Chapter 3: Principles of perceptual representation

If neural computation is the manipulation of representations, what are those representations? This chapter focuses on perceptual representations constructed by brain areas close to the sensory periphery. Rather than tabulating an exhaustive list of representations attested by brain activity, we organize them into a small set of general principles (efficiency, sparsity, prediction) which can be formulated as optimization problems. The relationships between different principles hint at the possibility of a unifying theory, although there are also some fundamental tensions between them.

Everything starts with representation: computations in the brain do not have access to the world itself, only to representations of the world. Consequently, there is evolutionary pressure on the brain to construct representations that are useful for computation. What kinds of representations are useful for what kinds of computations? The approach of this chapter is to define a set of optimality principles based on different computational goals. Each principle implies certain properties about the neural representations that we can measure experimentally.

For the purposes of this chapter, we will take 'neural representation' to mean the set of tuning functions $\{f_d(s)\}$ for a population of neurons, where d indexes neurons and s is a stimulus. A practical reason for focusing on tuning functions is that this is by far the most abundant source of published data. However, this is not the only way to think about representation. For example, neurons could potentially represent information through spike times (see Chapter 2). We will not cover alternative representational formats in this chapter, but we will touch upon them at other points in the book.

Recall from Chapter 2 that a tuning function maps states (s) to expected firing rates. In this chapter, we sometimes refer to s as the stimulus, in accordance with standard terminology in perceptual neuroscience.

Another caveat is that this chapter will not focus on how representations are learned or used, deferring these topics to later chapters. The primary goal is to establish a connection between abstract optimality principles and neurophysiological data.

1 *Representations optimized for fitness*

Like all biological structures, brains have evolved through a process of natural selection. In every generation, some individuals reproduce more than others, thereby propagating their genotype (and the phenotype encoded by the genotype) to the next generation. Here we will just talk about phenotypes to keep things simple. In particular,

A *genotype* is an individual's complete genetic material, and a phenotype is the complete set of observable characteristics of the individual, specified by a (possibly stochastic) genotype-to-phenotype mapping.

we will take the phenotype to be the vector-valued tuning function f that maps states (or stimuli) to expected firing rates.

As an example, suppose $s = 1$ corresponded to the presence of a predator, and $s = 0$ corresponded to its absence. A high-fitness tuning function f^+ would map these different states to distinct patterns of neural activity, $f^+(1) \neq f^+(2)$. In contrast, a low-fitness tuning function f^- would have low discriminability, $f^-(1) \approx f^-(2)$.

The fitness $\psi(f)$ of a phenotype f is a real-valued scalar summarizing the reproductive advantage conferred upon an organism by having that phenotype. We can formalize this as the *growth rate* of the phenotype frequency over time (assuming deterministic dynamics for simplicity):

$$\psi(f) = \frac{d}{dt} \log N(f), \tag{1}$$

where $N(f)$ is the number of individuals in a population with phenotype f. The above definition leads to dynamics of the following form:

This is known as the *replicator equation* (Schuster and Sigmund, 1983).

$$\dot{P}(f) = P(f)[\psi(f) - \bar{\psi}], \tag{2}$$

where $P(f) = N(f)/N$ is the proportion of the population (with size N) that has phenotype f, and $\bar{\psi}$ is the average fitness of the population. This equation tells us that a phenotype will spread in the population when its fitness is superior to the average fitness; the speed of this spread depends on the proportion of the population that already has that phenotype.

A particular phenotype distribution P is a *rest point* if, once reached, it does not change over time. It is *evolutionarily stable* if any small perturbation decreases fitness (i.e., no 'mutant' can successfully invade the population). An important result in the theory of evolutionary dynamics is that if P is evolutionarily stable then it is also a rest point (Taylor and Jonker, 1978); the converse is not true in general.

Technically, Taylor and Jonker (1978) require P to satisfy a mild regularity condition.

In principle, we could run the evolutionary dynamics forward and see what kind of representations emerge. However, in practice the lack of closed-form solutions makes it difficult to draw general conclusions about neural representations. In the following sections, we instead use other proxies for fitness which are more analytically tractable.

2 *Representations optimized for efficient coding*

One way to increase fitness is to encode more information about sensory signals in the neural representation. This comes with an energetic cost, because increasing the fidelity of neural transmission or adding more neurons requires the provision of additional metabolic resources. These trade-off are formalized in the principle of *efficient coding*: neural representations are optimized to communicate information subject to a set of resource constraints.

