Functional diversity among sensory neurons from efficient coding principles

Julijana Gjorgjieva[1], Markus Meister[2], Haim Sompolinsky[3,4]

1 Max Planck Institute for Brain Research, Frankfurt, Germany
2 Division of Biology, California Institute of Technology, Pasadena, CA, USA
3 Center for Brain Science, Harvard University, Cambridge, MA, USA
4 The Edmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem, Israel

* gjorgjieva@brain.mpg.de

Abstract

In many sensory systems the neural signal is coded by the coordinated response of heterogeneous populations of neurons. What computational benefit does this diversity confer on information processing? We derive an efficient coding framework assuming that neurons have evolved to communicate signals optimally given natural stimulus statistics and metabolic constraints. Incorporating nonlinearities and realistic noise, we study optimal population coding of the same sensory variable using two measures: maximizing the mutual information between stimuli and responses, and minimizing the error incurred by the optimal linear decoder of responses. Our theory is applied to a commonly observed splitting of sensory neurons into ON and OFF that signal stimulus increases or decreases, and to populations of monotonically increasing responses of the same type, ON. Depending on the optimality measure, we make different predictions about how to optimally split a population into ON and OFF, and how to allocate the firing thresholds of individual neurons given realistic stimulus distributions and noise, which accord with certain biases observed experimentally.

Introduction

The efficient coding hypothesis states that sensory systems have evolved to optimally transmit information about the natural world given limitations on their biophysical components and constraints on energy use [1]. This theory has been applied successfully to explain the structure of neuronal receptive fields in the mammalian retina [2,3] and fly lamina [4,5] based on the statistics of natural scenes. Similar arguments have been made to explain why early sensory pathways often split into parallel channels that represent different stimulus variables, for example different auditory waveforms [6], or local visual patterns [7]. Even neurons that encode the same sensory feature often split further into distinct types. One such commonly encountered diversification is into ON and OFF types: ON cells fire when the stimulus increases and OFF cells when it decreases. This basic ON-OFF dichotomy is found in many modalities, including vertebrate vision [8], invertebrate vision [9], thermosensation [10], and chemosensation [11]. Furthermore, among the neurons that encode the same sensory variable with the same sign, one often encounters distinct types that have different response thresholds, for example, among touch receptors [12] and electroreceptors [13]. The same principle seems to apply several synapses downstream from the receptors [14], and even in the organization of the motor periphery, where motor neurons that activate the same muscle have a broad range of response thresholds [15]. In the present article we consider this sensory response diversification among neurons that represent the same variable and explore whether it can be understood based on a nonlinear version of efficient population coding.

One reason why the ON and OFF pathways have evolved may be to optimize information about both increments and decrements in stimulus intensity by providing excitatory signals for both [16]. For instance, if there were only one ON neuron, then such a cell would need high baseline firing rate to encode stimulus decrements, which can be very costly. We, and others have previously addressed the benefits for having ON and OFF cells in a small population of just two cells [17–19]. However, it remains unclear how a population of many neurons could resolve this issue by tuning their thresholds so that they jointly code for the stimulus. Since ON and OFF neurons often exhibit a broad distribution of firing thresholds [12–15], an important question is thus, what distribution of thresholds yields the most efficient coding. Here we study optimal information transmission in sensory populations comprised of different mixtures of ON and OFF neurons, including purely homogeneous populations with neurons of only one type, e.g. ON, that code for a common stimulus variable by diversifying their thresholds.

Traditionally, efficient population coding has either optimized linear features in the presence of noise [2, 3, 20, 21], or nonlinear processing in the limit of no noise or infinitely large populations [22–25]. We simultaneously incorporate neuronal nonlinearities and realistic noise at the spiking output, which have important consequences in finite populations of neurons, as encountered biologically. We develop the problem parametrically in the neuronal noise and the distribution of stimuli that the cells encode, allowing us to make general predictions applicable to different sensory systems.

What quantity might neural populations optimize? We consider two alternative measures of optimal coding that are in common use [22, 26–30]: first we maximize the mutual information between stimulus and response without any assumptions about how this information should be decoded, and second we optimize the estimate of the stimulus obtained by a linear decoder of the response. The two criteria lead to different predictions both on the optimal ON/OFF ratio and the distribution of optimal thresholds. When constraining the maximal firing rate of each cell, we find that counter to our expectations the mutual information is identical for any mixture of ON and OFF cells once the thresholds of all cells are optimized. This result is independent of the shape of the stimulus distribution and the level of neuronal noise. However, the total mean spike count is the lowest for the population with equal numbers of ON and OFF cells, making this arrangement optimal in terms of bits per spike. Optimizing the linear decoder requires determining not only the cells' thresholds, but also the decoding weights in order to minimize the mean square error between the stimulus and its estimate. Under this criterion, the optimal ON/OFF mixture and cells' thresholds depend on the asymmetries in the stimulus distribution and the noise level, and can account for certain biases observed experimentally in different sensory systems. We also make distinct predictions for the optimal distribution of thresholds under the two optimality measures, noise level and stimulus distributions, providing insight into the diverse coding strategies of these populations across different sensory modalities and species where these differences are encountered.

Results

Population coding model

We develop a theoretical framework to derive the coding efficiency and response properties of a population of sensory neurons representing a common stimulus (Fig. 1A). We specifically consider populations with responses of opposite polarity, ON and OFF, which increase or decrease their response as a function of the common sensory variable; thus, our theory applies to any sensory system where ON and OFF pathways have been observed, for example, heat-activated and cold-activated ion channels in thermosensation [31, 32], mechanosensory neurons [33, 34], or retinal ganglion cells which code for the same spatial location and visual feature with different thresholds [18]. As a special case, we consider populations of neurons with a single polarity, which increase their response as a function of the common sensory variable, for example, olfactory receptor neurons that code for the same odor at different concentrations [35–37]. Each model neuron encodes information about a common scalar stimulus through the spike count observed during a short coding window. The duration of this coding window, T, is chosen based on the observed dynamics of neuronal responses, which is typically in the range of 10-50 ms [28, 38]. Neuronal spike counts are stochastic and their mean is modulated by the stimulus through a discrete response function with a finite number of responses. Discretization in neural circuits occurs on many levels [39]; for example, previous experimental studies have found that sensory neurons use discrete firing rate levels to represent continuous stimuli [4, 27, 40]. Furthermore, theoretical work has shown that the optimal neuronal response functions are discrete under different measures of efficiency [27, 41–43].

The best way to discretize a neural signal depends on many factors, including noise, stimulus statistics and biophysical constraints [39]. Under the constraint of short coding windows encountered in many sensory areas, optimizing a single response function results in a discretization with two response levels, i.e. a binary response function [27, 28, 41–43]. Binary response functions also offer a reasonable approximation of neural behavior in several systems [28, 44, 45]. Therefore, we assumed that ON (OFF) neurons fire Poisson spikes

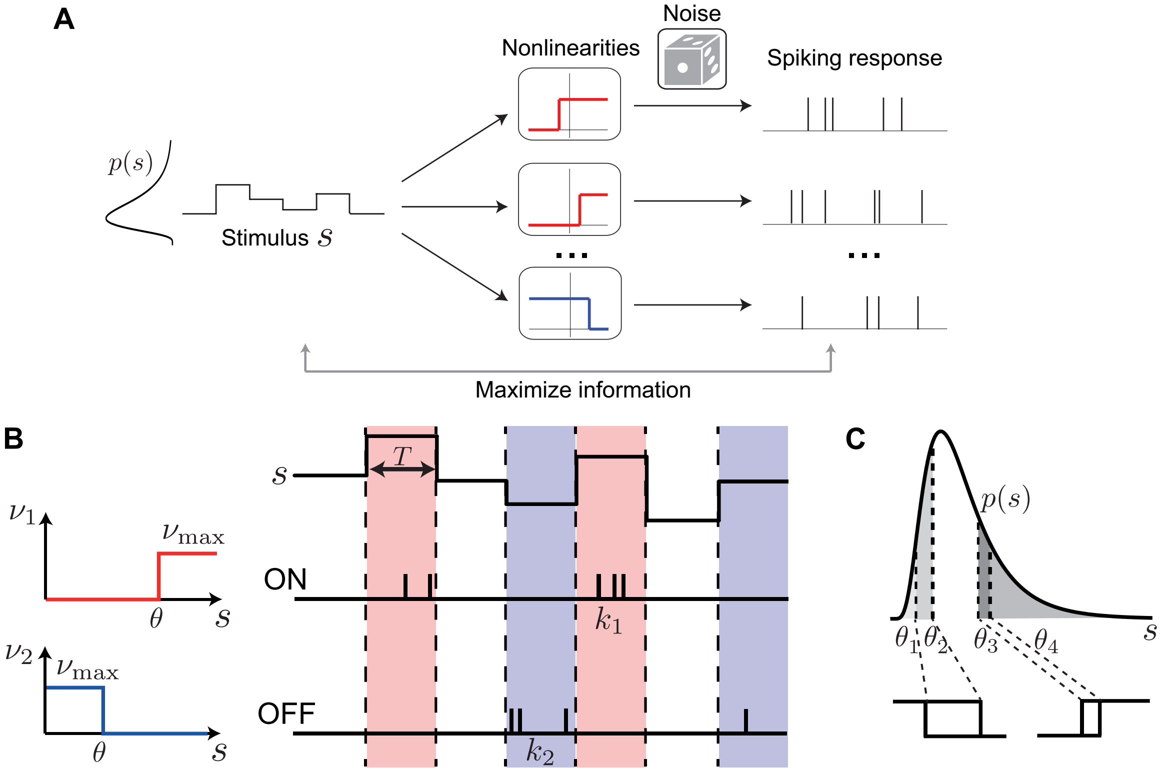

Figure 1. Neuron model and population coding framework. A. Framework schematic. A stimulus s from a probability distribution $p(s)$ is encoded by the spiking responses of a population of ON (red) and OFF (blue) cells. We optimize the cells' nonlinearities by maximizing the mutual information between stimulus and spiking response. **B**. Each cell is described by a binary response nonlinearity ν with a threshold θ and maximal firing rate $\nu_{\max}$. During a coding window of fixed duration T the stimulus is constant and the spike count k is drawn from a Poisson distribution with a mean rate ν. **C**. When measuring coding efficiency using the mutual information between stimulus and spike count response, the neurons' thresholds can be interpreted as quantiles of the original stimulus distribution, thus mapping an arbitrary stimulus distribution $p(s)$ into a uniform distribution (four thresholds shown).

with an average mean count $\nu_{\max}$ whenever the stimulus intensity is above (below) their threshold θ_i, and zero otherwise, i.e. $\nu_i(s) = \nu_{\max}\Theta(s - \theta_i)$ for ON neurons and $\nu_i(s) = \nu_{\max}\Theta(\theta_i - s)$ for OFF neurons (Fig. 1B), where Θ is the Heaviside function.

Maximal mutual information for mixtures of ON and OFF neurons

What should be the number of ON vs. OFF cells and the distribution of their firing thresholds in a population of neurons that optimally represent a given stimulus? To answer these questions, we first maximize the Shannon mutual information between stimulus and population response, in search of a simple efficient coding principle that could explain ON-OFF splitting and, more generally, threshold diversification. We perform the optimization while constraining the expected spike count $R = \nu_{\max}T$ for each cell. Biophysically, such a constraint on the maximal firing rate arises naturally from refractoriness of the spike-generating membrane. We have analytically proven the following theorem (see Methods):

Equal Coding Theorem: For a population of any number N of ON and OFF Poisson neurons coding a one-dimensional stimulus in a fixed time window T by binary rate functions with maximal firing rate $\nu_{\max}$, the mutual information is identical for all ON/OFF mixtures when the thresholds are optimized, for all N, $\nu_{\max}$ and stimulus distributions.

Specifically, the maximal information is achieved in the case when the ON and OFF cells do not overlap, so that all ON thresholds are bigger than all OFF thresholds. For example, consider a mixed population of ON and OFF cells. To calculate the information conveyed by this entire population, we imagine first observing only the ON cells, and in a second step the remaining OFF cells. If one of the ON cells fired a spike, we know the stimulus is in that cell's response range, and therefore we do not learn additional information from observing the OFF cells. If none of the ON cells fired, we gain additional information from observing the OFF cells. One can make the same argument if the remaining cells are all ON cells, or indeed any other mixture. Careful consideration shows that the maximal information gained from that remaining cell population is the same whether they are ON cells or OFF cells (see Methods and S1 Text). Hence, the homogeneous and any mixed ON-OFF population achieve the same maximal information.

The value of this maximal information depends on the expected spike count, $R = \nu_{\max} T$. We introduce the parameter $q = e^{-R}$, which ranges from $q = 0$ in the noiseless limit of high firing rate, and $q = 1$ in the high noise limit of low firing rate. We show that for any ON-OFF mixture (including the homogeneous with only ON cells), the maximal information achieved with optimized thresholds is (Fig. 2A, see Methods)

$$I = \log\left(1 + N(1-q)q^{q/(1-q)}\right). \tag{1}$$

We further ensured that the conclusion of equal coding holds in a population of two cells independent of the Poisson noise model we assumed, which has zero noise when the the firing rate of a neuron is zero. Specifically, we investigated information transmission introducing a spontaneous firing rate under the same Poisson model, as well as empirically measured sub-Poisson noise from salamander retinal ganglion cells [17, 28] (see S1 Text).

The above equation 1 allows us to exactly compute the maximal information that would be reached by a population of neurons as a function of the number of neurons N and the level of noise q assuming optimality, without resorting to expensive numerical calculations [46]. Even if real biological systems do not perform optimally, this quantity could be used as an upper bound for the largest possible information that the system could transmit under the appropriate constraints. In the noiseless limit, $R \to \infty$ (i.e. $q = 0$), where the neurons are deterministic, I reaches its upper bound $I = \log(N + 1)$. The effect of noise is most prominent when R is of order $1/N$, so that the total spike count RN is of order 1, implying that the signal-to-noise of the entire population is of order 1. We call this the **high noise regime**, and here we obtain $I \to \log(RN/e + 1)$, where e denotes $\exp(1)$.

Optimal distribution of thresholds

We next asked what distribution of thresholds in the population of ON and OFF cells achieves this maximal mutual information. In the case of a discrete rate function, we can replace θ_i by the corresponding cumulative threshold (fraction of stimuli below threshold), which essentially maps the stimulus distribution into a uniform distribution from 0 to 1 (Fig. 1C). Since the stimulus dependence enters only through these values, the maximal mutual information is independent of the stimulus distribution $p(s)$, provided that the stimulus cumulative distribution is continuous. Instead, the information depends on the areas of $p(s)$ between consecutive thresholds. It is therefore useful to define the optimal threshold intervals $p_i = \int_{\theta_i}^{\theta_{i+1}} \mathrm{d}s\, p(s)$ where the neurons' thresholds are ordered $\theta_1 \leq \ldots \leq \theta_N$ (and we define the special $\theta_0 = -\infty$ and $\theta_{N+1} = \infty$). We find a surprisingly simple structure for the optimal p_i (Fig. 2B,C). The optimal thresholds divide stimulus space into intervals of equal area, which depend on the noise level, q,

$$p_i = p = \int_{\theta_i}^{\theta_{i+1}} \mathrm{d}s\, p(s) = \frac{1-q}{q^{-q/(1-q)} + N(1-q)} \tag{2}$$

for all i, *except* for the two *'edge'* intervals,

$$p_{\text{edge}} = \int_{-\infty}^{\theta_1} \mathrm{d}s\, p(s) = \int_{\theta_N}^{\infty} \mathrm{d}s\, p(s) = \frac{p}{1-q}, \tag{3}$$

and the *'silent'* interval that separates the ON and OFF thresholds, $p_0 = 1 - (N-2)p - 2p_{\text{edge}}$ (see Methods). Note that for the homogeneous population, $p_0' = 1 - (N-1)p - p_{\text{edge}}$. We call this optimal threshold structure the **infomax** solution.

