Efficient population codes

Synonyms
Coding efficiency of neural populations, Coding accuracy of neural populations, Information transmission by neural populations

Definition
Natural stimulation caused by objects in the surrounding world do not stimulate single sensory receptors in isolation but lead to the activation of large numbers of neurons simultaneously. Thus, typical stimulus variables of interest are represented only implicitly in activation patterns across large neural populations. These patterns are statistical in nature since repeated presentation of the same stimulus usually leads to highly variable responses. The large dimensionality and randomness of the neural responses make it difficult to assess how well different stimuli can be discriminated. Depending on how effectively neurons share the labor of encoding the accuracy with which stimuli are represented can change dramatically. Thus, studying the efficiency of population codes is important for our understanding of both which information is encoded in neural populations and how it is encoded.

Detailed description
The question for the efficiency of population codes naturally arises if one seeks to relate the psychometric performance in a sensory discrimination task to the amount of information available from the activation of neural populations. For illustration, we consider an orientation discrimination task (Figure 1a) for which the behavioral performance can be summarized by the psychometric function (Figure 1b). The latter quantifies the discrimination error as a function of the stimulus parameter of interest. If one records from a neural population one can use decoders to solve the same task on the basis of the neural responses and compare the resulting neurometric performance to the psychometric one. In addition, one can try to predict the behavioral choice from the neural responses which is quantified using so-called choice probabilities.

For several tasks it has been observed that the neurometric function obtained for a single neuron or few is sufficient to explain the behavioral, psychometric performance. Given that the population of sensory neurons encoding for the stimulus contains many more neurons than just the recorded ones, this observation indicates that the encoding or the decoding of the task-relevant information from this sensory population is highly suboptimal: In case of suboptimal encoding, the information conveyed by each of these neurons is extremely redundant such that the entire population does not convey more information about the task than small subsets of neurons in this population (Zohary et al. 1994). In case of suboptimal decoding, the sensory population carries more information about the task than its individual neurons but the read-out of downstream neurons is unable to extract this distributed information (Beck et al. 2012).

The relationship between neural representation and behavioral performance can be studied using probabilistic modeling. In the beginning we will

Figure 1: In each trial of an orientation discrimination task two gratings are presented as depicted in a) where one of the two gratings is tilted to the right and the task is to say which of the two gratings is tilted. The task performance can be quantified by measuring the error frequency as a function of the the tilt angle. The resulting neurometric function is sketched in b).
assume that the stimulus entity of interest is parameterized by a one-dimensional variable $x$ (e.g. contrast, orientation, etc). Later, we will also discuss the encoding of high-dimensional stimuli such as natural images in the context of normative approaches to population coding that investigate optimal encoding strategies from an engineering perspective.

Throughout this chapter neural responses will be denoted by a vector $r$ which oftentimes describes the spike counts of the different neurons but could similarly describe other properties such as response latencies, interspike intervals or alike as well. Since the same stimulus usually leads to variable responses the relationship between stimulus and population response is determined by the conditional probability distribution $p(r|x)$. This conditional probability function specifies the encoding of the task relevant stimulus variable.

**Ideal observer analysis (Minimum discrimination error)**

In order to assess how much information in principle can be extracted from a neuron or a population it is necessary to determine the performance of an ideal observer. In a binary discrimination task with prior probabilities $p(x = x_1) = \lambda$ and $p(x = x_2) = 1 - \lambda$ for the two possible stimuli $x_1, x_2$ the discrimination error is minimized by the maximum a posteriori (MAP) estimator $\hat{x}_{MAP} = x_1$ if $p(x_1|r) > p(x_2|r)$ and $\hat{x}_{MAP} = x_2$ if $p(x_2|r) > p(x_1|r)$. Correspondingly, the minimum discrimination error (MDE) is given by (see Figure 2):

$$\text{MDE} = \int \min\{\lambda p(r|x_1), (1-\lambda) p(r|x_2)\} \, dr$$

In case of Gaussian distributions $p(r|x) = \mathcal{N}(r|f(x), C)$ with fixed noise covariance matrix $C$, one can show that the optimal classifier is linear and takes the following form:

$$\hat{x}_{MAP} = x_1$$

if $(\hat{\mu}_1 - \hat{\mu}_2)\top C^{-1}(\bar{x} - \langle \hat{\mu} \rangle) > \log \frac{\lambda}{1-\lambda}$, $\hat{x}_{MAP} = x_2$ otherwise

with $\mu_1 = f(x_1), \mu_2 = f(x_2)$, and $\langle \hat{\mu} \rangle = \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)$.

