 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 [5, 4]:

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

Here, $I(\boldsymbol{s}, \boldsymbol{n})$ is the mutual information between the stimulus and response, the second term penalises the average population spike count, and $\mu > 0$ controls the information-energy trade-off.

Note that $\Omega_i$ can be complex functions of the stimulus (*e.g.* multimodal, discontinuous), representing *any* general computation. Adjusting gains does not restrict the computations which can be performed

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

by downstream populations, as adjustments in gains can be compensated for by reciprocal adjustments in readout weights. Thus, our framework applies to populations located deep in the sensory processing pathway. This makes our treatment more general than other efficient coding frameworks (*e.g.* [4]), which place tight constraints on the shape and configuration of the tuning curves.

## 2 Information-energy trade-off leads to homeostasis

Analytic maximisation of mutual information is intractable. We make two approximations to derive a closed form expression for the optimal gains (see App. A.1 for details). For the Gaussian noise model, the conditional entropy is given by $2H[\boldsymbol{n}|\boldsymbol{s}] = \sum_{j=1}^{N} \ln(g_i) + \text{const. in } \boldsymbol{g}$. The marginal entropy, $H[\boldsymbol{n}]$, however, is intractable, and we replace it by an upper bound, namely the entropy of a Gaussian with the same covariance:

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

$I(\boldsymbol{s}, \boldsymbol{n}) = H[\boldsymbol{n}] - H[\boldsymbol{n}|\boldsymbol{s}]$, so $2I(\boldsymbol{n}, \boldsymbol{s}) \leq \ln\left(\det\left(I + P \,\mathrm{diag}(\boldsymbol{g} \odot \boldsymbol{\omega})\right)\right) + \text{const. in } \boldsymbol{g}$, where

$$P \equiv \mathrm{diag}(\boldsymbol{\omega})^{-1} \mathrm{Cov}[\boldsymbol{\Omega}(\boldsymbol{s})] \mathrm{diag}(\boldsymbol{\omega})^{-1} = \mathrm{diag}(\mathrm{CV}) \,\rho \,\mathrm{diag}(\mathrm{CV}), \tag{2}$$

$$\boldsymbol{\omega} \equiv \mathbb{E}[\boldsymbol{\Omega}(\boldsymbol{s})], \qquad \rho \equiv \mathrm{Corr}[\boldsymbol{\Omega}(\boldsymbol{s})] = \mathrm{Corr}[\boldsymbol{h}(\boldsymbol{s})]. \tag{3}$$

and $\odot$ denotes element-wise product, $\mathrm{Corr}[\cdot]$ denotes the Pearson correlation matrix, and $\mathrm{CV}_i$ is the coefficient of variation of $\Omega_i(\boldsymbol{s})$. Up to an additive constant, we obtain the upper bound

$$\mathcal{L}^0(\boldsymbol{g}) \leq \mathcal{L}(\boldsymbol{g}) := \mu \ln\left(\det\left(I + P \,\mathrm{diag}(\boldsymbol{g} \odot \boldsymbol{\omega})\right)\right) - \sum_{i=1}^{N} g_i \omega_i, \tag{4}$$

We subsequently optimise $\mathcal{L}$ (see App. A.2 for details). $\frac{d\mathcal{L}}{dg_i} = \frac{\mu}{g_i}(I + (P \,\mathrm{diag}(\boldsymbol{g} \odot \boldsymbol{\omega}))^{-1})_{ii}^{-1} - \omega_i$. Under the condition

$$\mu \,\mathrm{CV}_i^2 \gg [\rho^{-1}]_{ii}, \tag{5}$$

a first-order Neumann expansion in $(P \,\mathrm{diag}(\boldsymbol{g} \odot \boldsymbol{\omega}))^{-1}$ yields $\frac{d\mathcal{L}}{dg_i} \approx \frac{\mu}{g_i}(1 - \frac{[P^{-1}]_{ii}}{\omega_i g_i}) - \omega_i$. Setting this to zero and reusing Eq. (5) yields the following first-order approximation for the optimal gains

$$g_i \approx g_i^{(1)} \equiv \frac{\mu}{\omega_i}\left(1 - \frac{[\rho^{-1}]_{ii}}{\mu \,\mathrm{CV}_i^2}\right). \tag{6}$$