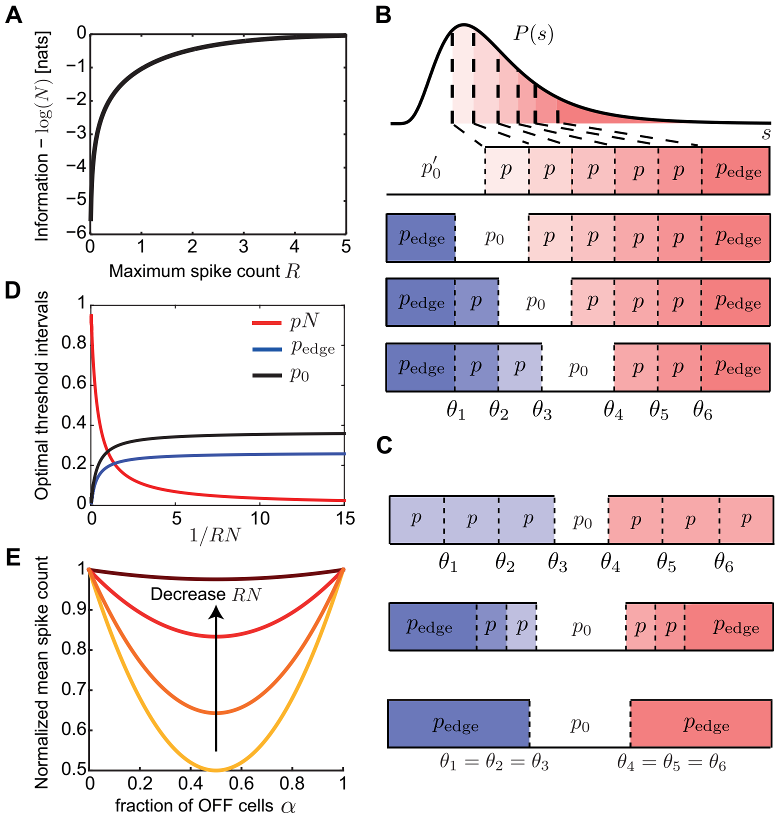

Figure 2. Mutual information when constraining the expected spike count. A. The mutual information between stimulus and response for any mixture of N ON and OFF cells is identical when constraining the expected spike count, R. **B**. The optimal threshold intervals for all possible mixtures of ON (red) and OFF (blue) cells in a population of $N = 6$ cells that achieve the same mutual information about a stimulus from an arbitrary distribution $p(s)$. **C**. The optimal threshold intervals for the equal ON-OFF mixture in a population of $N = 6$ cells and different values of R (equivalently, noise); see also D. Top: low noise ($RN \to \infty$); middle: intermediate noise ($RN = 1$); bottom: high noise ($RN \to 0$). **D**. The optimal threshold intervals as a function of $1/RN$. **E**. The mean spike count required to transmit the same information (see A) by populations with a different fraction of OFF cells (α), normalized by the mean spike count of the homogeneous population with $\alpha = 0$. The different curves denote $RN = \{0.1, 1, 5, 100\}$.

We consider several limiting cases: first, a large population $N \gg 1$ and maximal firing rate per neuron R, which is much larger than $1/N$, i.e. $1 - q = \mathcal{O}(1)$. We call this the **large population regime**. In this regime, $p_{\rm edge} = p = p_0 = 1/(N + 1)$, so the N thresholds divide stimulus space into $N + 1$ equal intervals (Fig. 2C top, D). In this large population regime, we can rewrite the optimal thresholds as a continuous function of cumulative stimulus space; we replace θ_i with $\theta(x)$, where $x = i/N$ is the threshold index between 0 and 1. Then the optimal thresholds equalize the area under the stimulus density, $x(\theta) = \int_{-\infty}^{\theta} p(\theta')\,\mathrm{d}\theta'$. Therefore, the population of cells achieves 'histogram equalization' in that it uses all the available response symbols at equal frequency, as has been shown before for a single cell with many discrete signaling levels in the limit of no noise [4, 47].

In contrast, when the noise is high so that $RN \to 0$, the system performs redundant coding so that each p_i is infinitesimally small and the only two substantial threshold intervals are the edge intervals $p_{\rm edge} = 1/e$, and the silent interval, $p_0 = 1 - 2/e$, which separates the ON and OFF thresholds (Fig. 2C bottom, D). This p_0 is the only non-noisy response state where the firing rate of each cell is zero. This implies that the optimal solution is to place all ON thresholds at roughly the same value, and similarly all OFF thresholds at another value (Fig. 2C bottom). This solution maximizes redundancy across neurons in the interest of noise reduction [2,48,49], consistent with various experimental and theoretical work [2,48,49]. Interestingly, for a small population of two cells, we (and others) have previously shown that in the presence of additional input noise before the signal passes through the nonlinearity, this ‘redundant coding’ regime exists for larger range of noise values [17–19].

In summary, we have derived the total mutual information and distribution of optimal thresholds in a population of binary neurons coding for the same stimulus variable for any stimulus distribution and noise level. While our results agree with previous work in the limit of no noise and an infinite population, we make unique and novel predictions – notably the surprisingly regular structure of the threshold intervals and invariance in information transmission for any ON/OFF mixture – in populations of any number of neurons and with sizable noise relevant for majority sensory systems.

Optimizing information per spike predicts equal ON/OFF mixtures

Our analysis so far showed that maximizing the information equally favors all ON-OFF mixtures independent of the noise level, although the exact distribution of population thresholds at which this information is achieved depends on noise. However, different sensory systems show dominance of OFF [50], dominance of ON [34,51], or similar numbers of ON and OFF [34]. Therefore, we next explored what other criteria might be relevant for neural systems under the efficient coding framework. We considered that neural systems might not just be optimized to encode as much stimulus information as possible, but might do so while minimizing metabolic cost. Therefore, for each ON-OFF mixed population achieving the same total information (Fig. 2A), we calculated the mean spike count used to achieve this information. In the large population regime, if α denotes the fraction of OFF cells, the mean spike count per neuron is $r(\alpha) = R(\alpha^2 + (1-\alpha)^2)/2$ (Methods). This mean spike count per neuron is minimized at $\alpha = 1/2$, where it is half of the mean spike count for the homogeneous population, $r(0) = R/2$ (Fig. 2E). This implies that it is most efficient to split the population into an equal number of ON and OFF cells. As the noise increases, the relative benefits of the equally mixed relative to the homogeneous population decrease (Fig. 2E). In the high noise regime, all mixtures produce roughly the same mean spike count per neuron of R/e (Fig. 2E). Therefore, if a sensory system is optimized to transmit maximal information at the lowest spike cost, our theory predicts similar numbers of ON and OFF neurons, which is consistent with ON-OFF mixtures encountered in some sensory systems [34].

Minimizing mean square error of the optimal linear readout

The efficient coding framework does not specify which quantity neural systems optimize to derive their structure. Until now, we have used the mutual information as a measure of coding efficiency because it tells us how well the population represents the stimulus without regard for how it can be decoded. An alternative criterion for coding efficiency is the ability of downstream neurons to decode this information. A simple biologically plausible decoding mechanism, commonly used in previous studies, is linear decoding [22,26,29,30,52]. Does this alternative measure of efficiency generate the same predictions for how sensory populations coding for the same stimulus variable should allocate their resources to ON vs. OFF neurons? Here, we examine the accuracy of a downstream neuron that estimates the stimulus value s using a weighted sum of spike counts n_i of the upstream population of neurons with thresholds θ_i (Fig. 3A)

$$y = \sum_{i=1}^{N} w_i n_i + w_0. \tag{4}$$

The weights w_i, constant w_0 and thresholds θ_i are optimized to minimize the mean square estimation error (MSE).

Accuracy of the optimal linear readout without noise

We first consider the scenario of low noise ($q \to 0$, or equivalently, $R \to \infty$), in which case the limitation on the accuracy of the stimulus reconstruction comes solely from the discreteness of the rate functions of each cell in the population (Fig. 3B). Unlike maximizing the information, when minimizing the MSE both weights and thresholds depend on the stimulus distribution $p(s)$ (Fig. 3).

Interestingly, we find that in this low noise limit, the optimal MSE is proportional to $1/N^2$ and is the same for all ON/OFF mixtures, including the homogeneous population with all cells of the same type (Fig. 3C,D; see Methods). The optimal decoding weights are given by

$$w_i = \langle s\rangle_i - \langle s\rangle_{i-1} \tag{5}$$

where $\langle s\rangle_i$ are the centers of mass of intervals of $p(s)$ intersected by neighboring thresholds

$$\sum_{0\leq j\leq i} w_j = \frac{\int_{\theta_i}^{\theta_{i+1}} \mathrm{d}s\, s\, p(s)}{\int_{\theta_i}^{\theta_{i+1}} \mathrm{d}s\, p(s)} = \langle s\rangle_i,\ 1\leq i\leq N \tag{6}$$

where we have defined $\theta_{N+1}=\infty$. The optimal thresholds are the average of two neighboring centers of mass

$$\theta_i = \frac{1}{2}(\langle s\rangle_i + \langle s\rangle_{i-1}). \tag{7}$$

The constant term and the stimulus interval not coded by any cell depend on the ON/OFF mixture (Fig. 3C,D; Methods). This gives a recursive relationship that from a set of initial thresholds converges to the optimal solution (see Methods).

To see how this threshold distribution is different than the one predicted by the mutual information, we first consider the large population regime. As for the mutual information, we can rewrite the thresholds θ_i as a continuous function $\theta(x)$ of the cumulative stimulus values $x = i/N$ between 0 and 1. Interestingly, the optimal thresholds equalize not the area under the stimulus density, as in the case of the mutual information, but the area under its one-third power

$$x(\theta) = Z\int_{-\infty}^{\theta} p(\theta')^{1/3}\,\mathrm{d}\theta' \tag{8}$$

where Z is a normalization factor. This result has been previously derived in the context of minimizing the distortion introduced in a pulse-coupled-modulation system due to quantization [53] (reviewed in [54]), as well as in the context of neural coding which maximizes the L_p reconstruction error of the maximum likelihood decoder, of which the mean squared error is the special case for $p = 2$ [25].

We invert the relationship in equation 8 to derive the optimal thresholds $\theta(x)$. Since the optimal MSE depends on the stimulus distribution, from now on we consider the Laplace distribution $p(s) = 1/2\,e^{-|s|}$, which arises when evaluating natural stimulus distributions [23, 55] and has a higher level of sparseness than the Gaussian distribution. In this case, the optimal thresholds become (Fig. 3E,F; see Methods):

$$\theta(x) = \begin{cases} 3\log(2x), & x\leq \frac{1}{2} \\ -3\log(2(1-x)), & x > \frac{1}{2}. \end{cases} \tag{9}$$

The thresholds derived from maximizing information are the same except that the pre-factor is 1 instead of 3, making them less spread out in the tails (Fig. 3F). In particular, the largest thresholds (in magnitude) are $\pm 3\log(2N)$ when optimizing the MSE, three times as large as in the infomax case, $\pm\log(2N)$. To highlight the different predictions for the optimal thresholds under the two efficiency measures, we also plot the cumulative optimal thresholds $\int_{-\infty}^{\theta(x)} p(\theta')\mathrm{d}\theta'$ (Fig. 3G). While the optimal strategy when maximizing the information is to emphasize stimuli with higher likelihood of occurring, minimizing the MSE of the optimal linear readout pushes thresholds logarithmically towards relatively rare stimuli near the tails of the stimulus distribution (Fig. 3F,G).

Taken together, we conclude that in the absence of noise our theory derives equal performance of all ON and OFF mixtures under the two optimality criteria, information maximization and minimizing the optimal linear readout. However, a key difference between the two criteria is the theoretically predicted optimal distribution of thresholds.

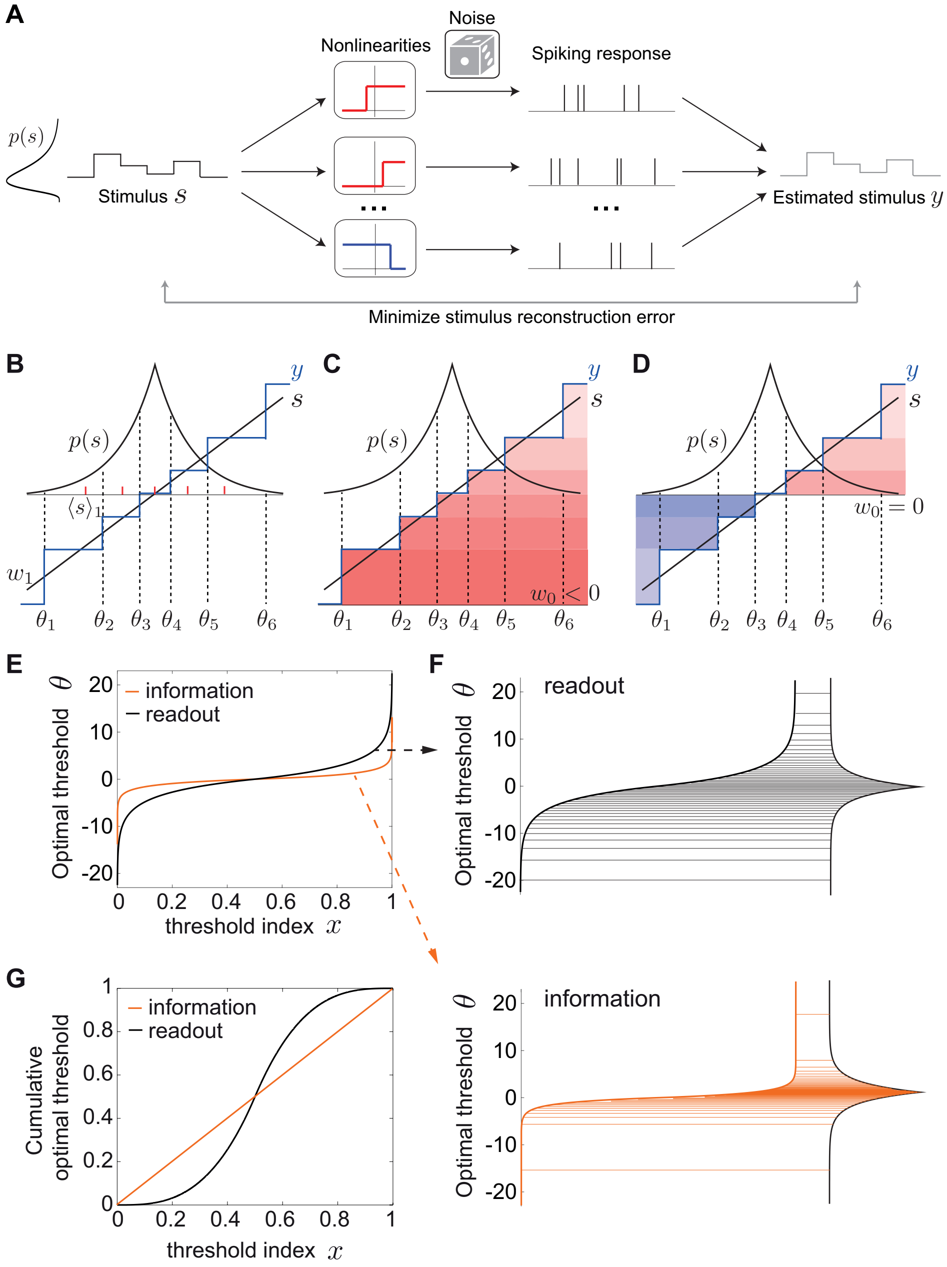

Figure 3. Optimal linear decoding of stimuli. **A**. Framework schematic. A stimulus s from a probability distribution $p(s)$ is encoded by the spiking responses of a population of ON (red) and OFF (blue) cells. We optimize the cells' nonlinearities by minimizing the mean squared error (MSE) between the original stimulus s and the linearly reconstructed stimulus y from the spiking response. **B**. Minimizing the MSE between a stimulus s (black) and its linear estimate y (blue) by a population of (6) ON and OFF cells, in the absence of noise. We show the optimal weight w_1 and the center of mass $\langle s \rangle_1$ of the first threshold interval (red dashes). **C,D**. Any ON-OFF population can achieve the same error with the same set of optimal thresholds and weights but a different constant, w_0. **C**. 6 ON cells ($w_0 < 0$). **D**. 3 OFF and 3 ON cells ($w_0 = 0$). **E**. The optimal thresholds equalize not the area under the stimulus density (as in the case of the mutual information), but the area under its one-third power (Eq. 8). The optimal thresholds are shown for the Laplace distribution. **F**. The information maximizing thresholds partition the Laplace distribution into intervals that code for stimuli with higher likelihood of occurrence (bottom), while minimizing the MSE pushes thresholds to favor rarer stimuli near the tails of the distribution (top). Threshold distributions are the same as in E. **G**. The cumulative optimal thresholds $\int_{-\infty}^{\theta(x)} p(\theta')\mathrm{d}\theta'$ (compare to E).