The application of efficient coding to the brain was initiated by Barlow (1961).

What are the resource constraints? First, neurons have minimum and maximum firing rates. Second, spike counts are discrete, which

means that firing rates calculated over some fixed interval of time are discrete; thus, firing rates can only distinguish a finite number of different input levels. Third, irreducible sources of noise (e.g., thermal noise that affects ion channels) limit the precision with which small changes in a neuron's input lead to commensurate changes in the firing rate. These constraints mean that a neuron has an upper bound on how much information it can communicate about its inputs. The efficient coding principle states that the neuron should be configured to operate at this upper bound.

To formalize this principle, let's start with a single neuron conceptualized as a communication channel for a scalar stimulus s. It receives inputs ("messages") about s that it communicates to downstream neurons via its firing rate (the channel output). Assume that the firing rates can distinguish M different input levels. Information is usually measured in "bits" (binary digits); with M firing rate levels, a neuron can communicate up to $\log M$ bits per sample. This upper bound is achieved when each firing rate is used with equal frequency across the distribution of inputs. To see why, we need to define information rate more precisely.

Intuitively, a communication channel is informative to the extent that its outputs change the receiver's beliefs about its inputs. In the context of neural transmission, a neuron's activity x (the spike count vector) is informative to the extent that it allows downstream neurons to reduce their uncertainty about the stimulus s. We can quantify this uncertainty in the following way. The "surprisal" of observing s (measured in bits) is defined as $-\log p(s)$, where we have assumed that the base of the logarithm is 2. If a stimulus is perfectly predictable, $p(s) = 1$, then its surprisal will be 0; if it's a toss-up, the surprisal will be 1. Uncertainty can then be quantified as the average surprisal, or *entropy*, capturing the idea that one is more uncertain about stimuli that one can't predict well:

Returning to our earlier example, if we rarely encounter a predator, $p(s) = 0.001$, then surprisal will be about 10 bits when we encounter the predator. the rest of the time, surprisal will be close to 0 bits. So the average surprisal (entropy) will be about 0.01 bits.

$$\mathcal{H}[s] = \mathbb{E}[-\log p(s)] = -\sum_s p(s) \log p(s). \tag{3}$$

In the same fashion, we can quantify uncertainty about s after observing neural activity x in terms of the *conditional entropy*:

$$\mathcal{H}[s|x] = \mathbb{E}[-\log p(s|x)] = -\sum_x p(x) \sum_s p(s) \log p(s|x). \tag{4}$$

The uncertainty reduction afforded by observing x is the difference between these two entropies, the *mutual information*:

Mutual information is also known as *relative entropy*.

$$\mathcal{I}[s; x] = \mathcal{H}[s] - \mathcal{H}[s|x]. \tag{5}$$

Another way of thinking about uncertainty reduction is through the lens of Bayesian updating. Before observing x, the receiver starts with

a prior $p(s)$, and then updates this to $p(s|x)$ using Bayes' rule. The degree of change from the prior to the posterior can be quantified by the Kullback-Leibler (KL) divergence:

$$\mathcal{D}[p(s|x)||p(s)] = \sum_s p(s|x) \log \frac{p(s|x)}{p(s)}. \tag{6}$$

The expectation of the KL divergence under $p(x)$ is equal to the mutual information:

$$\mathcal{I}[s;x] = \sum_x p(x)\mathcal{D}[p(s|x)||p(s)]. \tag{7}$$

Thus, the mutual information characterizes the average degree of change in the posterior over s after observing x.

The mutual information is a symmetric function of its arguments, which means we can equivalently write it as:

$$\mathcal{I}[s;x] = \mathcal{H}[x] - \mathcal{H}[x|s]. \tag{8}$$

This is convenient for thinking about neural coding. The first term measures the variability of firing rates across the distribution of inputs. The second term measures transmission noise. If we assume that transmission noise is negligible, then $\mathcal{H}[x|s]$ approaches 0. Maximizing information then corresponds to maximizing output entropy. This is achieved when the output response distribution $p(x)$ is uniform, which can be implemented by setting the tuning function to be the cumulative distribution function (CDF) of the stimulus distribution $p(s)$:

$$f(s) \propto P(s) = \int_{s' \leq s} p(s')ds'. \tag{9}$$

This is known as the *probability integral transform*. In image processing it is known as *histogram equalization*.