The corresponding zeroth-order solution, $g_i^{(0)} \equiv \mu/\omega_i$, gives exact *homeostasis*, as under these gains

$$\text{average rate of neuron } i = \mathbb{E}[h_i(\boldsymbol{s})] = g_i^{(0)} \omega_i = \mu$$

which is constant, both between neurons and across environments (as specified by $Z(\boldsymbol{s})$). Since $[\rho^{-1}]_{ii} \geq 0$, $g_i^{(1)} \leq g_i^{(0)}$. Averaging the factor in parentheses in Eq. 6 over neurons, we additionally define $\bar{g}_i^{(1)} = \frac{\mu}{\omega_i}\left(1 - \frac{1}{N}\sum_j (\mu \,\mathrm{CV}_j^2)^{-1} [\rho^{-1}]_{jj}\right)$, which yields uniform mean rates across neurons and, as long as $\sum_j (\mu \,\mathrm{CV}_j^2)^{-1}[\rho^{-1}]_{jj}$ is constant across environments, homeostasis of firing rates.

## 3 Validation of results

In our first-order expansion above, we used the condition (5). How likely is this to hold in the cortex?

- We see from the solution that $\mu$ scales the average spike count of the neurons over the rate coding time-interval. Typical cortical firing rates are 10 Hz [1], so assuming a coding interval of 0.1 sec., we obtain $\mu \approx 1$. Taking our model neurons to represent a cluster of $m$ cortical neurons with identical tuning, $\mu$ is further scaled up by $m$.
- $\rho$ is the *signal correlation matrix* of neurons. Stringer *et al.* [9] have found that in V1, the eigenvalues of $\rho$ obey a power law decay of approximately $1/n$. This leads to an average the average of $[\rho^{-1}]_{ii}$ of approximately $\ln(N)/2$. Even for large values of $N = \mathcal{O}(10^8)$ this remains bounded above by 10

- The coefficient of variation CV can be seen as a measure of sparseness of responses. Consider the toy-model $\Omega_i(\boldsymbol{s}) \sim \mathrm{Bern}(p_i)$ (for an $\Omega_i(\boldsymbol{s})$ with two output levels: 0 and 1). In this case, $\mathrm{CV}_i^2 = (1-p_i)/p_i$, which is large for small $p_i$. At $p_i = 0.1$ we obtain $\mathrm{CV}_i^2 = 9$. Values of $p_i$ in this range have been observed in empirical studies [1].

This makes it clear when we should expect homeostasis – when firing rates are high but selective, and signal correlation structure corresponds to a high-dimensional geometry, as *e.g.* observed in [9]. When these conditions are violated, homeostasis deviates from optimally.

We numerically compare the performance of the zeroth-order homeostatic code $g_i^{(0)} = \mu/\omega_i$, the first order correction, $g_i^{(1)}$, and the homeostatic approximation to the latter, $\bar{g}_i^{(1)}$, against gains $g_i^{opt}$ which have been numerically optimised by performing simple gradient ascent on the objective $\mathcal{L}(\boldsymbol{g})$, initialised as $g_i^{(1)}$. An environment is specified by $\rho$, $\omega_i$, and $\mathrm{CV}_i$ (via $Z(\boldsymbol{S})$). We consider a sequence of environments parameterised by $\epsilon \in [0,1]$. In environment $\epsilon$, $\omega_j(\epsilon) = 4 - \epsilon\cos(2\pi j/N)$; this can be thought of as an environmental shift in which some stimuli become up to $25\%$ more or less prevalent. We took $N = 100$, $\mu = 10$, and $\mathrm{CV}_i(\epsilon) = 3$ for all environments and neurons.

$\rho(\epsilon)$ is obtained by normalising a positive-definite covariance matrix $\Sigma(\epsilon)$ which has a $1/n$ eigenspectrum, in line with the findings of [9]. The orthonormal eigen-bases of the end points $\Sigma(0)$ and $\Sigma(1)$ are sampled randomly and independently, and we smoothly interpolate between these to obtain the eigen-bases for $\Sigma(\epsilon)$, $\epsilon \in (0,1)$. (see App. A.3 for details).