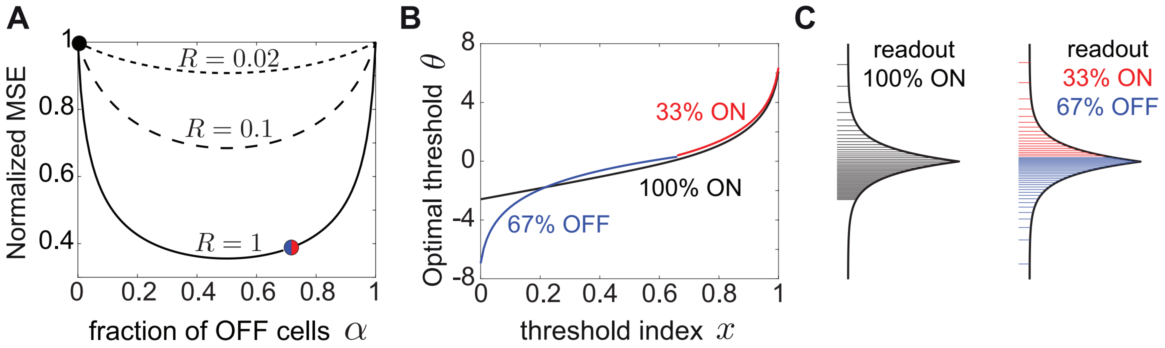

Figure 4. Optimal linear decoding of stimuli with noise depends on the ON/OFF mixture. A. The MSE as a function of the fraction of OFF cells in the population, α, for a different expected spike count, R. The MSE was normalized to the MSE for the homogeneous population of all ON cells. The MSE is shown for $N = 100$ cells and for the Laplace distribution. Symbols indicate the MSE values realized with the thresholds in B and C. **B**. The optimal thresholds for the homogeneous population (black) partition the Laplace stimulus distribution starting with a much larger first threshold than the mixed population with 2/3 OFF cells (blue) and 1/3 ON cells (red). **C**. The optimal thresholds for the Laplace distribution for a homogeneous population (black) and a mixed population with 2/3 OFF cells (blue) and 1/3 ON cells (red). In B and C, $R = 1$. Note the difference in the optimal threshold distribution between the mixed ON/OFF and the homogeneous ON population, especially for small $x = i/N$ (logarithmic in blue vs. linear in black).

Mixed ON/OFF populations in the presence of noise

In biologically realistic scenarios with non-negligible noise, however, we find that mixed ON/OFF populations show a dramatic improvement of the MSE over predominantly homogeneous populations (Fig. 4A). For the Laplace distribution we have considered so far, and different noise values, we find that the optimal fraction of OFF cells in the population is $\alpha = 1/2$. Although there is a unique best ON/OFF mixture, the best linear stimulus reconstruction achieved by other populations with ON-OFF mixtures closer to the optimal 1/2 mixture is similar (i.e. the MSE around $\alpha = 1/2$ is flat). The worst stimulus reconstruction is achieved by the homogeneous population with all cells of ones type (all ON or all OFF), which has the highest MSE. As the noise q decreases (R increases) further, this difference in performance between the mixed and homogeneous populations becomes quite dramatic, see for example $R = 1$ (Fig. 4A).

In addition to the big difference in coding performance between mixed and homogeneous populations, incorporating biologically realistic noise also affects the theoretically derived distribution of optimal thresholds (Fig. 4B). While in mixed populations the thresholds are distributed logarithmically towards relatively rare values at the tails of the stimulus distribution (Eq. 9; see Fig. 4B,C and Methods), for the homogeneous population the optimal thresholds exhibit a distinct asymmetry. A large fraction of thresholds are distributed linearly as a function of their index, while the remaining thresholds are distributed logarithmically as before:

$$\theta(x) = \begin{cases} -(1-x)\log(RN), & 0 < x \leq \left(1 - \frac{\sqrt{2}}{\log(RN)}\right)^{-1} \\ -2\log(1-x), & \left(1 - \frac{\sqrt{2}}{\log(RN)}\right)^{-1} < x \leq 1, \end{cases} \tag{10}$$

although the noise has the effect of concentrating the thresholds near more likely stimuli, increasing the redundancy of the code. Moreover, the smallest threshold for the homogeneous population is much larger than the smallest threshold for any mixed population, suggesting that there is a large region of stimuli that is not coded by any cell in the homogeneous case (Fig. 4C), which is the reason for the significantly lower MSE.

In summary, using the MSE of the optimal linear decoder as a measure of efficiency can fundamentally alter our conclusions about how to split a population into ON and OFF cells and how to distribute the population thresholds to achieve the optimal stimulus reconstruction. At biologically realistic noise levels, coding by mixed ON-OFF populations is much better than by a homogeneous population, with qualitatively distinct optimal threshold distributions.

The optimal ON-OFF mixture of the linear readout depends on the asymmetry in the stimulus distribution

Since the MSE as a measure of efficiency depends on the stimulus distribution, we asked how the stimulus distribution can affect optimal population coding. The distribution of natural stimuli may be asymmetric around the most likely stimulus. For example, the distribution of contrasts in natural images, and the intensity of natural sounds are indeed skewed towards more negative values [20, 56–61]. Therefore, we instead consider an asymmetric Laplace distribution $p(s) \propto e^{s/\tau_-}$ for $s < 0$ and $p(s) \propto e^{-s/\tau_+}$ for $s \geq 0$ where we take $\tau_- > \tau_+$. Minimizing the MSE one finds that the optimal way to divide a population into ON and OFF respects these stimulus asymmetries. Increasing the negative stimulus bias τ_-/τ_+ favors more OFF cells (Fig. 5A,B). The optimal thresholds for these different stimulus biases are best compared in the cumulative space of stimulus (Fig. 5C). Increasing the bias also pushes the thresholds towards more negative stimulus values, which occur with higher probability than positive stimuli.

At a fixed level of stimulus bias, increasing the noise further accentuates the asymmetry in the optimal ON-OFF mixture (Fig. 5D,E). As the noise becomes non-negligible, the optimal thresholds lose the logarithmic spread at the tails of the stimulus distribution and begin to code for more likely stimuli that occur with a higher probability. At the same time, a larger region of stimulus values near the median is no longer coded by any cells, i.e. the gap between ON and OFF thresholds becomes larger (Fig. 5F). Had we considered the limit of zero noise or infinitely large populations as previous studies [22–25], we would not have been able to identify these differences between the optimal thresholds that result in conditions of biologically realistic noise and finite populations.

In summary, our theory predicts different optimal ON-OFF numbers at which the lowest MSE is achieved depending on asymmetries in the stimulus distribution and the noise level. Indeed in nature, the relative predominance of ON and OFF cells in diverse sensory systems can be different (Table 1). Therefore, if we know the natural stimulus distribution being encoded by a population and the bounds on cells' firing rates, we can predict the optimal ON and OFF numbers, as well as the response thresholds of the cells and compare them to experimental observations.

Predicting stimulus distributions from experimentally measured thresholds

Here we propose to reverse our efficient coding framework and starting from an experimentally measured distribution of thresholds, to predict the distribution of natural stimuli that the thresholds could be optimized to encode (Fig. 6A). This could be particularly relevant for sensory systems where the distribution of the sensory variable being encoded is unknown. We decided to test this approach on odor concentration coding in the olfactory system of *Drosophila* larvae given recently published data [37]. The first stage of olfactory processing in *Drosophila* larvae is implemented by a population 21 olfactory receptor neurons (ORNs), which code for a broad space of odorants and concentrations [37]. We hypothesized that these ORNs might have distributed their thresholds at different concentrations to optimally encode any particular odor. In the classical efficient coding approach, knowing the distribution of odor concentrations would allow us to predict the optimal thresholds. In the reversed approach that we use here, knowing the distribution of thresholds allows us to predict the distribution of concentrations of a known odor (Fig. 6A).

A recent study estimated these thresholds by recording from the entire ORN population [37]. The responses for 34 odorants over a five-fold magnitude in concentration were well described by a common Hill function with a shared steepness, but different activation thresholds. Pooling all thresholds across the different odorants and concentrations revealed a power law distribution. To use this threshold distribution in our theoretical framework, where a range of thresholds codes for the intensity of a single stimulus, we had to make a critical assumption. Specifically, we assumed that the population thresholds spanning the range of concentrations for any one odor are a shuffled version of the population thresholds for other odorants. This was justified by an analysis of a related data set [35], in which the distribution of ORN firing rates was found to be stereotyped across different odors [36].

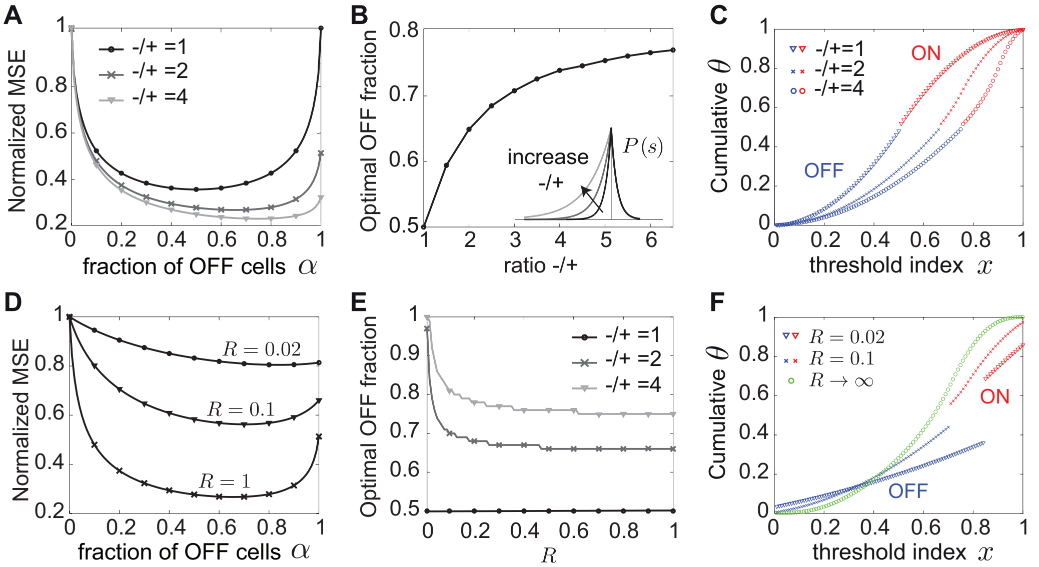

Figure 5. The optimal ON/OFF mixture derived from the linear readout is tuned to asymmetries in the stimulus distribution. A. The MSE as a function of the fraction of OFF cells (α) normalized to that for the homogeneous population of all ON cells ($\alpha = 0$). The MSE is shown for an asymmetric Laplace distribution with varying negative to positive bias $-/+$, expected spike count $R = 1$ and $N = 100$ neurons. **B**. The optimal fraction of OFF cells as a function of stimulus bias of the asymmetric Laplace distribution and $R = 1$. **C**. The optimal thresholds for the ON-OFF mixtures (50%, 66% and 75%) in A that yield the lowest MSE, while varying negative to positive bias $-/+ = \{1, 2, 4\}$. **D**. Same as A but for an asymmetric Laplace distribution with a negative bias $-/+ = 2$ and varying R (equivalently, noise). **E**. The optimal fraction of OFF cells as a function of R for different stimulus bias of the asymmetric Laplace distribution. **F**. The optimal thresholds for the ON-OFF mixtures (84%, 70% and *any*) in D that yield the lowest MSE, while varying $R = \{0.02, 0.1, \infty\}$.

Using our optimal coding framework with a population of only ON neurons (since ORNs have monotonically increasing response functions with concentration), we derived the mostly likely stimulus distribution of odorant concentrations for each of the two efficiency measures. The predicted distribution of odorant concentrations follows a power law distribution with an exponent determined by the efficiency measure. Given a measured distribution of thresholds which follows a power law with an exponent of -0.58 (Methods, Fig. 6B,C) and assuming an infomax code we predict that the distribution of odorant concentrations should also be a power law with an exponent of -0.58 (Methods, Fig. 6D). In contrast, assuming a code that minimizes the stimulus reconstruction error, the distribution of odorant concentrations should be a power law with an exponent of -1.74 (Methods, Fig. 6D). Indeed, many processes like convection and turbulence can generate power law dynamics [62], but the exact exponents will need to be determined, for instance by measuring the volatiles from natural environments [63, 64]. Although complex temporal dynamics in the stimulus can further complicate ORN coding of fluctuating odorant concentrations, the measured temporal filter across ORNs is remarkably stereotyped, suggesting that the olfactory code is similar between static and dynamic odor environments.

We note that in our analysis we explicitly assume that the goal of the olfactory system is to estimate the concentration of any one odor with high fidelity, therefore it is only valid for experiments where only one odor is present. However, the optimization problem faced by the olfactory system might be different, i.e. to determine which, of many, odors are present. Therefore, it is possible that the optimal thresholds in these two cases may be different.