We can think of this tuning curve as a rank transformation, where the firing rate for a stimulus corresponds to its normalized rank in the stimulus distribution. These ranks will change quickly in high-density regions of the stimulus space, so the neuron will be most sensitive to changes in these regions.

As an example, consider the distribution of brightness contrasts in a natural image (Figure 1). Although the range of contrasts is rather large, the distribution is concentrated in a narrow range. The tuning curve derived from the CDF shows that sensitivity is optimized around the mode of the distribution, flattening out for very low and very high contrast values.

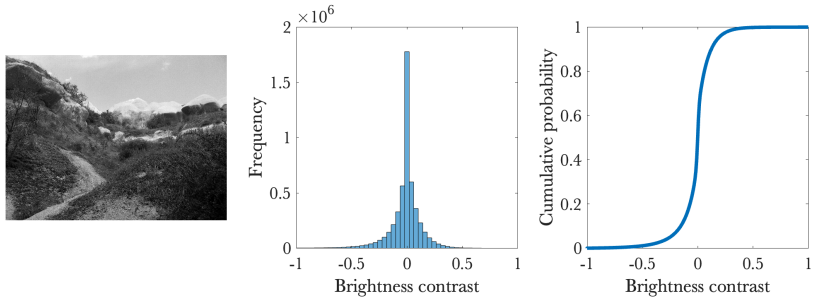

Figure 1: **Brightness contrast statistics**. The distribution of brightness contrast in a natural image and an estimate of the cumulative distribution function.

Laughlin (1981), in a classic study of efficient coding, connected the CDF for brightness contrast to neural responses of the blowfly's large monopolar cells, which are analogous to bipolar cells in the vertebrate retina. These neurons exhibit a graded, sigmoidal response to brightness contrast that is well-matched to the CDF (Figure 2). This is remarkable considering that there are no free parameters; the tuning curve is derived entirely from image statistics.

In this example, efficiency is gained by only coding variations around the mean—absolute magnitude is discarded. This can lead to powerful illusions, such as the one shown in Figure 3. The same stimulus is mapped onto different firing rates depending on the stimulus distribution, such that a brightness difference is perceived where there is no objective difference.

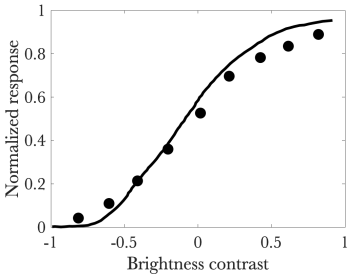

Figure 2: **Efficient coding in the blowfly eye**. The circles show the normalize responses of large monopolar cells at different contrast levels. The line shows the cumulative distribution function of contrast estimated from images of the fly's natural environment. Adapted from Laughlin (1981).

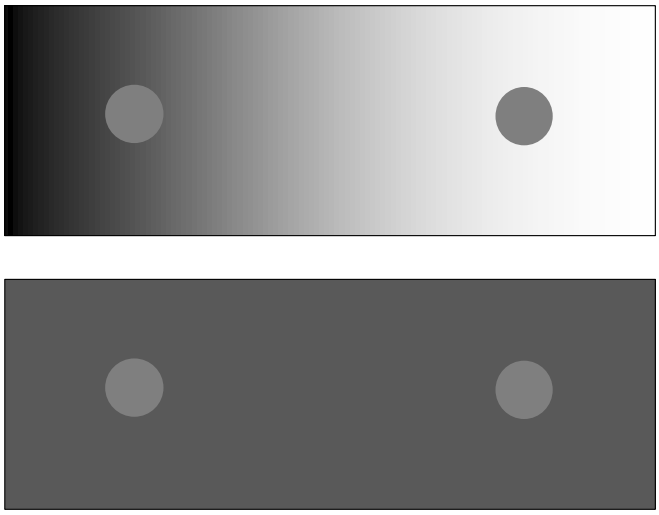

Figure 3: **A brightness illusion**. (Top) The disc on the bright background appears darker than the disc on the dark background. (Bottom) When displayed on a uniform background, the two discs have the same brightness.