To compare the performance of each approximate solution, $\boldsymbol{g}^{app}(\epsilon)$, we use the measure

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

which can be interpreted as the improvement in $\mathcal{L}(\cdot;\epsilon)$ achieved by the adaptive $\boldsymbol{g}^{app}(\epsilon)$ over the unadapted optimal gains in the original environment $\boldsymbol{g}^{opt}(0)$, relative to the same improvement obtained by the optimally adaptive gains. $C(\epsilon)$ for different approximations are plotted in Fig. 1.

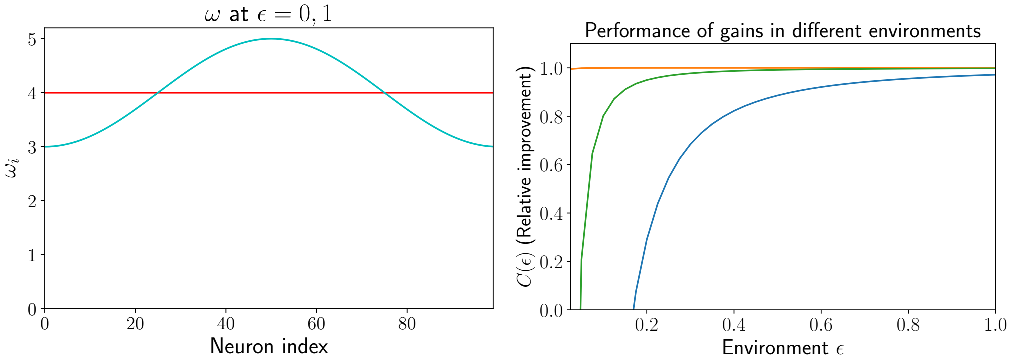

Figure 1: Left: Profiles of rates unmodulated by gains, $\boldsymbol{\omega}$, for the $\epsilon = 0$ (red) and $\epsilon = 1$ (cyan) environments. Right: $C(\epsilon)$ as a function of $\epsilon$ (averaged across 50 realisations of $\rho(\epsilon)$) for the three approximations $\boldsymbol{g}^{(0)}(\epsilon)$ (blue), $\boldsymbol{g}^{(1)}(\epsilon)$ (orange), and $\bar{\boldsymbol{g}}^{(1)}(\epsilon)$ (green). Here, $N = 100$, $\mu = 10$, and $\mathrm{CV}_i(\epsilon) = 3$ (for all neurons and environments).

We first note that the first order approximation (6) performs essentially as well as numerical optimisation of $\mathcal{L}$. The homeostatic correction $\bar{\boldsymbol{g}}^{(1)}(\epsilon)$ also performs close to optimally for $\epsilon \geq 0.1$. Moreover, as soon as $\epsilon$ exceeds about 0.2, exact homeostasis $\boldsymbol{g}^{(0)}$ is superior to no adaptation. Finally, we also computed the mean relative errors $\frac{1}{N}\sum_j |g_j^{app} - g_j^{opt}|/g_j^{opt}$ of the approximate solutions. These were approximately constant in $\epsilon$ to 3 significant figures, and took the values 0.0294, 0.000291, 0.00615, for $\boldsymbol{g}^{(0)}$, $\boldsymbol{g}^{(1)}$, and $\bar{\boldsymbol{g}}^{(1)}$, respectively.