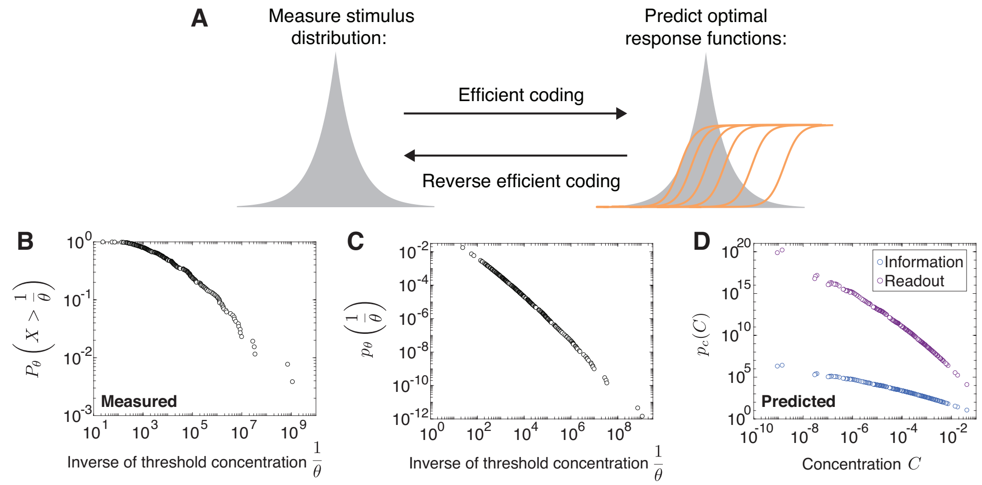

Figure 6. Deriving a distribution of stimulus intensities from experimentally measured thresholds. A. Our efficient coding framework enables us to predict the optimal distribution of thresholds given a known stimulus distribution. By reversing our framework, we derive the stimulus distribution from a distribution of measured thresholds assuming optimal coding under the two optimality criteria. **B**. Log-log plot of the cumulative distribution of the inverse of thresholds from measured dose-response curves of the entire population of ORNs in the *Drosophila* larva olfactory system [37]. This is well described by a power law with exponent -0.42. **C**. The probability distribution of the inverse of optimal thresholds derived from the data in B. This is well described by a power law with exponent -0.58. **D**. Predicted distribution of concentrations across different odorants when assuming optimal coding by maximizing information or minimizing the MSE of the best linear decoder. This is well described by a power law with exponents -0.58 and -1.74, respectively. The proportionality constant is not shown.

Discussion

Information in neural circuits is processed by many different cell types, but it remains a challenge to understand how these distinct cell types work together. Here we treat a puzzling aspect of neural coding, how do discrete cell types conspire to collectively encode a single relevant variable in responses of opposite polarity? To evaluate such a population code we built on the framework of efficient coding and extended it in several novel ways: by considering nonlinear processing, biologically realistic levels of noise, short coding windows, and the coordination of responses in populations of any size – factors which may vary across sensory systems. We then derived two aspects of the population code, namely how to optimally split a population into ON and OFF cells, and how to allocate the thresholds of the individual neurons as a function of the noise level, the stimulus distribution and the optimality measure.

Optimal ON-OFF mixtures and comparison to experimental data

We considered two different measures of coding efficiency that are in common use [22, 26–30]: the mutual information between stimulus and responses, and the mean squared error of the linearly reconstructed stimulus. The first aspect of our predictions applies to the expected mixture of ON and OFF cells. If one chooses mutual information as the efficiency measure, then all ON/OFF mixtures in the population perform identically once the thresholds are adjusted (Fig. 2). This result holds independent of the noise level and the shape of the stimulus distribution, and generalizes for response functions with any number of discrete firing rate levels. However, the number of spikes required for this performance, and thus the metabolic cost, differs greatly depending on the ON/OFF ratio. If one considers the information per spike as the relevant measure, then a system with equal number of ON and OFF cells is most efficient.

When we require the stimulus to be read out by an optimal linear readout, different ON/OFF mixtures also achieve similar coding performance but only in the absence of noise (Fig. 3). In the biologically relevant regimes of non-negligible noise, noise has a dramatic influence on the optimal performance realized by different

ON/OFF mixtures (Fig. 4). Populations with a similar number of ON and OFF cells have a much smaller decoding error than populations dominated by one cell type. The extreme case of the homogeneous population performs substantially worse that any mixed population (Fig. 4). In the case of asymmetries in the stimulus distribution, as encountered in many natural sensory stimulus distributions [20, 56–61], minimizing the linear reconstruction error predicts that the optimal ON/OFF mixture should be tuned to these asymmetries and the amount of noise (Fig. 5).

How do these predictions accord with known neural codes? Since our theory applies to populations of sensory neurons that code for the same stimulus variable, we need to consider sensory systems where this is the case. Analyzing raw stimulus values, such as the light intensity in a natural scene or the intensity of natural sounds, results in distributions which are skewed towards negative stimuli [20, 56–61]. Our linear decoding theory then predicts that more resources should be spent on OFF. Indeed, in the fly visual system, the OFF pathway is overrepresented in the circuit for computations that extract motion vision, with the L1 neurons being responsible for the processing of ON signals, while both L2 and L3 neurons for OFF [9, 50]. These neurons are repeated in each cartridge, thus together code for the same spatial location. Hence, at least for the fly visual system, our efficient coding results are in accord with naturally encountered ON/OFF ratios. In contrast, the vertebrate retina represents a visual stimulus with spikes across diverse types of retinal ganglion cells, which differ in their spatial and temporal processing characteristics [65, 66]. Certain types of ganglion cell come in ‘paramorphic pairs,’ meaning an ON-type and an OFF-type that are similar in all other aspects of their visual coding. A previous study by Ratliff et al. (2010) derived the optimal numbers of ON and OFF retinal ganglion cells for encoding natural scenes assuming maximal information transmission, as a function of the spatial statistics in these natural stimuli. In their model, every ganglion cell in the population encodes a different stimulus variable, because it looks at a different spatial location. In contrast, our theory can only be applied to populations that code for the same same stimulus feature, which may in fact contain only one of each type (ON and OFF), but requires further experiments to determine the exact numbers. To properly account for all thirty types of retinal ganglion cells will require more complete models that include the spatial dimension and the encoding of different visual features.

Besides the visual system, there are other examples in biology where different numbers of ON and OFF cells are encountered, and where our theory more naturally applies with populations of neurons encoding a one-dimensional stimulus (Table 1). Single neurons in monkey somatosensory cortex show diverse ON and OFF responses to the temporal input frequency of mechanical vibration of their fingertips. While most neurons in primary somatosensory cortex (S1) tune with a positive slope to the input frequency (ON), about half of the neurons in secondary somatosensory cortex (S2) behave in the opposite way (OFF) [33, 34]. Opposite polarity pathways are also observed in thermosensation, where receptor proteins activated directly by positive and negative changes in temperature enable the detection of thermal stimuli. Four mammalian heat-activated and two cold-activated ion channels have been shown to function as temperature receptors [31, 32]. Given these observations of ON-OFF asymmetries, one is led to conclude that information per spike may not be the cost function that drove evolution of this system, since that would predict equal numbers of ON and OFF cells. Thus, whether these different experimental observations are consistent with maximizing mutual information, optimal linear decoding, or yet a different objective function or task (e.g. [67–69]), remains to be seen (see also our discussion on the generality of assumptions). Other neuronal systems are candidates for similar analysis, for instance, auditory nerve fibers [44], motor cortex [70], and primary vestibular neurons [71].

Optimal threshold distributions and comparison to experimental data

Beyond predicting ON-OFF numbers, which has been the main focus of different models about the vertebrate retina [20], we also predict the structure of response thresholds. Generally, maximizing information implements an optimal strategy which emphasizes stimuli that occur with higher probability (Fig. 3, 4). In the limit of low noise, this is consistent with the well-known strategy of ‘histogram equalization’ [4], but we generalize this result to any amount of biologically realistic noise. Importantly, the optimal interval size depends on the level of noise with larger noise favoring smaller threshold intervals, implying a strategy closer to redundant coding. In contrast to the information, minimizing the mean square error of the linear readout implements a more conservative strategy that utilizes more cells in the encoding of rarer stimuli due to a larger error penalty

Table 1. List of experimentally measured ON and OFF neuron numbers in different sensory systems.

Sensory system	ON/OFF numbers
Primary somatosensory cortex S1 (primate) [34]	ON dominance
Secondary somatosensory cortex S2 (primate) [34]	ON $\approx$ OFF
Visual system (insect) [50]	OFF dominance
Olfactory system (mammalian, insect) [35, 72, 73]	Unknown
Thermosensory system (mammalian, insect) [31, 32]	OFF dominance
Mechanosensory system (mammalian) [51]	ON dominance

(Fig. 3, 4).

Our theoretical framework applies to the case when the distribution of stimuli encoded by the cells is known, and the only problem is to estimate the value of the stimulus by appropriately distributing the cells' thresholds. In the case of vision, for example, this implies estimating the light intensity or contrast level. A direct test of our theoretical predictions for the optimal thresholds would require simultaneous measurement of the population response thresholds, which is within reach of modern technology [66]. In the meantime, we reversed our theoretical approach and starting from an experimentally measured distribution of thresholds, we predicted the distribution of natural stimuli that the thresholds might optimally encode. We applied this approach for the population of ORNs in the olfactory system of *Drosophila* larvae. However, unlike vision, applying our framework to olfaction presents a different problem. Here, the goal of the olfactory system is primarily to determine whether or not an odor is present, not its concentration. Therefore, our analysis is only appropriate when only one odor is present, and it can be inferred with high certainty. In this case, we assumed that the ORNs code for the distribution of concentrations of the present odor by diversifying their thresholds. The ORNs' experimentally described tuning curves were identical in shape, with response thresholds following a power law distribution [37]. Since the probability distribution of ORN firing rates is stereotyped across different odors [36], we assumed that the thresholds coding the range of concentrations for any one odor are a shuffled version of the thresholds for other odorants. We derived the stimulus distribution of concentrations for any one tested odor, under the two optimality measures, the maximal mutual information and the minimal error of the best reconstructed stimulus (Fig. 6). This threshold distribution was also a power law with an exponent dependent on the efficiency measure. Whether these distributions correspond to distributions of odorant concentrations found in natural olfactory environments remains to be tested, and techniques for collecting the volatiles from natural encountered odors now exist [63, 64]. These distributions would be strongly influenced by processes like convection and turbulence, which can give rise to power law dynamics [62]. Although these are dynamical variables that fluctuate in time, we propose that the distributions can be build by pooling different aspects of the dynamics over extended time periods. In that context, our theoretical framework would apply to populations which have been adapted to these distributions over those long periods of time. It is possible that when considering a different optimization problem implemented by the olfactory system which aims to determine which, of many, odors are present, a very different distribution of thresholds than those our theory predicts would be optimal.

Generality of assumptions and relationship to previous work

The two efficiency measures that we have used are entirely agnostic about the content of signal transmission. However, faithful encoding of signals is not the only fitness requirement on a sensory system, for example, some stimuli may have greater semantic value than others. Or, the aim may be to extract task-relevant sensory information as in the case of the Information Bottleneck framework [67, 68], or to achieve optimal inference of behaviorally-relevant properties in dynamic stimulus environments [69]. Other recent approaches, such as Bayesian efficient coding, optimize an arbitrary error function [74]. Since our framework aims to encode a stimulus as best as possible, we propose that it may be most appropriate for early sensory processing, where stimulus representation might be the goal.

The efficient coding hypothesis was originally proposed by Attneave [75] and Barlow [1], who studied deterministic coding, in the absence of noise. Since then, many studies have investigated efficient coding strategies under different conditions. Atick and Redlich introduced noise and demonstrated that efficient coding can be used to explain the center-surround structure of receptive fields of retinal ganglion cells, which changes to center-only structure as the signal-to-noise increases [2, 76]. Including nonlinear processing in the limit of low noise produced Gabor-like filters encountered in the primary visual cortex [13]. However, we now know that already the very first stages of processing in many sensory systems are nonlinear, consist of many parallel pathways and exhibit substantial amount of noise [77] – important aspects of coding that we simultaneously incorporate in our analysis. Our work differs from a previous report on ON and OFF cells in the vertebrate retina which proposed a simplified noise model implemented by assuming a finite number of signaling levels (i.e. firing rates), which does not incorporate spiking [20].

We considered a Poisson noise model of spiking which is commonly used in many studies. Our results are especially relevant in the high noise regime, which corresponds to short coding windows commonly encountered in biology, for instance, a few spikes per coding window [28, 38, 78]. In the low noise regime when the coding window is sufficiently long, or there is a large number of neurons, our results agree with previous studies on infomax and the optimal linear readout [4, 25, 79]. Efficient coding in the high noise regime has previously been examined, but only in terms of the transfer function of a single neuron, which was shown to be binary [27, 43]. We go beyond this work and provide analytical solutions for how a population of neurons should coordinate their response ranges to optimally represent a given stimulus in the realistic regimes of short encoding times.

We used a binary rate function to describe single neuron responses because it gives our problem analytical tractability and it still represents a significant departure from previous efficient coding frameworks based on linear processing [2, 3, 20, 21], long coding windows or infinitely large populations [22–25]. Indeed, discretization in neural circuits is a common phenomenon that is not only relevant for sensory coding, but also for neuropeptide signaling, ion channel distributions and information transmission in genetic networks [39, 80]. Considering more general nonlinearities is currently only tractable with numerical simulations or in the case of optimizing a local efficiency measure, the Fisher information, which may not accurately quantify coding performance in finite size populations or biologically realistic noise (e.g. low firing rates or short coding windows) [25, 49, 81–83].

Summary

Given the ubiquity of ON/OFF pathway splitting in different sensory modalities and species, our framework provides predictions for the optimal ON/OFF mixtures and the functional diversity of sensory response properties that achieve this optimality in many sensory systems based on the distribution of relevant sensory stimuli, the noise level and the measure of optimality. Our theoretical approach is sufficiently general and is not fine-tuned to the specifics of any one experimental system. The different predictions that we make depending on the model assumptions could help determine the specific optimality criteria operating in different sensory systems where different ON-OFF mixtures and tuning properties have been observed. Directed experiments to compare the predicted and measured threshold distributions will test whether the efficient coding criteria proposed here are a likely constraint shaping the organization and adaptation of sensory systems.

Materials and Methods

Mutual information and proof of the Equal Coding Theorem

First we prove the Equal Coding Theorem for a general population with N binary neurons. Without loss of generality we assume that the neurons' thresholds are:

$$\theta_1 \leq \ldots \leq \theta_N \tag{11}$$

and we define the special $\theta_0 = -\infty$ and $\theta_{N+1} = \infty$. The Shannon mutual information between the stimulus s and the spiking response $\boldsymbol{n}$ of the population is the difference between response and noise entropy:

$$I(s, \boldsymbol{n}) = H(\mathbf{n}) - H(\boldsymbol{n}|s) = -\langle \log p(\boldsymbol{n}) \rangle_{\boldsymbol{n}} + \sum_{i=1}^{N} \langle \log p(n_i|s) \rangle_{n_i, s} \tag{12}$$

where $\langle \cdot \rangle_x$ denote averages over the distribution $p(x)$ and $p(\boldsymbol{n}) = \langle p(\boldsymbol{n}|s) \rangle_s$. We assume that stimulus encoding by all neurons is statistically independent conditional on s so that

$$p(\boldsymbol{n}|s) = \prod_{i=1}^{N} p(n_i|s). \tag{13}$$

Given the Poisson noise model, knowing the stimulus s unambiguously determines the response firing rate ν; for instance, for an ON cell if $s < \theta$, $\nu = 0$ and if $s \geq \theta$, $\nu = \nu_{\max}$. We can replace $p(n_i|s)$ with $p(n_i|\nu)$ which is Poisson distributed: $p(n_i|\nu) = \frac{[\nu T]^{n_i}}{n_i!} e^{-\nu T}$.