Furthermore, if $\lambda = 1 - \lambda = 0.5$, the MDE is given by

$$\mathcal{F} \left( \frac{d'}{2} \right) = \int_{-\infty}^{d'} \mathcal{N}(z|0,1) \, dz$$

where $d' := \sqrt{(\hat{\mu}_1 - \hat{\mu}_0)\top C^{-1}(\hat{\mu}_1 - \hat{\mu}_0)}$

In the absence of noise correlations, the noise covariance matrix $C$ is diagonal. Its dimensionality equals the number of neurons $N$ and $d'$ is of the order of $\sqrt{N}$. Thus, for uncorrelated noise the minimum discrimination error converges to zero in the limit of large $N$. In the presence of certain types of noise correlations, however, the error may saturate at a nonzero level. For example, this can happen in case of input noise which obviously cannot be removed by subsequent signal transmission (unless by the use of contextual side information).

As an illustration consider the following linear encoding $r = f(x + n_x) + n_r$ with $f'(x) = wx$, where $n_x \sim \mathcal{N}(0, \sigma_x^2)$ denotes the input noise and $n_r \sim \mathcal{N}(0, \sigma_r^2 I)$ constant additive output noise. In this case, the total noise covariance matrix reads $w\sigma_x^2 w\top + \sigma_r^2 I$, and its largest eigenvalue is given by $\sigma_x^2 + \sigma_r^2$ with

![Figure 2: The decision boundary of the MAP estimator is located at the intersection of the two curves. Correspondingly, the minimum discrimination error is given by the shaded area.](image)
eigenvector \( \mathbf{w} \). The remaining \((n - 1)\) dimensions belong to the same degenerated eigenvalue \( \sigma_r^2 \). Using the Sherman-Morrison formula it is possible to determine the inverse of this covariance matrix

\[
C^{-1} = \frac{1}{\sigma_r^2} \left( I - \frac{\sigma_r^2 \mathbf{w} \mathbf{w}^\top}{\sigma_r^2 + \sigma_x^2 \mathbf{w}^\top \mathbf{w}} \right)
\]

and

\[
d' = \Delta x \cdot \sqrt{\frac{A}{1 + \sigma_x^2 A}} \quad \text{with} \quad A := \frac{\mathbf{w}^\top \mathbf{w}}{\sigma_r^2},
\]

where \( A \) is of the order of \( N \). Accordingly, for finite input noise \( \sigma_x > 0 \), the population sensitivity \( d' \) converges to \( \frac{\Delta x}{\sigma_x} \) in the limit \( N \to \infty \). This can be one reason why behavioral performance may be not much better than one would predict on the basis of the neurometric performance of very few neurons. By estimating \((\Delta x/d')^2\) as a function of \( N \) for empirical data one can use the following relationship

\[
\left( \frac{\Delta x}{d'} \right)^2 = \sigma_x^2 + \frac{1}{A} \xrightarrow{N \to \infty} \sigma_x^2
\]

for estimating the effective amount of input noise (see also Figure 3).


**Discrimination, Mutual information, Minimum mean square estimation**

The minimum discrimination error is only one way to evaluate the coding accuracy of a neural population. Other popular measures are based on the mutual information between stimulus and response or minimum mean square estimation. In contrast to the discrimination approach, these other measures make assumptions about the stimulus distribution in order to obtain an average performance. If one assumes a binary stimulus distribution the mutual information between stimulus and response is also called the Jensen-Shannon divergence which provides an upper and lower bound on the Minimum discrimination error (see Fig. 4) but is not related to it in a one-to-one fashion (Berens et al. 2009). By the Jensen-Shannon information \( H_B^{-1}(1 - I_{JS}) \leq MDE \leq \frac{1}{2}(1 - I_{JS}) \) where the Jensen-Shannon information is equal to the mutual information \( I_{JS} = I[x : r] \) between the stimulus \( x \) and the responses \( r \).
For a Gaussian stimulus distributions, if we still assume a linear encoding and a Gaussian noise model, the joint distribution of stimulus and response is Gaussian and mutual information as well as the minimum mean square error can be computed analytically. For nonlinear encodings and non-Gaussian noise models these quantities are much more difficult to assess and approximation approaches are important.

**Fine discrimination and Fisher information**

We cannot expect that neural encoding is accurately captured by a linear Gaussian model. Nevertheless, in the case of fine discrimination, when $|x_2 - x_1|$ is sufficiently small, $d'$ can often be used to quantify approximately how well the two stimulus alternatives can be discriminated.