# 4 Bayes-ratio coding

In this section, we neglect correction terms and work with the homeostatic approximation $g_i^{(0)} = \mu/\omega_i$, which yields $h_i(\boldsymbol{s}) = \mu\Omega_i(\boldsymbol{s})/\mathbb{E}[\Omega_i(\boldsymbol{S})]$. So far we have been silent on the computation encoded by the representational curves $\boldsymbol{\Omega}(\boldsymbol{s})$. Here, we apply the homeostatic code to a form of Bayesian encoding, the distributed distributional code (DDC) [11]. In a DDC, which is based on an internal generative model of stimuli with latent variables $\boldsymbol{z}$, $\Omega_i(\boldsymbol{s})$ is the posterior expectation (given the observed $\boldsymbol{s}$) of a so-called kernel function $k_i(\boldsymbol{z})$. We make an ideal-observer assumption, wherein the internal generative model generates the true stimulus distribution, $Z(\boldsymbol{s})$. In this case, the homeostatic adaptation yields $h_i(\boldsymbol{s}) = \mu\frac{\mathbb{E}[k_i(\boldsymbol{z})|\boldsymbol{s}]}{\mathbb{E}[k_i(\boldsymbol{z})]}$. For the special case $k_i(\boldsymbol{z}) = \delta(\boldsymbol{z}^i - \boldsymbol{z})$, this yields *Bayes-ratio coding*:

$$h_i(\boldsymbol{s}) = \mu\frac{\Pi(\boldsymbol{z}^i|\boldsymbol{s})}{\pi(\boldsymbol{z}^i)} = \mu\frac{f(\boldsymbol{s}|\boldsymbol{z}^i)}{Z(\boldsymbol{s})}, \tag{7}$$

where $\Pi$ and $\pi$ denote the posterior and prior distributions over $\boldsymbol{z}$, and $f(\boldsymbol{s}|\boldsymbol{z})$ is the generative model's likelihood. As shown in App. A.4, Bayes-ratio coding can be achieved via divisive normalisation, a canonical cortical operation [3], with adaptive normalisation weights proportional to the prior $\pi(\boldsymbol{z}^i)$ (allowed to vary across environments) and feedforward inputs given by the likelihood function.

Additionally, Bayes-ratio coding can be used to explain certain stimulus-specific adaptation effects. It is typical of stimulus specific adaptation in V1 [2] that orientation tuning curves display a repulsion and suppression around the over-represented orientation of the adaptor stimulus. To model the findings of [2] with Bayes-ratio coding, we took stimulus and latent variable spaces to be the orientation space $[-90, 90)$, and the likelihood $f(s|z^i) = \psi(s - z^i; 10)$ where $\psi(\cdot; k)$ a von Mises density with precision $k$, where $z^i$ are uniformly spaced over $[-90, 90)$. We consider a uniform pre-adaptation prior $\pi^{pre}$, and a post-adaptation prior $\pi^{post}(z) = \frac{0.6}{180} + 0.4\psi(z; 2)$ representing an adaptor stimulus at $z = 0°$. The pre- and post-adaptation tuning curves are plotted in figure 2, and exhibit the suppression and repulsion of the adapted tuning curves around the over-represented stimulus. This result generalises to a homeostatic DDC with finite-width unimodal kernel functions.

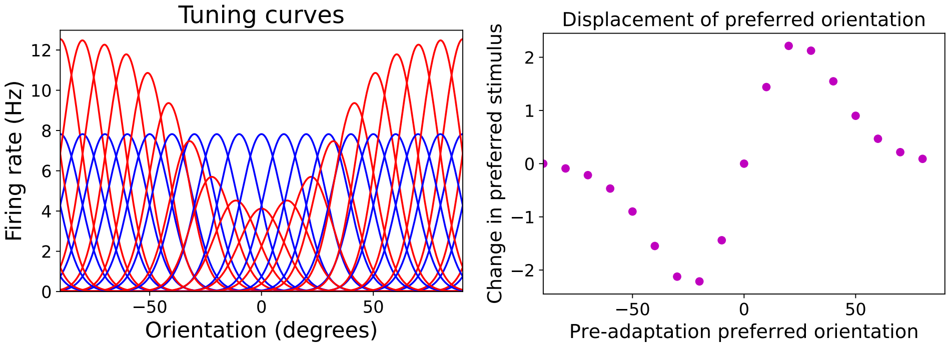

Figure 2: Left: The tuning curves pre- (blue) and post-adaptation (red). Right: The displacement of the preferred orientation of neurons (in degrees) as a function of the pre-adaptation preferred orientation, demonstrating the repulsion of tuning curves from the adaptor orientation.