We prove that $I(s; \boldsymbol{n}) = I(\boldsymbol{\nu}, \boldsymbol{n})$. To see this, we write

$$\begin{aligned} I(s, \boldsymbol{n}) &= \sum_{\boldsymbol{n}} \int_s \mathrm{d}s\, p(s)\, p(\boldsymbol{n}|s) \log \frac{p(\boldsymbol{n}|s)}{p(\boldsymbol{n})} && (14) \\ &= H(\boldsymbol{n}) + \sum_{\boldsymbol{n}} \int_s \mathrm{d}s\, p(s)\, p(\boldsymbol{n}|s) \log p(\boldsymbol{n}|s). && (15) \end{aligned}$$

Using Eq. 13, this becomes

$$\begin{aligned} I(s, \boldsymbol{n}) &= H(\boldsymbol{n}) + \sum_{\boldsymbol{n}} \int_s \mathrm{d}s\, p(s) \prod_j p(n_j|s) \sum_i \log p(n_i|s) && (16) \\ &= H(\boldsymbol{n}) + \sum_i \sum_{n_i} \int_s \mathrm{d}s\, p(s) p(n_i|s) \log p(n_i|s) && (17) \end{aligned}$$

Similarly, we derive

$$\begin{aligned} I(\boldsymbol{\nu}, \boldsymbol{n}) &= \sum_{\boldsymbol{n}} \sum_{\boldsymbol{\nu}} p(\boldsymbol{\nu})\, p(\boldsymbol{n}|\boldsymbol{\nu}) \log \frac{p(\boldsymbol{n}|\boldsymbol{\nu})}{p(\boldsymbol{n})} && (18) \\ &= H(\boldsymbol{n}) + \sum_{\boldsymbol{n}} \sum_{\boldsymbol{\nu}} p(\boldsymbol{\nu})\, p(\boldsymbol{n}|\boldsymbol{\nu}) \log p(\boldsymbol{n}|\boldsymbol{\nu}) && (19) \end{aligned}$$

which since

$$p(\boldsymbol{n}|\boldsymbol{\nu}) = \int_s \mathrm{d}s\, p(\boldsymbol{n}|s) p(s|\boldsymbol{\nu}) = \prod_i \int_s \mathrm{d}s\, p(n_i|s) p(s|\boldsymbol{\nu}) = \prod_i p(n_i|\nu) = \prod_i p(n_i|\nu_i) \tag{20}$$

becomes

$$\begin{aligned} I(\boldsymbol{\nu}, \boldsymbol{n}) &= H(\boldsymbol{n}) + \sum_{\boldsymbol{n}} \sum_{\boldsymbol{\nu}} p(\boldsymbol{\nu})\, p(\boldsymbol{n}|\boldsymbol{\nu}) \sum_i \log p(n_i|\nu_i) && (21) \\ &= H(\boldsymbol{n}) + \sum_i \sum_{n_i} \sum_{\nu_i} p(\nu_i) p(n_i|\nu_i) \log p(n_i|\nu_i). && (22) \end{aligned}$$

Now, for a given i and a corresponding given spike count n_i, which without loss of generality we assume is an ON cell with threshold θ_i, we take the second term from Eq. 17 and split the integral:

$$\int_s \mathrm{d}s\, p(s)p(n_i|s)\log p(n_i|s) = \int_{-\infty}^{\theta_i} \mathrm{d}s\, p(s)p(n_i|s)\log p(n_i|s) + \int_{\theta_i}^{\infty} \mathrm{d}s\, p(s)p(n_i|s)\log p(n_i|s) \tag{23}$$

$$= \int_{-\infty}^{\theta_i} \mathrm{d}s\, p(s)p(n_i|\nu_i = 0)\log p(n_i|\nu_i = 0) + \int_{\theta_i}^{\infty} \mathrm{d}s\, p(s)p(n_i|\nu_i = \nu_{\max})\log p(n_i|\nu_i = \nu_{\max})$$

$$= \sum_{\nu_i} p(\nu_i)p(n_i|\nu_i)\log p(n_i|\nu_i) \tag{24}$$

because we can just integrate out the s. Therefore, from Eqs. 17, 22 and 24, we get $I(s;\boldsymbol{n}) = I(\boldsymbol{\nu},\boldsymbol{n})$. Note that for a single cell, Nikitin et al. [43] also proved the same equality of information using a different approach.

For a binary response function with two firing rate levels, 0 and $\nu_{\max}$, we can lump together all states with *nonzero* spike counts into a single state which we denote as $\mathbf{1}$. Correspondingly, the state with zero spikes is $\mathbf{0}$. Hence, we can evaluate the mutual information between stimulus and spiking response using the following expressions for the spike count probabilities:

$$\begin{aligned} p(\mathbf{0}|\nu = 0) &= 1, & p(\mathbf{1}|\nu = 0) &= 0, \\ p(\mathbf{0}|\nu = \nu_{\max}) &= q, & p(\mathbf{1}|\nu = \nu_{\max}) &= 1 - q, \end{aligned} \tag{25}$$

where $q = e^{-R}$ and $R = \nu_{\max}T$ denote the level of noise in the system.

We can derive the expression for the mutual information between stimulus and response given the N intervals

$$u_i = \int_{\theta_{N+1-i}}^{\infty} \mathrm{d}s\, p(s), \quad i = 1, \ldots, m \tag{26}$$

for the m ON cells and

$$u_i = \int_{-\infty}^{\theta_{i-m}} \mathrm{d}s\, p(s), \quad i = m+1, \ldots, N \tag{27}$$

for the OFF cells, see Figure 7.

We prove the Equal Coding Theorem by showing that the mutual information coded by a population of N ON cells is the same as that for any arbitrary mixture of ON and OFF cells, for instance, a population with m ON cells (with indices $1, 2, \ldots m$) and $N - m$ OFF cells (with indices $m+1, \ldots, N$). At the optimal solution the ON cells have larger thresholds than the OFF cells. This is due to the assumed Poisson noise model, where the states at which a given cell's firing rate is 0 are non-noisy and determine the stimulus with certainty. The total information can be described as the information from observing the ON cells $1, \ldots m$, plus any additional information gained from observing cells $m+1, \ldots, N$. Below we demonstrate that this additional information is identical independent of whether the $N - m$ cells are ON (in which case the population is homogeneous and comprised of all ON cells) or OFF type (in which case the population is mixed). This turns out to be the case, as long as the thresholds of the additional $N - m$ cells are appropriately adjusted.

If a spike was observed from ON cells $1, 2, \ldots m$, then no additional information is gained from cells $m+1, \ldots, N$, independent of their type because their firing rate is constant over the entire stimulus interval in which cells $1, \ldots, m$ fire. Then, the total mutual information achieved by all N cells, $I_N(s, \boldsymbol{n})$, is equal to the mutual information obtained from observing the m ON cells, $I_m(s, \boldsymbol{n})$:

$$I_N(s, \boldsymbol{n}; \{u_1, \ldots, u_N\}) = I_m(s, \boldsymbol{n}; \{u_1, \ldots, u_m\}) \tag{28}$$

where we explicitly denote the dependence of the mutual information on the threshold intervals, u_i's. If no spike was observed from the ON cells $1, 2, \ldots m$, then we get additional information from the remaining cells $m+1, \ldots, N$, but we need to consider the change in the stimulus distribution posterior to seeing no spike. Such a change in the stimulus distribution is equivalent to adjusting the thresholds of cells $m+1, \ldots, N$, and

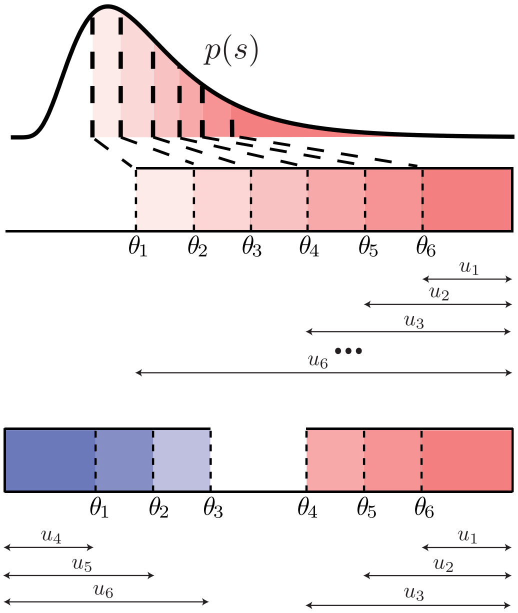

Figure 7. Thresholds θ_i and intervals between thresholds u_i for a population of 6 cells. Top: a homogeneous population with 6 ON cells; bottom: a mixed population with 3 ON and 3 OFF cells.

as a result, the threshold intervals. If none of the ON cells $1, \dots m$ fired, then, we can formally write the total information as follows (see Fig. 7):

$$I_N(s, \boldsymbol{n}; \{u_1, \dots, u_N\}) = I_m(s, \boldsymbol{n}; \{u_1, \dots, u_m\}) + Q_m I_{m+1,\dots,N|1,\dots,m}(s, \boldsymbol{n}; \{u'_{m+1}, \dots, u'_N\}) \tag{29}$$

where Q_m is the probability that none of the ON cells $1, \dots m$ fired, and $I_{m+1,\dots,N|1,\dots,m}$ is the additional mutual information gained from the remaining $N-m$ cells with adjusted thresholds, and consequently threshold intervals, u'_i.

By writing the information in this manner, we have only assumed that the first m cell are ON, but have not assumed anything about the type of the additional $N-m$ cells. In fact, for any ON-OFF mixture given by the number of ON cells m, one can choose the same thresholds $\theta_1, \dots, \theta_m$ (and thus thresholds intervals $u_1, \dots, u_m$) for the first m ON cells, and then change the thresholds $\theta_{m+1}, \dots, \theta_N$ (and thus threshold intervals $u_{m+1}, \dots, u_N$) of the remaining $N-m$ cells so as to produce the same adjusted threshold intervals, $u'_{m+1}, \dots, u'_N$.

How can this readjustment be done for the different mixtures? If no spike was observed from the ON cells $1, \dots, m$, then the stimulus distribution to be coded by the remaining cells changes from the prior $p(s)$ to a new posterior distribution

$$p(s|\mathbf{0}) = p(\mathbf{0}|s)\frac{p(s)}{p(\mathbf{0})} = p(s)\frac{p(\mathbf{0}|s)}{Q_m}. \tag{30}$$

1. **If the remaining $N-m$ cells are ON**, the region of reduced $p(s)$ is entirely within the response region. Thus, the revised probability of having the stimulus in the response region is

$$\begin{aligned} u'_i &= \int_{\theta_i}^{+\infty} \mathrm{d}s\, p(s|\mathbf{0}) = \frac{1}{Q_m}\int_{\theta_i}^{+\infty} \mathrm{d}s\, p(\mathbf{0}|s)p(s) \\ &= \frac{1}{Q_m}\int_{1-u_i}^{1} \mathrm{d}x\, p(\mathbf{0}|x) = \frac{u_i - (1-Q_m)}{Q_m}. \end{aligned} \tag{31}$$

where $x = \int_0^s p(s')\mathrm{d}s'$.

2. **If the remaining $N-m$ cells are OFF**, the region of reduced $p(s)$ is entirely outside their response region. Thus, their revised probability is

$$u_i' = \int_0^{\theta_i} \mathrm{d}s\, p(s|\mathbf{0}) = \frac{1}{Q_m}\int_0^{\theta_i} \mathrm{d}s\, p(\mathbf{0}|s)p(s) = \frac{1}{Q_m}\int_0^{u_i} \mathrm{d}x\, p(\mathbf{0}|x) = \frac{u_i}{Q_m}. \tag{32}$$

Therefore, the readjustment of the threshold intervals can be done differently for a homogeneous population when the remaining $N-m$ cells are all ON, vs. a mixed population when the remaining $N-m$ cels are all OFF. Since m can be anything between 1 and N, this covers all possible mixtures of ON and OFF cells, where

$$u_i' = \begin{cases} \frac{u_i-(1-Q_m)}{Q_m}, & \text{homogeneous population with } N \text{ ON cells} \\ \frac{u_i}{Q_m}, & \text{mixed population with } m \text{ ON cells and } N-m \text{ OFF cells} \end{cases} \tag{33}$$

To find the maximal mutual information one needs to maximize Eq. 29 with respect to all the thresholds (i.e. threshold intervals). Since the homogeneous population of N ON cells and the mixed population of m ON cells and $N-m$ OFF cells share the same m ON cells, maximizing the total mutual information I_N in Eq. 29 is equivalent to maximizing the additional mutual information $I_{m+1,\ldots,N|1,\ldots,m}$ gained from the remaining $N-m$ cells with adjusted threshold intervals according to Eq. 33. This explains why the maximum information is identical between the purely homogeneous population with N ON cells and a mixed population where $N-m$ cells are OFF.

Thresholds when optimizing the mutual information: a homogeneous population

Next we derive the optimal thresholds for the homogeneous population with N ON cells, and later derive the thresholds of the $N-m$ OFF cells after swapping.

If the thresholds are ordered in ascending order as assumed above, then $u_1 < u_2 < \ldots < u_N$ (S3 Figure). The mutual information of N ON cells can be written as follows. First, for a population of $N=1$ cells this has the form

$$I_1 = H(n_1) - H(n_1|s) = h(u_1(1-q)) - u_1 h(1-q), \tag{34}$$

where h is the entropy of a binary variable, $h(u) = -u\log u - (1-u)\log(1-u)$. For a population of $N=2$ cells it has the form

$$I_2 = I_1 + P(n_1=0)I_{2|1} = g(u_1) + (1-u_1(1-q))g\left(u_2^{(1)}\right) \tag{35}$$

where we have defined $g(u) = h(u(1-q)) - uh(1-q)$. Here, $u_2^{(1)}$ denotes the revised value of u_2 following the observation of cell 1. In general, we use $u_i^{(j)}$ to denote the revised value of u_i after the observation that cell $j<i$ did not spike. Therefore, for a population of $N=3$ cells it has the form

$$I_3 = g(u_1) + (1-u_1(1-q))\left[g\left(u_2^{(1)}\right) + \left(1-u_2^{(1)}(1-q)\right)g\left(u_3^{(2)}\right)\right]. \tag{36}$$

Generalizing this for N cells, the information is

$$I_N = g(u_1) + (1-u_1(1-q))\left[g\left(u_2^{(1)}\right) + \ldots\left(1-u_{N-1}^{(N-2)}(1-q)\right)g\left(u_N^{(N-1)}\right)\ldots\right]. \tag{37}$$

The revised values of $u_i^{(j)}$ for $i=2,\ldots N$ and $j=1,\ldots i-1$ follow based on readjusting the thresholds depending on the observation of cells $1,\ldots N-1$ one at a time. For example, following the observation that cell 1 did not spike, the effective values of $u_2, u_3, \ldots u_N$ are revised to

$$u_i^{(1)} = \frac{u_i - u_1(1-q)}{1-u_1(1-q)}, \qquad \text{for } i = 2,\ldots N. \tag{38}$$

Following the observation that cell 2 did not spike, $u_3^{(1)}, u_4^{(1)} \ldots u_N^{(1)}$ are further revised to

$$u_i^{(2)} = \frac{u_i^{(1)} - u_2^{(1)}(1-q)}{1-u_2^{(1)}(1-q)}, \qquad \text{for } i = 3,\ldots N. \tag{39}$$

This process continues, until the observation of cell $N-1$ with the final set of $u_N^{(N-2)}$ being revised to

$$u_N^{(N-1)} = \frac{u_N^{(N-2)} - u_{N-1}^{(N-2)}(1-q)}{1 - u_{N-1}^{(N-2)}(1-q)}. \tag{40}$$