2.1 *Noise*

So far, we have assumed that noise is negligible. However, this assumption is questionable. Consider again the blowfly large monopolar cells, which receives input from photoreceptors. The precision with which these photoreceptors can detect light is limited by several sources. One is the noise from photon counts: because they are quantum particles, the number of photons hitting a photoreceptor has non-negligible variability when aggregated over the timescale relevant for vision. A second source of noise comes from the transduction of photons into membrane voltage (Lillywhite and Laughlin, 1979). Taken together, these noise sources imply that $\mathcal{H}[x|s] > 0$, and thus the optimal channel needs to balance entropy maximization with noise suppression. The latter can be accomplished by smoothing the noisy signals (e.g., by temporal averaging) before computing the CDF (Atick and Redlich, 1992). Consistent with this hypothesis, van Hateren (1992) showed that under conditions of low background illumination (when photon noise is expected to be higher), neural responses to light flashes are slower and more prolonged, indicative of temporal averaging.

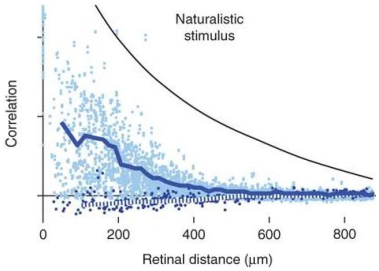

Figure 4: **Decorrelation in the retina**. Pairwise correlations between retinal ganglion cells. The solid blue line shows the correlation for cells of the same type; the dashed blue line shows the correlation for cells of different types. The black line shows the correlation between pixels, a proxy for correlations present in the photoreceptor array. Adapted from Pitkow and Meister (2012).

2.2 *Efficient coding with multiple neurons*

The efficient coding principle presented thus far applies only to a single neuron. What is the optimal efficient code for a population of neurons? One strategy for answering this question (see Atick, 1992; Nadal and Parga, 1994) is to start by noting that if each neuron (d) is tuned to a different stimulus (s_d), then we're back to the single-stimulus setting described above. This realizes a *factorial code*, where the marginal firing rate distribution decomposes into a product of factors:

$$p(x) = \prod_d p(x_d). \tag{10}$$

The entropy then decomposes additively:

$$\mathcal{H}[x] = \sum_d \mathcal{H}[x_d]. \tag{11}$$

As a consequence, the joint efficient coding problem separates into independent efficient coding problems. Thus, low redundancy of tuning curves allows neurons to solve a simpler efficient coding problem.

An architecture of this sort may exist in the retina, where the retinal ganglion cells (the stage of retinal processing prior to transmission into the brain) exhibit strong, though incomplete, decorrelation of firing (Figure 4). This appears to arise from two mechanisms

Even though retinal ganglion cells are not perfectly decorrelated, Nirenberg et al. (2001) showed that more than 90% of the information they carry about stimuli could be extracted by a decoder that ignored the correlations.

(Pitkow and Meister, 2012). One mechanism is the center-surround structure of the receptive fields (Figure 5), which implement a form of *predictive coding* (Srinivasan et al., 1982): each cell only responds to the extent that excitatory input (in the center of its receptive field) cannot be predicted by neighboring inputs (which send inhibitory drive, thereby creating the antagonistic surround). In this way, the cells respond primarily to uncorrelated prediction errors.

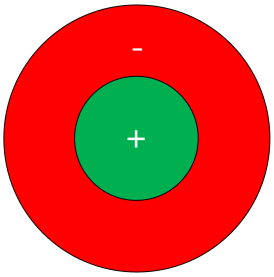

Figure 5: **Center-surround receptive field**. Stimuli in the center of the receptive field (green area) excite the neuron; stimuli in the surround of the receptive field (red area) inhibit the neuron.

To see how this works, consider a simple linear model in which a multidimensional stimulus $s = (s_1, \ldots, s_D)$ corresponds to a pattern of photoreceptor activity, where s_d reports the light intensity at retinotopic location d. The activity of nearby photoreceptors will tend to be highly correlated because of correlations in the light intensities. We model the activity of retinal ganglion cells (x) as a noisy linear combination of photoreceptors (although in reality the connection is indirect):

$$x_d = s_d - \sum_n w_{nd} s_n + \epsilon, \tag{12}$$

where n ranges over the neighborhood of d, and ϵ is uncorrelated noise. If the second term is a good predictor of the first term, then the elements of x will only reflect the uncorrelated prediction errors.