In a fine discrimination task, we study how well a stimulus at any location $x_1 = x$ can be discriminated from a stimulus with a slightly changed parameter value $x_2 = x + \Delta x$. Assuming that the tuning functions are smooth and the noise variance is finite, the discrimination performance approaches chance level $MDE \to 0.5$ in the limit of small differences in the stimulus parameter $\Delta x \to 0$. The asymptotic behavior in the limit of fine discrimination is more amenable to an analytical derivation of coding efficiency. Specifically, we can bound the minimum discrimination by the Jensen-Shannon information (see Fig.2) which for small $\Delta x$ can be approximated by the Fisher information:

$$ I_{JS}(\Delta x) \approx \frac{J(x)}{8} \Delta x^2 $$

Fisher information is defined as the variance of the score $V = \partial_x \log p(r|x)$ conditioned on the stimulus $x$:

$$ J(x) := \text{Var}[V|x] = E[V^2|x] $$

where the second equality holds because it generally holds $E[V|x] = 0$. Many papers on neural coding refer to the Cramer-Rao bound to motivate the evaluation of Fisher information which gives a lower bound on the mean squared error of any unbiased estimator:

$$ E[(\hat{x}_{unbiased}(r) - x)^2|x] \geq \frac{1}{J(x)}. $$

This bound, however, is only of intuitive use for the evaluation of encodings as it does not account for the fact that the signal $x$ must be a random variable and the precision with which $x$ can be decoded strongly depends on the stimulus distribution $p(x)$. Similarly, the local expansion of the Jensen-Shannon divergence must be wrong whenever $I_{JS}$ exceeds 1 bit since it generally holds $I_{JS} \leq 1$ bit. Thus the approximation for fine discrimination only holds if

$$ \Delta x^2 \ll \Delta x_c^2 \equiv \frac{8}{J(x)} $$

In the next section, we will discuss the evaluation of population codes if the stimulus distribution $p(x)$ is given.

**Coding efficiency and optimal encodings**

The question of coding efficiency critically depends on the stimulus distribution $p(x)$ which often requires additional considerations that go beyond the framework of discrimination. While for each pair of stimuli tested for in a discrimination task the prior distribution is reduced to two discrete alternatives, for continuous stimulus parameters in principle any pair of stimuli could be picked. Therefore, one often rather wants to assess the coding efficiency for a prior distribution that describes how the stimulus parameter is distributed under natural conditions. For example, if the stimulus parameter is contrast or speed one would assume that the natural distribution has a unimodal density over the positive real axis which takes its maximum at zero and is monotonically decreasing.
To be able to compare the performance of two coding schemes it is necessary to use a measure that provides a unique ordering relation. A direct ranking of neurometric functions, for example, is not possible unless in the special case if one function lies uniformly above or below the other one. If one was interested exclusively in fine discrimination one could compare just the slope of the neurometric function at chance level describing how fast the MDE drops from chance level and ignore the discrimination performance for larger differences in the stimulus parameter. In most cases however, one would rather be interested in the average performance when the stimuli are drawn from the natural distribution. The corresponding notions and assessments of coding efficiency can be very different. The optimization for fine discrimination is tightly related to optimizing Fisher information, whereas an optimization of the average discrimination error (Berens et al. 2011) is more similar (but not equivalent) to optimizing mutual information (Yarrow et al. 2012) or the minimum mean squared error.

There is a special condition for which the minimum mean squared error is determined by Fisher information. If the responses of the different neurons can be described as independent, identically distributed (i.i.d.) observations, or more generally if the conditions of the central limit theorem are fulfilled then Fisher information can be used to determine an asymptotic approximation of the decoding error Bethge et al. (2002). Optimal codes, however, generically exhibit highly heterogeneous encodings which preclude the use of the central limit theorem. In particular, one should note that the linear increase of Fisher information in case of i.i.d. observations corresponds to a very slow logarithmic growth in mutual information. In contrast, encodings that are optimized for mutual information usually achieve a much faster, linear growth of the mutual information. The two situations can be illustrated with the following two prototypical examples:

\[
p_{\text{iid}}(\mathbf{r}, \mathbf{x}) = p(\mathbf{x}) \prod_{k=1}^{N} p(r_k|f(\mathbf{x})) \quad \text{vs} \quad p_{\text{fac}}(\mathbf{r}, \mathbf{x}) = \prod_{k=1}^{N} p(r_k, f_k(\mathbf{x})).
\]

It is easy to derive that for the left example of i.i.d. coding mutual information grows only logarithmically, \( I[\mathbf{r}, \mathbf{x}] \propto \log(N) \), whereas for the right example of a factorial code (Nadal & Parga 1994) mutual information grows linearly, \( I[\mathbf{r}, \mathbf{x}] \propto N \). Thus, for large populations of neurons i.i.d. coding is highly inefficient compared to factorial coding.