We maximize the information in Eq. 37 with respect to each $u_i^{(j)}$. We can do this sequentially: first maximize I with respect to $u_N^{(N-1)}$, which results in maximizing $g\left(u_N^{(N-1)}\right)$. The maximum is obtained at

$$u_N^{(N-1)} = \frac{1}{(1-q) + q^{-q/(1-q)}} \tag{41}$$

yielding a maximal value of

$$\log\left(1 + (1-q)q^{q/(1-q)}\right). \tag{42}$$

Next, we maximize I with respect to $u_{N-1}^{(N-2)}$, which results in maximizing $g\left(u_{N-1}^{(N-2)}\right) + \left(1 - u_{N-1}^{(N-2)}(1-q)\right)\log\left(1 + (1-q)q^{-q/(1-q)}\right)$. The maximum is obtained at

$$u_{N-1}^{(N-2)} = \frac{1}{2(1-q) + q^{-q/(1-q)}} \tag{43}$$

yielding a maximal value of

$$\log\left(1 + 2(1-q)q^{q/(1-q)}\right). \tag{44}$$

Finally, we maximize I with respect to u_1, which results in maximizing $g(u_1) + (1 - u_1(1-q))\log\left(1 + (N-1)(1-q)q^{-q/(1-q)}\right)$. The maximum is obtained at

$$u_1 = \frac{1}{N(1-q) + q^{-q/(1-q)}} \tag{45}$$

yielding a maximal value of the mutual information as in Eq. 1 in the Results section

$$I = \log\left(1 + N(1-q)q^{q/(1-q)}\right). \tag{46}$$

Based on these derivations we can obtain the sequence of

$$u_i = \frac{1 + (i-1)(1-q)}{N(1-q) + q^{-q/(1-q)}}, \qquad \text{for } i = 1, \ldots N \tag{47}$$

where the difference between consecutive thresholds is given by Eq. 2

$$p = u_{i+1} - u_i = \frac{1-q}{N(1-q) + q^{-q/(1-q)}}, \qquad \text{for } i = 1, \ldots N-1 \tag{48}$$

and the 'edge' threshold is Eq. 3

$$p_{\text{edge}} = u_1 = \frac{1}{N(1-q) + q^{-q/(1-q)}}. \tag{49}$$

Thresholds when optimizing the mutual information: a mixed population

With the Equal Coding Theorem we showed that the information for any ON/OFF mixture is the same (Eq. 46). Next, we show how to derive the thresholds for a mixed population since we know that it will have the same mutual information as the homogeneous population. We do this by swapping $N-m$ of the ON cells into OFF cells, knowing that the thresholds of the ON cells in the new mixed population remain the same, and derive the thresholds for the swapped OFF cells. This means that we need to derive a new set of $u_{m+1}^{\text{mix}}, \ldots, u_N^{\text{mix}}$ for the

OFF population, while keeping $u_1, \dots u_m$ the same for the ON population. To do this, recall that the thresholds for the OFF cells follow different update rules every time an ON cell is observed (see Eq. 33). In particular,

$$u_i^{(k)} = \frac{u_i^{(k-1)}}{1 - u_k^{(k-1)}(1-q)}, \qquad \text{for } i = m+1, \dots N. \tag{50}$$

Additionally, following the observation of OFF cell k (where $k = m+1, \dots, N-1$)

$$u_i^{(k)} = \frac{u_i^{(k-1)} - u_k^{(k-1)}(1-q)}{1 - u_k^{(k-1)}(1-q)}, \qquad \text{for } i = k+1, \dots N. \tag{51}$$

Using these recursions and the values $u_i^{(j)}$ for the ON cells derived previously (Eq. 41 – 45) one can recover the thresholds:

$$u_i = \frac{1 + (i-1)(1-q)}{N(1-q) + q^{-q/(1-q)}}, \qquad \text{for } i = 1, \dots m \tag{52}$$

for the ON cells and

$$u_i^{\text{mix}} = \frac{1 + (m-i+1)(1-q)}{N(1-q) + q^{-q/(1-q)}}, \qquad \text{for } i = m+1, \dots N \tag{53}$$

for the OFF cells (in the mixed population case) where the difference between consecutive thresholds (except between the smallest ON and the largest OFF) is given by Eq. 48 and the 'edge' thresholds by Eq. 49. From here we can derive the *'silent'* interval between the smallest ON and the largest OFF that separates the ON and OFF thresholds, $p_0 = 1 - (N-2)p - 2p_{\text{edge}}$.

Mean firing rate when optimizing the mutual information

Given the optimal thresholds, the mean firing rate per neuron in a population with m ON cells is:

$$r = R\left[p_{\text{edge}} + \frac{p}{2}\left(\frac{m^2 + (N-m)^2}{N} - 1\right)\right] \tag{54}$$

In the large population regime, with $\alpha = m/N$ the fraction of ON cells, the mean firing rate per neuron is

$$r(\alpha) = \frac{R}{2}\left(\alpha^2 + (1-\alpha)^2\right), \tag{55}$$

however, in the high noise regime this becomes independent of α

$$r(\alpha) = \frac{R}{e}. \tag{56}$$

Optimal linear readout without noise

We present here the derivation for the homogeneous population with only ON cells when $R \to \infty$. The linear stimulus estimate of s (Eq. 4) can be written as:

$$y = \sum_{i=1}^{N} w_i \Theta(s - \theta_i) + w_0 \tag{57}$$

where w_i represent the decoding weights and the responses are given by the binary Heaviside functions with thresholds θ_i. Then the mean square error between the original and the estimated stimulus can be written as:

$$E = \langle (y-s)^2 \rangle. \tag{58}$$

In the case of the homogeneous population, we can emulate the constant term w_0 as the weight of an additional neuron with threshold $\theta_0 = -\infty$. Then

$$C_i = \langle \Theta(s - \theta_i) \rangle \quad \text{and} \quad U_i = \langle s\Theta(s - \theta_i) \rangle \tag{59}$$

so the error can be written as:

$$E = w^T C w - 2U^T w + \langle s^2 \rangle \tag{60}$$

where since $\langle \Theta(s - \theta_i)\Theta(s - \theta_j) \rangle = \langle \Theta(s - \max(\theta_i, \theta_j)) \rangle$, for $i \geq j$, we can write: $C_{ij} = C_i$. Optimizing with respect to the weight will gives us

$$w = C^{-1} U \tag{61}$$

which we can rewrite as (Eq. 6 in the Results section):

$$\sum_{j \leq i} w_j = \frac{\int_{\theta_i}^{\theta_{i+1}} \mathrm{d}s\, s\, p(s)}{\int_{\theta_i}^{\theta_{i+1}} \mathrm{d}s\, p(s)} = \langle s \rangle_i,\ 0 \leq i \leq N \tag{62}$$

with $\theta_{N+1} = \infty$ and (Eq. 5 in the Results section):

$$w_i = \langle s \rangle_i - \langle s \rangle_{i-1},\ i = 1, ..., N \quad \text{and} \quad w_0 = \langle s \rangle_0. \tag{63}$$

Optimizing with respect to the thresholds:

$$\sum_{j \leq i} w_j = \theta_i + \frac{w_i}{2} \tag{64}$$

which gives

$$\theta_i - \theta_{i-1} = \frac{1}{2}(w_i - w_{i-1}) \tag{65}$$

and from this we can derive (Eq. 7 in the Results section):

$$\theta_i = \frac{1}{2}\left(\langle s \rangle_i + \langle s \rangle_{i-1}\right),\ i = 1, ..., N. \tag{66}$$

Optimizing with respect to the constant term yields:

$$w_0 = \frac{\int_{-\infty}^{\theta_1} \mathrm{d}s\, s\, p(s)}{\int_{-\infty}^{\theta_1} \mathrm{d}s\, p(s)}. \tag{67}$$

To solve these equations numerically, we implement an iterative procedure that rapidly converges to the optimal solution: starting from an ansatz for the thresholds, we compute $\langle s \rangle_i$ and obtain w_i, which is used to derive the new set of thresholds.

In the case of the mixed population with ON and OFF cells, the optimal solution is one where the ON and OFF responses do not overlap; thus, there is no correlation between them. Therefore, we can treat each subpopulation separately, and in identical manner to the purely homogeneous case. The optimal weights and thresholds are identical to the homogeneous population population, with the exception of the constant term:

$$w_0^m = \frac{\int_{\theta_1{}^{\mathrm{OFF}}}^{\theta_1^{\mathrm{ON}}} \mathrm{d}s\, s\, p(s)}{\int_{\theta_1{}^{\mathrm{OFF}}}^{\theta_1^{\mathrm{ON}}} \mathrm{d}s\, p(s)} \tag{68}$$

where ${\theta_1}^{\mathrm{OFF}}$ denotes the largest OFF threshold and ${\theta_1}^{\mathrm{ON}}$ denotes the smallest ON threshold in the population.

Thresholds when optimizing the linear readout without noise

Now we consider the case of large N (for any mixture of ON and OFF cells) to derive the thresholds in the asymptotic limit where the threshold intervals (differences between neighboring thresholds) are small. We use a first order expansion of the stimulus distribution $p(s)$ around each threshold θ_j in the expressions for $\langle s \rangle_j$.

$$\langle s \rangle_j = \frac{\int_{\theta_j}^{\theta_{j+1}} \mathrm{d}s\, s\, p(s)}{\int_{\theta_j}^{\theta_{j+1}} \mathrm{d}s\, p(s)} = \theta_j + \frac{\int_{\theta_j}^{\theta_{j+1}} \mathrm{d}s (s-\theta_j) p(s)}{\int_{\theta_j}^{\theta_{j+1}} \mathrm{d}s\, p(s)} \approx \theta_j + \frac{\theta_{j+1}-\theta_j}{2} + \frac{p'(\theta_j)}{12 p(\theta_j)} (\theta_{j+1}-\theta_j)^2 \tag{69}$$

and similarly,

$$\langle s \rangle_{j-1} \approx \theta_j + \frac{\theta_{j-1}-\theta_j}{2} + \frac{p'(\theta_j)}{12 p(\theta_j)} (\theta_{j-1}-\theta_j)^2. \tag{70}$$

Combining Eq. 69 and Eq. 70 into Eq. 66, yields

$$\frac{2\theta_j - \theta_{j+1} - \theta_{j-1}}{4} = \frac{p'(\theta_j)}{24 p(\theta_j)} \left[(\theta_{j+1}-\theta_j)^2 + (\theta_{j-1}-\theta_j)^2 \right]. \tag{71}$$

Taking the continuous limit so that j maps onto x with $j=1$ corresponding to $x=0$, $j=N$ corresponding to $x=1$, and $\mathrm{d}x = 1/N$, we can write

$$\theta_{j+1} - \theta_j = \mathrm{d}x\, \theta'(x) \tag{72}$$

$$\theta_{j+1} - 2\theta_j + \theta_{j-1} = (\mathrm{d}x)^2\, \theta''(x) \tag{73}$$

turning Eq. 71 into:

$$\theta{"}(x) = g(x)\, (\theta'(x))^2. \tag{74}$$

We can further define:

$$g(x) = -\frac{dp(\theta(x))/\mathrm{d}\theta}{3p(\theta(x))} = \frac{G(x)}{\theta'(x)} \text{ where } G(x) = -\frac{d}{3\mathrm{d}x} \log p(\theta(x)). \tag{75}$$

Denoting $y(x) = \theta'(x)$ gives the differential equation

$$y'(x) = G(x)\, y(x) \tag{76}$$

which has the solution

$$\log y = \int^x \mathrm{d}u\, G(u) + c \tag{77}$$

and consequently we obtain the differential equation

$$\theta'(x) = \frac{c}{p(\theta(x))^{1/3}} \tag{78}$$

where c is a constant. This can be inverted into

$$x(\theta) = c \int_{-\infty}^{\theta} p(\theta')^{1/3} \mathrm{d}\theta' + c' \tag{79}$$

where $x = i/N$ is the threshold index. We can determine the constants c and c' from the boundary conditions:

$$x(-\infty) = 0, \; x(\infty) = 1 \tag{80}$$

such that (as Eq. 8 in the Results section),

$$x(\theta) = Z \int_{-\infty}^{\theta} p(\theta')^{1/3} \mathrm{d}\theta', \qquad Z^{-1} = \int_{-\infty}^{\infty} p(\theta')^{1/3} \mathrm{d}\theta'. \tag{81}$$

Inverting this relationship, we can obtain the threshold distribution $\theta(x)$ as a function of the index $x = i/N$. An expression for the optimal thresholds for the Laplace distribution: $p(s) = 1/2 e^{-|s|}$ is provided in Eq. 9 in the Results section.

Optimal linear readout with noise: homogeneous population

For convenience, we normalize the linear readout

$$y = \frac{1}{R}\sum_i w_i n_i + w_0. \tag{82}$$

The error can be written as before (Eq. 60) with different correlations

$$C_{ij} = \frac{1}{R}\langle\langle n_i\rangle_n \langle n_j\rangle_n\rangle + \frac{1}{R^2}\delta_{ij}\langle\langle n_i\rangle_n\rangle = \langle\Theta(s-\theta_i)\Theta(s-\theta_j)\rangle + \frac{1}{R}\delta_{ij}\langle\Theta(s-\theta_i)\rangle \tag{83}$$

If we define, as before:

$$C_i = \langle\Theta(s-\theta_i)\rangle \tag{84}$$

then for $\langle\Theta(s-\theta_i)\Theta(s-\theta_j)\rangle = \langle\Theta(s-\max(\theta_i,\theta_j)\rangle$, and for $i \geq j$:

$$C_{ij} = C_i + \frac{1}{R}\delta_{ij}C_i \tag{85}$$

and

$$U_i = \langle s\Theta(s-\theta_i)\rangle - w_0\langle\Theta(s-\theta_i)\rangle. \tag{86}$$

Optimizing with respect to the weights:

$$w = C^{-1}U \tag{87}$$

and

$$w_0 = \langle s\rangle - \sum_{i=1}^{N} w_i\langle\Theta(s-\theta_i)\rangle. \tag{88}$$

Optimizing with respect to the thresholds:

$$\theta_i = w_0 + \sum_{j\leq i} w_j - \frac{w_i}{2}\left(1 - R^{-1}\right), \quad i = 1, ..., N. \tag{89}$$

To solve these equations numerically, we implement an iterative procedure that rapidly converges to the optimal solution: starting from an ansatz for the thresholds, we compute C and U and obtain w from Eq. 87, which is used to derive the new set of thresholds.