# 5 Conclusion

We developed a theory of optimal gain modulation for combatting noise in neural representations. We demonstrated that, when mean neural firing rates are not too small and responses are sufficiently sparse with 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. Lastly, we demonstrated that, when applied to Bayesian DDCs as an example of neural representation, homeostasis leads to characteristic suppression and repulsion of tuning curves as observed in V1. In future research, we will extend our analysis to the more realistic case of Poisson spiking noise (*i.e.* $n_i|\boldsymbol{S} \sim \mathrm{Poisson}(h_i(\boldsymbol{S}))$), and investigate the extent to which homeostasis deviates from optimally as the condition $\mu\mathrm{CV}_i \gg [\rho^{-1}]_{ii}$ ceases to hold.

## References

[1] David Attwell and Simon B. Laughlin. “An Energy Budget for Signaling in the Grey Matter of the Brain”. In: *Journal of Cerebral Blood Flow & Metabolism* 21.10 (Oct. 2001). Publisher: SAGE Publications Ltd STM, pp. 1133–1145. ISSN: 0271-678X. DOI: 10.1097/00004647-200110000-00001. URL: https://doi.org/10.1097/00004647-200110000-00001 (visited on 07/30/2022).

[2] Andrea Benucci, Aman B. Saleem, and Matteo Carandini. “Adaptation maintains population homeostasis in primary visual cortex”. eng. In: *Nature Neuroscience* 16.6 (June 2013), pp. 724–729. ISSN: 1546-1726. DOI: 10.1038/nn.3382.

[3] Matteo Carandini and David J. Heeger. “Normalization as a canonical neural computation”. en. In: *Nature Reviews Neuroscience* 13.1 (Jan. 2012). Number: 1 Publisher: Nature Publishing Group, pp. 51–62. ISSN: 1471-0048. DOI: 10.1038/nrn3136. URL: https://www.nature.com/articles/nrn3136 (visited on 08/09/2022).

[4] Deep Ganguli and Eero P. Simoncelli. “Efficient Sensory Encoding and Bayesian Inference with Heterogeneous Neural Populations”. en. In: *Neural Computation* 26.10 (Oct. 2014), pp. 2103–2134. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/NECO_a_00638. URL: https://direct.mit.edu/neco/article/26/10/2103-2134/8022 (visited on 08/21/2022).

[5] William Levy and Robert Baxter. “Energy Efficient Neural Codes”. In: *Neural computation* 8 (May 1996), pp. 531–43. DOI: 10.1162/neco.1996.8.3.531.

[6] R. Linsker. “Self-organization in a perceptual network”. In: *Computer* 21.3 (Mar. 1988). Conference Name: Computer, pp. 105–117. ISSN: 1558-0814. DOI: 10.1109/2.36.

[7] Timothy O’Leary et al. “Cell Types, Network Homeostasis, and Pathological Compensation from a Biologically Plausible Ion Channel Expression Model”. In: *Neuron* 82.4 (May 2014), pp. 809–821. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2014.04.002. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4109293/ (visited on 08/21/2022).

[8] Zahid Padamsey et al. “Neocortex saves energy by reducing coding precision during food scarcity”. In: *Neuron* 110 (Nov. 2021). DOI: 10.1016/j.neuron.2021.10.024.

[9] Carsen Stringer et al. “High-dimensional geometry of population responses in visual cortex”. en. In: *Nature* 571.7765 (July 2019). Number: 7765 Publisher: Nature Publishing Group, pp. 361–365. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1346-5. URL: https://www.nature.com/articles/s41586-019-1346-5 (visited on 07/07/2022).

[10] Gina G. Turrigiano and Sacha B. Nelson. “Homeostatic plasticity in the developing nervous system”. en. In: *Nature Reviews Neuroscience* 5.2 (Feb. 2004). Number: 2 Publisher: Nature Publishing Group, pp. 97–107. ISSN: 1471-0048. DOI: 10.1038/nrn1327. URL: http://www.nature.com/articles/nrn1327 (visited on 08/21/2022).

[11] Eszter Vertes and Maneesh Sahani. *Flexible and accurate inference and learning for deep generative models*. arXiv:1805.11051 [cs, stat]. May 2018. DOI: 10.48550/arXiv.1805.11051. URL: http://arxiv.org/abs/1805.11051 (visited on 08/20/2022).