A second decorrelation mechanism is the non-linearity in the response function, such that cells do not respond appreciably until the excitatory drive is very strong. This has the effect of making responses sparse (only a small proportion of cells are active for any given stimulus), and generally reduces correlations between cells. As we will discuss below, sparsity is itself a representational principle used by the brain.

2.3 *Efficient coding with a convolutional population*

We can obtain further analytical insight into efficient codes by making stronger assumptions about neural populations. Here we summarize a model derived by Ganguli and Simoncelli (2014), who posited an idealized "convolutional" population of neurons with identical, uniformly spaced tuning functions—i.e., each idealized tuning function is a shifted copy of a "prototype" tuning function $\tilde{f}$, such that the population obeys a tiling property:

$$\sum_d f_d(s - s_d) \approx 1, \tag{13}$$

where $\{s_d\}$ is a set of evenly spaced points in stimulus space (what we'll call the *stimulus lattice*). The firing rates are then rescaled the preferred stimuli warped to maximize an approximation of the mutual information (the Fisher information) subject to an upper bound

The log of Fisher information may be either an upper bound or a lower bound (up to a constant) on the mutual information, depending on assumptions about firing rate variability (Wei and Stocker, 2016). For Poisson neurons with high firing rates, variability is approximately Gaussian, in which case Fisher information is a lower bound.

(G) on the average firing rate summed across the population:

$$f^* = \underset{f}{\operatorname{argmax}}\, \mathbb{E}[\log J(s)] \tag{14}$$

$$\text{subject to: } \mathbb{E}\left[\textstyle\sum_d f_d(s)\right] \leq G, \tag{15}$$

where both expectations are taken with respect to $p(s)$. The Fisher information $J(s)$, implicitly a function of f, is defined as:

$$J(s) = \sum_x p(x|s) \frac{\partial^2}{\partial s^2} \log p(x|s), \tag{16}$$

where x is the spike count vector for the population. For independent Poisson neurons, this becomes:

$$J(s) = \sum_d \frac{f_d'(s)^2}{f_d(s)}, \tag{17}$$

where $f_d'(s)$ is the derivative of the tuning curve for neuron d. This objective function is convenient because it can be expressed in terms of the tuning functions and their derivatives, enabling an analytical solution.

Each tuning function is parametrized according to:

$$f_d(s) = g(s_d^*) \tilde{f}(\Gamma(s) - s_d), \tag{18}$$

where $g(s)$ is a gain function (controlling the scale of the tuning function), $\Gamma(s)$ is a warping function (controlling the shape of the tuning function), and $s_d^* = \Gamma^{-1}(s_d)$ is the preferred stimulus of neuron d after warping. The warping function is derived from an underlying density function $\gamma(s)$:

The warping function is the CDF of the density function.

$$\Gamma(s) = \int_{s' \leq s} \gamma(s') ds'. \tag{19}$$

Intuitively, the density function controls to what extent the population of neurons is tuned to a particular state, via both the spacing of their preferred stimuli and their tuning widths. This can be seen more easily by taking a first-order Taylor expansion of $\Gamma(s)$ around s_d^*:

$$f_d(s) \approx g(s_d^*) \tilde{f}(\gamma(s_d^*)(s - s_d^*)), \tag{20}$$

which shows that high-density regions of the stimulus space will have both narrower spacing and narrower tuning.

The information-maximizing density and gain functions are given by:

$$\gamma(s) \propto p(s), \qquad g(s) \propto G. \tag{21}$$

This solution shows that high probability stimuli should be encoded with higher tuning density (i.e., there should be greater discriminability of neural activity for high probability stimuli, due to the fact that more neurons are tuned to that part of the stimulus space), and that the gain is constant across stimuli. Thus, different stimuli are encoded with varying precision depending on their prior probability, but each neuron participates roughly equally in the representation of the stimulus distribution—a hallmark of efficient coding, as we discussed earlier.

Under this solution, the optimal warping function $\Gamma(s)$ is proportional to the CDF of the prior distribution.