Thresholds when optimizing the linear readout with noise: homogeneous population

We provide an expression for the optimal thresholds for the general Laplace distribution:

$$p(s) = \begin{cases} A_+ e^{-s/\tau_+}, & s \geq 0, \\ A_- e^{s/\tau_-}, & s < 0. \end{cases} \tag{90}$$

The symmetric Laplace distribution is one example, $p(s) = 1/2e^{-|s|}$, with $A_+ = A_- = 1/2$ and $\tau_+ = \tau_- = 1$. In the limit of large population size N, we again derive the thresholds in the asymptotic limit where the threshold intervals are small. Assuming $\theta_1 < 0$,

$$x(\theta) = \frac{1}{\int_{\theta_1}^{0} \mathrm{d}u\,\sqrt{1 - A_-\tau_-\, e^{u/\tau_-}} + 2\tau_+\sqrt{A_+\tau_+}} \begin{cases} \int_{\theta_1}^{\theta} \mathrm{d}u\,\sqrt{1 - A_-\tau_-\, e^{u/\tau_-}}, & \theta \leq 0 \\ \int_{\theta_1}^{0} \mathrm{d}u\,\sqrt{1 - A_-\tau_-\, e^{u/\tau_-}} + 2\tau_+\sqrt{A_+\tau_+}(1 - e^{-\theta/2\tau_+}), & \theta > 0 \end{cases} \tag{91}$$

and assuming $|\theta_1|$ is large so that $\int_{\theta_1}^{\theta} \mathrm{d}u \sqrt{1 - A_- \tau_- e^{u/\tau_-}} \approx \theta - \theta_1$, we can approximate

$$x(\theta) \approx \frac{1}{-\theta_1 + \tau_+ \sqrt{A_+ \tau_+}} \begin{cases} \theta - \theta_1, & \theta \leq 0 \\ -\theta_1 + 2\tau_+ \sqrt{A_+ \tau_+}(1 - e^{-\theta/2\tau_+}), & \theta > 0 \end{cases} \tag{92}$$

inverting this relationship, the optimal thresholds are:

$$\theta(x) \approx \begin{cases} \theta_1 + (-\theta_1 + 2\tau_+ \sqrt{A_+ \tau_+})x, & 0 \leq x \leq \frac{1}{1 - 2\tau_+ \sqrt{A_+ \tau_+}/\theta_1}, \\ -2\tau_+ \log\left[\left(1 - \frac{\theta_1}{2\tau_+ \sqrt{A_+ \tau_+}}\right)(1 - x)\right], & \frac{1}{1 - 2\tau_+ \sqrt{A_+ \tau_+}/\theta_1} \leq x \leq 1. \end{cases} \tag{93}$$

To fully determine the optimal thresholds, this requires knowledge of the first threshold, θ_1. In the asymptotic limit, where the thresholds θ_i are close to each other, again expanding $p(s)$ around each threshold, we derive

$$\theta_1 \approx \tau_- \log\left(\frac{\log(RNA_- \tau_-)}{RNA_- \tau_-^2}\right). \tag{94}$$

Optimal linear readout with noise: mixed ON-OFF population

So far we have not explicitly treated the ON and OFF populations separately, because both when maximizing the mutual information for all noise levels, and minimizing the MSE in the limit of no noise, the performance and optimal thresholds were the same for all populations independent of the ON/OFF mixture. Now, we must treat the two populations separately.

Assume we have m ON cells and $N - m$ OFF cells. We order the thresholds in the following manner (since non-overlapping ON and OFF cells is the optimal solution),

$$\theta_{N-m}^{\mathrm{OFF}} \leq \theta_{N-m-1}^{\mathrm{OFF}} \leq ... \leq \theta_1^{\mathrm{OFF}} \leq \theta_1^{\mathrm{ON}} \leq \theta_2^{\mathrm{ON}} \leq ... \leq \theta_{m-1}^{\mathrm{ON}} \leq \theta_m^{\mathrm{ON}} \tag{95}$$

so that we can proceed in the same manner for each subpopulation as for the homogeneous population. The readout can be written as

$$y = \frac{1}{R} \sum_{i=1}^{N-m} w_i^{\mathrm{OFF}} n_i^{\mathrm{OFF}} + \frac{1}{R} \sum_{i=1}^{m} w_i^{\mathrm{ON}} n_i^{\mathrm{ON}} + w_0. \tag{96}$$

The error then is (assuming the optimal ON and OFF thresholds do not overlap – so that the ON-OFF cross-correlation is zero):

$$E = \langle (y - s)^2 \rangle = (w^{\mathrm{ON}})^T C^{\mathrm{ON}} w^{\mathrm{ON}} + (w^{\mathrm{OFF}})^T C^{\mathrm{OFF}} w^{\mathrm{OFF}} - 2(w^{\mathrm{ON}})^T U^{\mathrm{ON}} - 2(w^{\mathrm{OFF}})^T U^{\mathrm{OFF}} + \langle (s - w_0)^2 \rangle \tag{97}$$

$$C_{ij}^{\mathrm{ON}} = C_i^{\mathrm{ON}} \text{ and } C_{ij}^{\mathrm{OFF}} = C_i^{\mathrm{OFF}} \text{ for } i \geq j, \tag{98}$$

and

$$C_i^{\mathrm{ON}} = \langle \Theta(s - \theta_i^{\mathrm{ON}}) \rangle + \frac{1}{R} \delta_{ij} \langle \Theta(s - \theta_i^{\mathrm{ON}}) \rangle \tag{99}$$

$$C_i^{\mathrm{OFF}} = \langle \Theta(\theta_i^{\mathrm{OFF}} - s) \rangle + \frac{1}{R} \delta_{ij} \langle \Theta(\theta_i^{\mathrm{OFF}} - s) \rangle, \tag{100}$$

$$U_i^{\mathrm{ON}} = \langle s\Theta(s - \theta_i^{\mathrm{ON}}) \rangle - w_0 \langle \Theta(s - \theta_i^{\mathrm{ON}}) \rangle \tag{101}$$

$$U_i^{\mathrm{OFF}} = \langle s\Theta(\theta_i^{\mathrm{OFF}} - s) \rangle - w_0 \langle \Theta(\theta_i^{\mathrm{OFF}} - s) \rangle \tag{102}$$

Optimizing with respect to the weights we get very similar expressions for each subpopulation (ON and OFF) as for the homogeneous population:

$$w^{\mathrm{ON}} = (C^{\mathrm{ON}})^{-1} U^{\mathrm{ON}} \quad \text{and} \quad w^{\mathrm{OFF}} = (C^{\mathrm{OFF}})^{-1} U^{\mathrm{OFF}} \tag{103}$$

and optimizing the thresholds:

$$\sum_{j\le i} w_j^{\mathrm{ON}} = \theta_i^{\mathrm{ON}} - w_0 + \frac{w_i^{\mathrm{ON}}}{2}(1-R^{-1}),\ i = 1, ..., m \tag{104}$$

for the ON cells, and similarly for the OFF:

$$\sum_{j\le i} w_j^{\mathrm{OFF}} = \theta_i^{\mathrm{OFF}} - w_0 + \frac{w_i^{\mathrm{OFF}}}{2}(1-R^{-1}),\ i = 1, ..., N-m. \tag{105}$$

The difference from the homogeneous population is in the constant term:

$$w_0^m = \langle s\rangle - \sum_{i=1}^{m} w_i^{\mathrm{ON}} C_i^{\mathrm{ON}} - \sum_{j=1}^{N-m} w_j^{\mathrm{OFF}} C_j^{\mathrm{OFF}}. \tag{106}$$

Thresholds when optimizing the linear readout with noise: mixed ON-OFF population

We proceed in a similar fashion as with the homogeneous population to obtain the approximation in the case of large N: Let $f^{\mathrm{ON}} = m/N$ be the fraction of ON cells and $f^{\mathrm{OFF}} = (N-m)/N$ be the fraction of OFF cells in the population. We remap the thresholds, so that in the continuum limit $\theta_{N-m}^{\mathrm{OFF}} \le \theta_{N-m-1}^{\mathrm{OFF}} \le ... \le \theta_1^{\mathrm{OFF}}$ becomes $\theta^{\mathrm{OFF}}(x^{\mathrm{OFF}})$ and $\theta_1^{\mathrm{ON}} \le \theta_2^{\mathrm{ON}} \le ... \le \theta_{m-1}^{\mathrm{ON}} \le \theta_m^{\mathrm{ON}}$ becomes $\theta^{\mathrm{ON}}(x^{\mathrm{ON}})$. Thus, the threshold index $x = i/N \in [0,1]$ for the homogeneous population becomes $x^{\mathrm{ON}} = i/m \in [0, f^{\mathrm{ON}}]$ and $i = 1, 2, \ldots, m$ being the indices of the ON cels, and $x^{\mathrm{OFF}} = i/(N-m) \in [0, f^{\mathrm{OFF}}]$ and $i = 1, 2, \ldots, N-m$ being the indices of the OFF cells. Figure 8 illustrates the mapping.

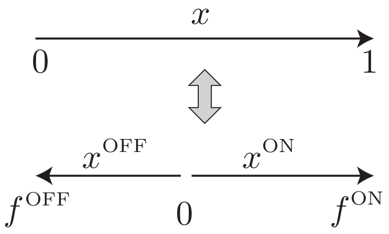

Figure 8. The mapping of the threshold indices from the homogeneous population with only ON cells to the mixed population with ON and OFF cells.

We provide an expression for the optimal thresholds for the general Laplace distribution (Eq. 90), and for a population that is unbalanced and has more ON cells, $f^{\mathrm{ON}} > f^{\mathrm{OFF}}$. If $\theta_1^{\mathrm{OFF}} \le \theta_1^{\mathrm{ON}} \le 0$, for the ON thresholds and weights the solution is similar to the case of the homogeneous population, i.e.

$$x^{\mathrm{ON}}(\theta^{\mathrm{ON}}) = \frac{f^{\mathrm{ON}}}{\int_{\theta_1^{\mathrm{ON}}}^{0} \mathrm{d}u\, \sqrt{1 - A_-\tau_-\, e^{u/\tau_-}} + 2\tau_+\sqrt{A_+\tau_+}} \int_{\theta_1^{\mathrm{ON}}}^{\theta^{\mathrm{ON}}} \mathrm{d}u\, \sqrt{1 - A_-\tau_-\, e^{u/\tau_-}}, \quad \theta^{ON} \le 0 \tag{107}$$

and

$$x^{\mathrm{ON}}(\theta^{\mathrm{ON}}) = f^{\mathrm{ON}} \left[1 - \frac{2\tau_+\sqrt{A_+\tau_+} e^{-\theta^{\mathrm{ON}}/2\tau_+}}{\int_{\theta_1^{\mathrm{ON}}}^{0} \mathrm{d}u\, \sqrt{1 - A_-\tau_-\, e^{u/\tau_-}} + 2\tau_+\sqrt{A_+\tau_+}} \right], \quad \theta^{\mathrm{ON}} > 0. \tag{108}$$

These expressions have to be inverted to obtain $\theta^{\mathrm{ON}}(x^{\mathrm{ON}})$, which has to be done numerically. We proceed very similarly for the OFF cells. Namely, if $\theta_1^{\mathrm{OFF}} < 0$ and assuming $|\theta_{N-m}^{\mathrm{OFF}}|$ is large

$$x^{\mathrm{OFF}}(\theta^{\mathrm{OFF}}) = f^{\mathrm{OFF}} \left(1 - e^{(\theta^{\mathrm{OFF}} - \theta_1^{\mathrm{OFF}})/2\tau_-} \right) \tag{109}$$

inverting this relationship is possible analytically

$$\theta^{\mathrm{OFF}}(x^{\mathrm{OFF}}) = \theta_1^{\mathrm{OFF}} + 2\tau_- \log\left(1 - x/f^{\mathrm{OFF}}\right). \tag{110}$$

To fully determine the optimal thresholds, this requires knowledge of the first ON and OFF thresholds, θ_1^{ON} and θ_1^{OFF}.

When the population is mixed so that neither population dominates, the first ON and OFF thresholds are order 1. Assuming that they are close in stimulus space, so that $\theta_1^{\mathrm{ON}} - \theta_1^{\mathrm{OFF}} \ll 1$, we can use the equations from optimizing the thresholds and weights to obtain the following equation which can be solved for $\phi = \theta_1^{\mathrm{OFF}}$:

$$X(\phi) = \frac{2\tau_- \sqrt{1 - A - \tau_- e^{-\phi/\tau_-}}}{\int_{\theta_1^{\mathrm{ON}}}^{0} \mathrm{d}u \sqrt{1 - A_-\tau_- e^{u/\tau_-}} + 2\tau_+ \sqrt{A_+\tau_+}} \left(\frac{1}{A_-\tau_-} e^{\phi/\tau_-} - 1\right) \tag{111}$$

For the symmetric Laplace distribution with $A_+ = A_- = 1/2$ and $\tau_+ = \tau_- = 1$, the equation to solve for ϕ reduces to (Fig. 9):

$$X(\phi) = \frac{2\sqrt{1 - \frac{1}{2}e^{-\phi}}}{\sqrt{2} + \int_{-\phi}^{0} \mathrm{d}u \sqrt{1 - \frac{1}{2}e^{u}}} \left(2e^{\phi} - 1\right). \tag{112}$$

As shown in Figure 9, when there is an equal number of ON and OFF cells, $X = 1$ and $\theta_1^{\mathrm{ON}} \approx \theta_1^{\mathrm{OFF}} \approx 0$. If there are 20% OFF cells and 80% ON cells in the population, then $X = (1/5)/(4/5) = 1/4$, and the first thresholds of each subpopulation are $\theta_1^{\mathrm{ON}} \approx \theta_1^{\mathrm{OFF}} = -0.79$. In the Results section we also considered asymmetric stimulus distributions where we varied the negative-to-positive bias τ_-/τ_+ and derived the solutions in a similar manner (Fig. 5).

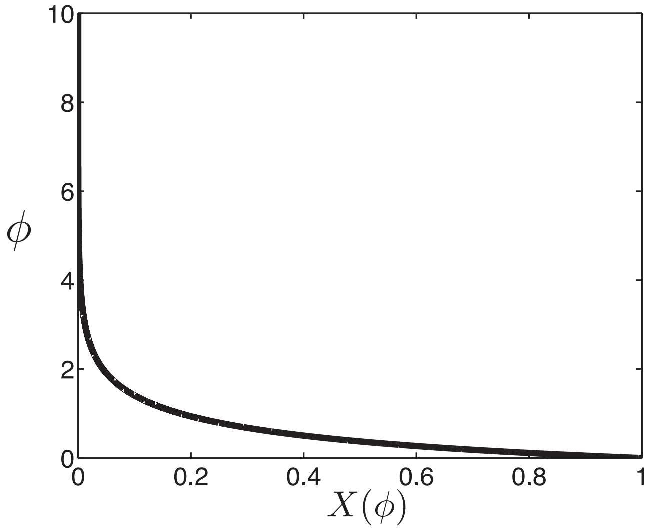

Figure 9. Determining the first thresholds for a mixed population of ON and OFF cells, $\phi = |\theta_1^{\mathrm{ON}}| \approx |\theta_1^{\mathrm{OFF}}|$ as a function of $X = f^{\mathrm{OFF}}/f^{\mathrm{ON}}$. For a symmetric Laplace distribution $p(s) = 1/2e^{-|s|}$.

Deriving the stimulus distribution from measured ORN thresholds

From the study of Si and colleagues we extracted the distribution of measured thresholds (referred to as EC_{50} values) [37]. The cumulative distribution of the inverse of thresholds is

$$P_\theta\left(X > \frac{1}{\theta}\right) \propto \left(\frac{1}{\theta}\right)^{-\lambda} \tag{113}$$

where $\lambda = 0.42$ (Fig. 6B). This enables us to derive the probability density function of the inverse of thresholds (Fig. 6C)

$$p_\theta\left(\frac{1}{\theta}\right) \propto \left(\frac{1}{\theta}\right)^{-\lambda-1}. \tag{114}$$

This distribution has a cut-off of $\theta_c = 4.22 \cdot 10^4$ as reported in [37]. From this, we can derive the distribution of measured thresholds

$$p_\theta(\theta) = \frac{1}{\theta^2} p\left(\frac{1}{\theta}\right) \quad \text{such that} \quad p_\theta(\theta) \propto \theta^{-\lambda+1}. \tag{115}$$

Next, we assume that these measured thresholds implement an optimal code first under the infomax criterion. Now, using the equation for the cumulative distribution of optimal thresholds in the large population limit, $x(\theta) = \int_{-\infty}^{\theta} p_c(z)\,\mathrm{d}z$, we can derive the stimulus distribution of odorant concentrations, p_c,

$$p_c(C) \propto C^{-\lambda+1} = C^{-0.58}. \tag{116}$$

However, if we assume that these measured thresholds implement an optimal code under the criterion of minimizing the mean squared error of the optimal linear decoder, $x(\theta) = \int_{-\infty}^{\theta} p_c^{1/3}(z)\,\mathrm{d}z$, then the stimulus distribution of odorant concentrations, p_c, is

$$p_c(C) \propto C^{3(-\lambda+1)} = C^{-1.74}. \tag{117}$$

These are both shown in Fig. 6D.