[12] Zachary M. Westrick, David J. Heeger, and Michael S. Landy. “Pattern Adaptation and Normalization Reweighting”. en. In: *Journal of Neuroscience* 36.38 (Sept. 2016). Publisher: Society for Neuroscience Section: Articles, pp. 9805–9816. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.1067-16.2016. URL: https://www.jneurosci.org/content/36/38/9805 (visited on 08/20/2022).

# A Appendix

## A.1 Upper bound on $\mathcal{L}^0$

Recall that $n_i|\boldsymbol{s} \sim N(h_i(\boldsymbol{s}), h_i(\boldsymbol{s}))$. In this case,

$$\begin{aligned}
H[n_i|\boldsymbol{s}] &= \int H[N(h_i(\boldsymbol{s}), h_i(\boldsymbol{s}))]Z(\boldsymbol{s})d\boldsymbol{s} \\
&= \int \frac{1}{2}\ln(2\pi e h_i(\boldsymbol{s}))Z(\boldsymbol{s})d\boldsymbol{s} \\
&= \frac{1}{2}\ln(2\pi e) + \frac{1}{2}\ln(g_i) + \frac{1}{2}\mathbb{E}[\ln(\Omega_i(\boldsymbol{s}))]
\end{aligned}$$

$$\begin{aligned}
H[\boldsymbol{n}|\boldsymbol{s}] &= \sum_{j=1}^{N} H[n_j|\boldsymbol{s}] && (8) \\
&= \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\sum_{j=1}^{N}\ln(g_i) + \sum_{j=1}^{N}\frac{1}{2}\mathbb{E}\left[\ln(\Omega_j(\boldsymbol{s}))\right] && (9)
\end{aligned}$$

We now use the fact that the entropy of any continuous distribution can be upper bounded by the entropy of a Gaussian with with the same covariance. The entropy of a $N(\boldsymbol{\mu}, \Sigma)$ random variable is $H[N(\boldsymbol{\mu}, \Sigma)] = \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\ln(\det(\Sigma))$. It therefore suffices to find the covariance of $\boldsymbol{n}$. We use the decomposition $\mathrm{Cov}(\boldsymbol{n}) = \mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})] + \mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}])$ and compute

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s}) &= \mathrm{diag}(\mathrm{Var}(\boldsymbol{n}|\boldsymbol{s})) \\
&= \mathrm{diag}(\boldsymbol{h}(\boldsymbol{s})) \\
\mathbb{E}[\mathrm{Cov}(\boldsymbol{n}|\boldsymbol{s})] &= \mathrm{diag}(\mathbb{E}[\boldsymbol{h}(\boldsymbol{s})]) \\
&= \mathrm{diag}(\boldsymbol{g} \odot \boldsymbol{\omega}) \\
&= \mathrm{diag}(\boldsymbol{g})\mathrm{diag}(\boldsymbol{\omega}) \\
\mathrm{Cov}(\mathbb{E}[\boldsymbol{n}|\boldsymbol{s}]) &= \mathrm{Cov}(\boldsymbol{h}(\boldsymbol{s})) \\
&= \mathrm{diag}(\boldsymbol{g})\mathrm{Cov}(\boldsymbol{\Omega}(\boldsymbol{s}))\mathrm{diag}(\boldsymbol{g})
\end{aligned}$$