In cases other than i.i.d. coding (or, more generally, whenever the conditions for the central limit theorem are not fulfilled) the behavior of Fisher information can be very different from that of mutual information. For example, it is possible to build factorial codes for which Fisher information grows only linearly in \( N \) just like in the case of i.i.d. coding but an exponential or super-exponential growth of Fisher information is possible as well. Conversely, it is easy to construct encodings for which Fisher information grows arbitrarily fast with \( N \) but this does not imply that mutual information grows linearly (Bethge et al. 2002). Fisher-optimal codes are exclusively optimized for fine discrimination and may exhibit extremely poor performance for any discrimination for which \( \Delta x^2 \geq \Delta x_t^2 \).

**Task-dependence of coding efficiency**

While it is obvious that coding efficiency is highly task dependent, it is important to understand the implications of this task-dependency for the interpretation of the available experimental data. As mentioned in the beginning of this chapter it is typically observed that the neurometric function obtained for a single neuron or few is sufficient to explain the behavioral, psychometric performance, and that this may hint to highly inefficient encoding or decoding. If the encoding is fixed (i.e. independent of the task), it is straightforward to show that the efficiency of the encoding will depend on the task. For illustration, let us assume we have the same number of neurons as stimulus dimensions and the following encoding with independent noise model:

\[
p_{\text{enc}}(\mathbf{r} | \mathbf{x}) = \prod_{k=1}^{N} p(r_k | x_k).
\]
Depending on which stimulus distribution $p(x)$ we use, one can obtain either a highly redundant iid coding scheme or a maximally efficient factorial code. Specifically, if we set $p(x) = \prod_{k=1}^{N} p(x_k)$ we obtain the maximally efficient factorial code

$$p_{\text{enc}}(r|x)p(x) = \prod_{k=1}^{N} p_{\text{enc}}(r_k|x_k)p(x_k) = \prod_{k=1}^{N} p(r_k, x_k),$$

whereas if we choose a stimulus distribution $p(x) = p(x_1)\prod_{k=2}^{N} \delta(x_k - x_1)$, for which it always holds $x_1 = x_2 = \cdots = x_N$, we then obtain the highly redundant i.i.d. coding scheme. The first example roughly corresponds to a task where any two images are to be discriminated while in the second case only full-field uniform gray images need to be discriminated.

The mutual information between stimulus $x$ and population response $r$ cannot be larger than the entropy of the stimulus distribution. For good reasons, typical laboratory tasks are inherently biased towards low entropy stimuli. More specifically, most experiments are designed such that it is not critical which neurons exactly are recorded. In vision, this is achieved by using global homogeneous stimuli such as global motion patterns or gratings. By varying only one parameter of these stimuli like the net velocity or orientation it is achieved that all neurons encode the same parameter in an i.i.d. fashion. Correspondingly, mutual information cannot grow faster than logarithmically with the number of neurons yielding a highly suboptimal population coding strategy.

For many tasks, however, it is likely that this type of suboptimality is not sufficient to explain why behavioral performance does not utilize more information than provided by a single or few neurons. For spatially homogeneous stimuli, the information of neurons with different receptive field locations should still accumulate such that doubling the number of neurons should reduce the squared error by a factor of two as predicted by an i.i.d. coding scheme. This fact, however, is not consistent with the observation that the information of a local population of neurons reaches behavioral performance. Thus, it seems that for typical low-level tasks the read-out cannot integrate information from different local regions.

In conclusion, both encoding and decoding are likely to render behavioral performance highly suboptimal in case of common laboratory tasks for which the decision space has low entropy and for which the stimulus size is rather large. In the future, important new insights about principles of neural coding and sensory decision making may be obtained by choosing more ecologically relevant tasks such as natural scene classification. The importance of studying natural tasks may be corroborated by the example of grid cell encoding in hippocampus. Despite the fact that the stimulus parameter is only two-dimensional—namely the x-y-position of the animal—the mutual information between this stimulus parameter and the grid cells scales proportional to the number of neurons (Sreenivasan & Fiete 2011, Mathis et al. 2012). Similarly, one would expect a linear growth of mutual information for the encoding of natural images in the visual system.

References