Supporting information

S1 Figure. Binary neurons with spontaneous firing rate and Poisson noise. A framework with binary neurons that have two firing rate levels, r if the stimulus is smaller (bigger) than a threshold, and R if the stimulus is bigger (smaller) than a threshold for ON (OFF) cells. We compare two systems, left: one ON (red) and one OFF (blue) cells, and right, two ON cells, where the information is maximized by optimizing the cells thresholds.

S2 Figure. Sigmoidal neurons with sub-Poisson experimentally measured noise. Two sigmoidal nonlinearities for an ON cell (red) and an OFF cell (blue), describing the firing rate as a function of stimulus with the maximum expected spike count R, the gain β, and the threshold θ. The shaded curve denotes the Laplace stimulus probability distribution.

S1 Text. Mutual Information for a system with two cells. The mutual information for different noise models.

S1 Table. Conditional probability matrix. Conditional probability matrix $p(k_1, k_2|s)$ for a mixed ON-OFF system.

S2 Table. Conditional probability matrix. Conditional probability matrix $p(k_1, k_2|s)$ for a homogeneous ON-ON system.

S3 Table. Mutual information for a two-cell system. Mutual information for a two-cell system with spontaneous firing rate and Poisson noise.

S4 Table. Mutual information for a two-cell system. Mutual information for a two-cell system with empirically measured sub-Poisson noise from salamander retinal ganglion cells.

Acknowledgments

All authors were supported by the NIH, the Gatsby Charitable Foundation and the Swartz Foundation. JG was supported by the Max Planck Society and a Burroughs-Wellcome Career Award at the Scientific Interface. JG thanks Shuai Shao for careful reading of the analytical calculations.

References

1. Barlow HB. Possible principles underlying the transformations of sensory messages. In: Sensory Communication. MIT Press; 1961. p. 217–234.

2. Atick JJ, Redlich AN. Towards a theory of early visual processing. Neural Comput. 1990;2:308–320.

3. Atick JJ, Redlich AN. What does the retina know about natural scenes? Neural Comput. 1992;4:196–210.

4. Laughlin SA. Simple coding procedure enhances a neuron's information capacity. Z Naturforsch C. 1981;36:910–912.

5. van Hateren JH. Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. J Comp Physiol A. 1992;171:157–170.

6. Smith EC, Lewicki MS. Efficient auditory coding. Nature. 2006;439:978–982.

7. Simoncelli EP, Olshausen BA. Natural image statistics and neural representation. Annu Rev Neurosci. 2001;24:1193–1216.

8. Kuffler SW. Discharge patterns and functional organization of mammalian retina. J Neurophysiol. 1953;16:37–68.

9. Joesch M, Schnell B, Raghu SV, Reiff DF, Borst A. ON and OFF pathways in *Drosophila* motion vision. Nature. 2010;468:300–304.

10. Gallio M, Ofstad TA, Macpherson LJ, Wang JW, Zuker CS. The coding of temperature in the *Drosophila* brain. Cell. 2011;144:614–624.

11. Chalasani SH, Chronis N, Tsunozaki M, Gray JM, Ramot D, Goodman MB, et al. Dissecting a circuit for olfactory behaviour in *Caenorhabditis elegans*. Nature. 2007;450:63–70.

12. Tsunozaki M, Bautista DM. Mammalian somatosensory mechanotransduction. Curr Opin Neurobiol. 2009;19:362–369.

13. Bell CC. Mormyromast electroreceptor organs and their afferent fibers in mormyrid fish. III. Physiological differences between two morphological types of fibers. J Neurophysiol. 1990;63:319–332.

14. Kastner DB, Baccus SA. Coordinated dynamic encoding in the retina using opposing forms of plasticity. Nat Neurosci. 2011;14:1317–1322.

15. Hodson-Tole EF, Wakeling JM. Motor unit recruitment for dynamic tasks: current understanding and future directions. J Comp Physiol B. 2009;179:57–66.

16. Schiller PH. The ON and OFF channels of the visual system. Trends Neurosci. 1992;15:86–92.

17. Gjorgjieva J, Sompolinsky H, Meister M. Benefits of pathway splitting in sensory coding. J Neurosci. 2014;34:12127–12144.

18. Kastner DB, Baccus SA, Sharpee TO. Critical and maximally informative encoding between neural populations in the retina. Proc Natl Acad Sci USA. 2015;112:2533–2538.

19. Brinkman BAW, Weber AI, Rieke F, Shea-Brown E. How do efficient coding strategies depend on origins of noise in neural circuits. PLoS Comp Biol. 2016;12:e1005150.

20. Ratliff CP, Borghuis BG, Kao YH, Sterling P, Balasubramanian V. Retina is structured to process an excess of darkness in natural scenes. Proc Natl Sci USA. 2010;107:17368–17373.

21. Doi E, J LG, Field GD, Shlens J, Sher A, Greschner M, et al. Efficient coding of spatial information in the primate retina. J Neurosci. 2012;32:16256–16264.

22. Seung HS, Sompolinsky H. Simple models for reading neuronal population codes. Proc Natl Acad Sci USA. 1993;90:10749–10753.

23. Bell AJ, Sejnowski TJ. The "Independent Components" of natural scenes are edge filters. Vision Res. 1997;37:3327–3338.

24. Brunel N, Nadal JP. Mutual information, Fisher information, and population coding. Neural Comput. 1998;10:1731–1757.

25. Wang Z, Stocker AA, Lee DD. Efficient neural codes that minimize L_p reconstruction error. Neural Comput. 2016;28:2656–2686.

26. Warland DK, Reinagel P, Meister M. Decoding visual information from a population of retinal ganglion cells. J Neurophysiol. 1997;78:2336–2350.

27. Bethge M, Rotermund D, Pawelzik K. Optimal neural rate coding leads to bimodal firing rate distributions. Network: Comput Neur Syst. 2003;14:303–319.

28. Pitkow X, Meister M. Decorrelation and efficient coding by retinal ganglion cells. Nat Neurosci. 2012;15:628–635.

29. Rieke F, Warland D, de Ruyter van Steveninck RR, Bialek W. Spikes: Exploring the neural code. Cambridge, MA: MIT Press; 1997.

30. Bialek W, de Ruyter van Steveninck RR, Warland D. Reading a neural code. Science. 1991;252:1854–1857.

31. Lumpkin EA, Caterina MJ. Mechanisms of sensory transduction in the skin. Nature. 2007;445:858–865.

32. Dhaka A, Viswanath V, Patapoutian A. TRP ion channels and temperature sensation. Annu Rev Neurosci. 2006;29:135–161.

33. Romo R, Brody CD, Hernández A, Lemus L. Neuronal correlates of parametric working memory in the prefrontal cortex. Nature. 1998;399:470–473.

34. Salinas E, Hernández A, Zainos A, Romo R. Periodicity and firing rate as candidate neural codes for the frequency of vibrotactile stimuli. J Neurosci. 2000;20:5503–5515.

35. Hallem EA, Carlson JR. Coding of odors by a receptor repertoire. Cell. 2006;125:143–160.

36. Stevens CF. A statistical property of fly odor responses is conserved across odors. Proc Natl Acad Sci. 2016;113:6737–6742.

37. Si G, Kanwal JK, Hu Y, Tabone CJ, Baron J, Berck M, et al. Invariances in a combinatorial olfactory receptor code. Neuron. 2019;in press:https://doi.org/10.1016/j.neuron.2018.12.030.

38. Uzzell VJ, Chichilnisky EJ. Precision of spike trains in primate retinal ganglion cells. J Neurophysiol. 2004;92:780–789.

39. Sharpee TO. Optimizing neural information capacity through discretization. Neuron. 2017;94:954–960.

40. Balasubramanian V, Kimber D, II MJB. Metabolically efficient information processing. Neural Computation. 2001;13:799–815.

41. Stein RB. The information capacity of nerve cells using a frequency code. Biophys J. 1967;7:797–826.

42. Shamai S. Capacity of a pulse amplitude modulated direct detection photon channel. IEE Proc Commun Speech Vis. 1990;137:424–430.

43. Nikitin AP, Stocks NG, McDonnell MD. Neural population coding is optimized by discrete tuning curves. Phys Rev Lett. 2009;103:138101.

44. Sachs MB, Abbas PJ. Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli. J Acoust Soc Am. 1974;56:1835–1847.

45. Lewen GD, Bialek W, de Ruyter van Steveninck RR. Neural coding of naturalistic motion stimuli. Network: Comput Neural Syst. 2001;12:317–329.

46. Strong SP, de Ruyter van Steveninck RR, Bialek W, Koberle R. On the application of information theory to neural spike trains. Pac Symp Biocomput. 1998;1998:621–632.

47. Nadal JP, Parga N. Nonlinear neurons in the low noise limit: a factorial code maximizes information transfer. Network Comput Neural Syst. 1994;5:565–581.

48. Puchalla JL, Schneidman E, Harris RA, Berry MJ. Redundancy in the population code of the retina. Neuron. 2005;46:493–504.

49. Tkacik G, Prentice JS, Balasubramanian V, Schneidman E. Optimal population coding by noisy spiking neurons. Proc Natl Acad Sci USA. 2010;107:14419–14424.

50. Silies M, Gohl DM, Fisher YE, Freifeld L, Clark DA, Clandinin TR. Modular use of peripheral input channels tunes motion-detecting circuitry. Neuron. 2013;79:111–127.

51. Bhattacharya MRC, Bautista DM, Wu K, Haeberle H, Lumpkin EA, , et al. Radial stretch reveals distinct populations of mechanosensitive mammalian somatosensory neurons. Proc Natlc Acad USA. 2008;105:20015–20020.

52. Dayan P, Abbott LF. Theoretical neuroscience: computational and mathematical modelling of neural systems. Cambridge, Massachusetts, London, England: The MIT Press; 2001.

53. Panter PF, Dite W. Quantizing distortion in pulse-count modulation with nonuniform spacing of levels. Proc IRE. 1951;39:44–48.

54. Gray RM, Neuhoff DL. Quantization. IEEE Trans Inf Theory. 1998;44:2325–2383.

55. Field DJ. What is the goal of sensory coding? Neural Comput. 1994;6:559–601.

56. Ruderman DL. The statistics of natural images. Network: Comput Neural Syst. 1994;5:517–548.

57. Dong DW, Atick JJ. Statistics of natural time-varying images. Network: Comput Neural Syst. 1995;6:345–358.

58. van Hateren JH. Processing of natural time series of intensities by the visual system of the blowfly. Vision Res. 1997;37:3407–3416.

59. Tadmor Y, Tolhurst DJ. Calculating the contrasts that retinal ganglion cells and LGN neurones encounter in natural scenes. Vision Res. 2000;40:3145–3157.

60. Singh NC, Theunissen FE. Modulation spectra of natural sounds and ethological theories of auditory processing. J Acoust Soc Am. 2003;114:3394–3411.

61. Geisler WS. Visual perception and the statistical properties of natural scenes. Annu Rev Psychol. 2008;59:167–192.

62. Catrakis HJ, Dimotakis PE. Scale distributions and fractal dimensions in turbulence. Phys Rev Lett. 1996;77:3795.

63. Dekker T, Ibba I, Siju KP, Stensmyr MC, Hansson BS. Olfactory shifts parallel superspecialism for toxic fruit in *Drosophila melanogaster*, sibling, *D. sechellia*. Curr Biol. 2006;16:101–109.

64. Linz J, Baschwitz A, Strutz A, Dweck HKM, Sachse S, Hansson BS, et al. Host plant-driven sensory specialization in *Drosophila erecta*. Proc Royal Soc B. 2013;280:20130626.

65. Sanes JR, Masland RH. The types of retinal ganglion cells: current status and implications for neuronal classification. Annu Rev Neurosci. 2015;38:221–246.

66. Baden T, Berens P, Franke K, Rosón R, Bethge M, Euler T. The functional diversity of retinal ganglion cells in the mouse. Nature. 2016;529:345–350.

67. Tishby N, Pereira F, Bialek W. The information bottleneck method. In: Proceedings 37th Allerton Conference on Communication, Control, and Computing; 1999. p. 368–377.

68. Palmer SE, Marre O, Berry MJ, Bialek W. Predictive information in a sensory population. Proc Natl Acad Sci USA. 2015;112:6908–6913.

69. Mlynarski W, Hermundstad A. Adaptive coding for dynamic sensory inference. BioRxiv. 2018;doi: http://dx.doi.org/10.1101/189506.

70. Oswald MJ, Tantirigama MLS, Sonntag I, Hughes SM, Empson RM. Diversity of layer 5 projection neurons in the mouse motor cortex. Front Cell Neurosci. 2013;7:174:doi: 10.3389/fncel.2013.0017.

71. Risner JR, Holt JR. Heterogeneous potassium conductances contribute to the diverse firing properties of postnatal mouse vestibular ganglion neurons. J Neurophysiol. 2006;96:2364–2376.

72. Tan J, Savigner A, Ma M, Luo M. Odor information processing by the olfactory bulb analyzed in gene-targeted mice. Neuron. 2010; p. 912–926.

73. Tichy H, Hinterwirth A, Gingl E. Olfactory receptors on the cockroach antenna signal odour ON and odour OFF by excitation. Eur J Neurosci. 2005;22:3147–3160.

74. Park IM, Pillow J. Bayesian Efficient Coding. BioRxiv. 2017;doi: http://dx.doi.org/10.1101/178418.

75. Attneave F. Some informational aspects of visual perception. Phychol Rev. 1954;61:183–193.

76. Atick JJ, Redlich AN. Convergent Algorithm for Sensory Receptive Field Development. Neural Comput. 1993;5:45–60.

77. Roska B, Meister M. The Retina Dissects the Visual Scene into Distinct Features. In: In: The New Visual Neurosciences. Cambridge, MA: The MIT Press; 2014. p. 163–182.

78. Pillow JW, Shlens J, Paninski L, Sher A, Litke AM, Chichilnisky EJ, et al. Spatio-temporal correlations and visual signalling in a complete neuronal population. Nature. 2008;454:995–999.

79. Nadal JP, Brunel N. Nonlinear feedforward networks with stochastic output: infomax implies redundancy reduction. Network Comput Neural Syst. 1998; p. 207–217.

80. Tkacik G, Walczak AM, Bialek W. Optimizing information flow in small genetic networks. Phys Rev E. 2009;80:031920.

81. Ganguli D, Simoncelli EP. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. Neural Computat. 2014;26:2103–2134.

82. Karklin Y, Simoncelli EP. Efficient coding of natural images with a populations of noisy Linear-Nonlinear neurons. In: Shawe-Taylor J, Zemel RS, Bartlett P, Pereira F, Weinberger KQ, editors. Adv Neural Inf Proc Syst 24. Cambridge, MA: MIT Press; 2011. p. 999–1007.

83. Pouget A, Deneve S, Ducom JC, Latham PE. Narrow versus wide tuning curves: What's best for a population code? Neural Comput. 1999;11:85–90.