Thus

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{n}) &= \mathrm{diag}(\boldsymbol{g})\mathrm{diag}(\boldsymbol{\omega}) + \mathrm{diag}(\boldsymbol{g})\mathrm{Cov}(\boldsymbol{\Omega}(\boldsymbol{s}))\mathrm{diag}(\boldsymbol{g}) \\
&= \mathrm{diag}(\boldsymbol{g})\mathrm{diag}(\boldsymbol{\omega})[I + \mathrm{diag}(\boldsymbol{\omega})^{-1}\mathrm{Cov}(\Omega(s))\mathrm{diag}(\boldsymbol{\omega})^{-1}\mathrm{diag}(\boldsymbol{\omega})\mathrm{diag}(\boldsymbol{g})] \\
&= \mathrm{diag}(\boldsymbol{g})\mathrm{diag}(\boldsymbol{\omega})[I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})] \\
\ln(\det(\mathrm{Cov}(\boldsymbol{n}))) &= \ln(\det(\mathrm{diag}(\boldsymbol{g}))) + \ln(\det(\mathrm{diag}(\boldsymbol{\omega}))) + \ln(\det(I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))) \\
&= \sum_{i=1}^{N}\ln(g_i) + \sum_{i=1}^{N}\ln(\omega_i) + \ln\left(\det\left(I + P\,\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})\right)\right)
\end{aligned}$$

where $P = \mathrm{diag}(\boldsymbol{\omega})^{-1}\mathrm{Cov}(\boldsymbol{\Omega}(\boldsymbol{s}))\mathrm{diag}(\boldsymbol{\omega})^{-1}$. Putting this together, we obtain:

$$H[\boldsymbol{n}] \leq \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\sum_{j=1}^{N}\ln(g_i) + \frac{1}{2}\sum_{i=1}^{N}\ln(\omega_i) + \frac{1}{2}\ln\left(\det\left(I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})\right)\right) \qquad (10)$$

Finally, substituting Eqs. (9) and (10 in the definition for the mutual information, we obtain the upper bound

$$\begin{aligned}
I(\boldsymbol{n}, \boldsymbol{s}) &= H[\boldsymbol{n}] - H[\boldsymbol{n}|\boldsymbol{s}] \\
&\leq \frac{1}{2}\sum_{i=1}^{N}\ln(\omega_i) - \mathbb{E}\left[\ln(\Omega_j(\boldsymbol{s}))\right] + \frac{1}{2}\ln\left(\det\left(I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})\right)\right) \\
&= -\frac{1}{2}\sum_{i=1}^{N}\mathbb{E}\left[\ln\left(\frac{\Omega_i(\boldsymbol{s})}{\omega_i}\right)\right] + \frac{1}{2}\ln\left(\det\left(I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})\right)\right),
\end{aligned}$$

or $2I(\boldsymbol{n}, \boldsymbol{s}) \leq \ln\left(\det\left(I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})\right)\right) +$ const. in $\boldsymbol{g}$.

### A.2 Approximate maximisation of $\mathcal{L}$

We now consider optimising

$$\mathcal{L}(\boldsymbol{g}) = \mu \ln\left(\det\left(I + P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})\right)\right) - \sum_{i=1}^{N} g_i \omega_i$$

Taking the derivative with respect to $g_i$, we obtain

$$\frac{d\mathcal{L}}{dg_i} = \frac{\mu}{g_i}(I + (P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))^{-1})^{-1}_{ii} - \omega_i$$

To obtain an approximate solution for this we use a first order Neumann expansion for $(I + (P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))^{-1})^{-1}$. This requires that $(P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))^{-1}$ has a small norm, i.e. $P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})$ is large in norm. The Neumann expansion gives:

$$(I + (P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))^{-1})^{-1} \approx 1 - \mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g})^{-1}P^{-1}$$

Plugging this in, we obtain

$$\frac{d\mathcal{L}}{dg_i} \approx \frac{\mu}{g_i}(1 - \frac{[P^{-1}]_{ii}}{\omega_i g_i}) - \omega_i \tag{11}$$

Setting equation (11) to zero, and solving to find the approximate maximiser of $\mathcal{L}(\boldsymbol{g})$, we get:

$$g_i = \frac{\mu}{2\omega_i}\left(1 + \sqrt{1 - \frac{4[P^{-1}]_{ii}}{\mu}}\right) \tag{12}$$

We further approximate by taking a first order Taylor expansion of the square root to obtain equation (6):

$$g_i \approx \frac{\mu}{\omega_i}\left(1 - \frac{\rho^{-1}_{ii}}{\mu \mathrm{CV}_i^2}\right)$$

where $\rho$ is the correlation matrix $\mathrm{Corr}(\boldsymbol{\Omega}(\boldsymbol{s}))$ and $\mathrm{CV}_i$ is the coefficient of variation, $\frac{\sqrt{\mathrm{Var}(\Omega_i(\boldsymbol{s}))}}{\omega_i}$

The expansions employed in this argument require that $(P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))^{-1}$ is small in norm, and that $\rho^{-1}_{ii} \ll \mu \mathrm{CV}_i^2$. Note that $(P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))_{ij} = \mathrm{CV}_i \rho_{ij} \mathrm{CV}_j g_j \omega_j$. But we know that $g_j = \mathcal{O}(\mu/\omega_j)$ and therefore $g_j\omega_j = \mathcal{O}(\mu)$. Accordingly, we obtain that $(P\mathrm{diag}(\boldsymbol{\omega} \odot \boldsymbol{g}))^{-1} = \mathcal{O}\left(\frac{\rho^{-1}}{\mu \mathrm{CV}^2}\right)$. This means that the condition for truncating the Neumann expansion is indeed Eq. (5).

### A.3 Specification of environment for $\epsilon$

To generate $\rho(\epsilon)$, we first generate a covariance matrix $\Sigma(\epsilon)$, and let $\rho(\epsilon)$ be the corresponding correlation matrix. The procedure for generating $\Sigma(\epsilon)$ is as follows.

We randomly and independently sample two $N \times N$ random Gaussian matrices $A_0, A_1 \sim \mathcal{N}_{N\times N}(0, I_{N\times N})$ and obtain symmetric matrices $S_0 = A_0 + A_0^T$ and $S_1 = A_1 + A_1^T$. 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 $O(N)$.

Now let $S(\epsilon) = (1-\epsilon)S_0 + \epsilon S_1$. Almost surely these matrices have a non-degenerate spectrum, and hence a unique representation as $S(\epsilon) = U(\epsilon)\Lambda(\epsilon)U(\epsilon)^T$ where $\Lambda(\epsilon)$ is diagonal with strictly decreasing eigenvalues and $U(\epsilon)$ is orthogonal. Moreover, $U(\epsilon)$ depends continuously on $\epsilon$.

Finally, we define $\Sigma(\epsilon) = U(\epsilon)DU(\epsilon)^T$, where $D = \mathrm{diag}(1, 1/2, 1/3, \ldots, 1/N)$.

### A.4 Bayes-ratio coding and divisive normalisation

Given a collection of feed-forward inputs, $F_i(\boldsymbol{s})$, divisive normalisation computes the response (and thus the tuning curve) of neuron $i$ as

$$h_i(\boldsymbol{s}) = \gamma \frac{F_i(\boldsymbol{s})^n}{\sigma^n + \sum_j w_j F_j(\boldsymbol{s})^n} \tag{13}$$

where $w_j$ are a collection of *normalisation weights*, and $\gamma, \sigma \geq 0$ 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 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 $\boldsymbol{z}^i$. Then we have:

$$\begin{aligned}\sum_j w_j F_j(\boldsymbol{s}) &= \sum_j f(\boldsymbol{s}|\boldsymbol{z}^{(j)})\pi(\boldsymbol{z}^{(j)})\delta\boldsymbol{z}^{(j)} \\ &\approx \int f(\boldsymbol{s}|\boldsymbol{z})\pi(\boldsymbol{z})d\boldsymbol{z} \\ &= Z(\boldsymbol{s})\end{aligned}$$

hence

$$\begin{aligned}h_i(\boldsymbol{s}) &= \mu\frac{F_i(\boldsymbol{s})}{\sigma + \sum_{j=1}^{N} w_j F_j(\boldsymbol{s})} \\ &\approx \mu\frac{f(\boldsymbol{s}|\boldsymbol{z}^i)}{\sigma + Z(\boldsymbol{s})}\end{aligned}$$

Taking the limit as $\sigma \to 0$, we obtain Eq. 7:

$$h_i(\boldsymbol{s}) = \mu\frac{f(\boldsymbol{s}|\boldsymbol{z}^i)}{Z(\boldsymbol{s})}.$$

Therefore, provided $\sigma$ is small compared to $Z(\boldsymbol{s})$, 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).