Several empirical implications follow from this analysis. One is that the average firing rate for a neuron, $\mathbb{E}[f_d(s)] = \int_s p(s) f_d(s) ds$, should be approximately constant across the population, due to the constant gain function. Figure 6 shows the distribution of average firing rates across a population of neurons in auditory cortex. We can see that the distribution is strongly peaked and skewed; a large proportion of the probability mass is concentrated around 1 Hz. Notably, the stimulus-evoked distribution was found to be very similar to the spontaneous distribution (not shown here), consistent with the idea that average firing rate is invariant.

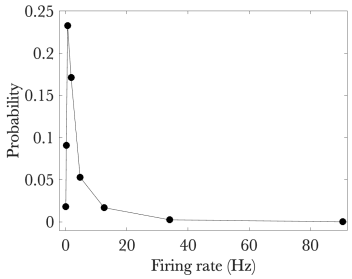

Figure 6: **Firing rate distribution in auditory cortex following presentation of auditory stimuli**. Adapted from Hromádka et al. (2008).

A second implication is that the distribution of preferred stimuli should match the prior distribution. This is because the optimal warping function is the CDF of the stimulus distribution, and the preferred stimuli are obtained by taking the inverse CDF evaluated at each stimulus on the stimulus lattice. This generates samples from $p(s)$, an algorithm known as *inverse transform sampling* (the inverse of the probability integral transform described above). Orientation of edges provides a useful case study, because the distribution in natural images is non-uniform (Figure 7, left); cardinal orientations (vertical and horizontal) are more prevalent than oblique orientations. The distribution of preferred orientations in primary visual cortex (V1) closely matches this distribution (Figure 7, right), as predicted by the efficient coding theory.

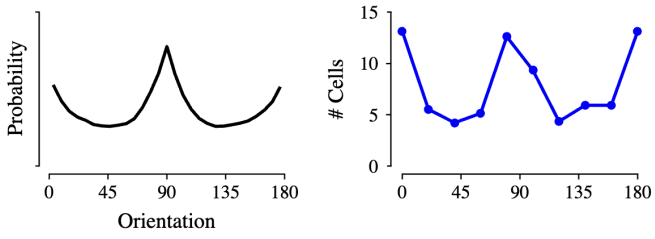

Figure 7: **Efficient coding of orientation in primary visual cortex (V1)**. (Left) Orientation distribution derived from natural images. (Right) Distribution of preferred stimuli from orientation-tuned cells recorded in Macaque V1 (Mansfield, 1974). Reproduced from Ganguli and Simoncelli (2010).

Figure 8 shows how this works in greater detail. We use the parametric form $p(s) \propto 2 - |\sin(s)|$, which is a decent approximation to the distribution of edge orientations in natural images (Girshick et al., 2011; Wei and Stocker, 2015). The prototype population is a set

of cosine tuning curves of the form $f_d(s) = \exp[\cos(s - s_d^*)/\nu]$, where s_d^* is the preferred orientation for neuron d and ν is the tuning width (shared by all neurons). The prototype population is transformed into an efficient code, which exhibits a concentration of tuning curves in the high-probability region of orientation space—much like what is seen in V1.

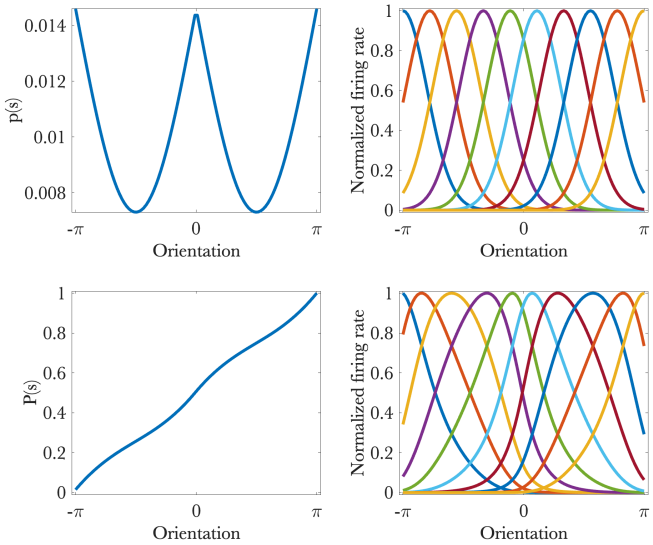

Figure 8: **Efficient orientation coding**. (Left) Probability density (top) and CDF (bottom) of orientation. (Right) Prototype population (top) transformed into an efficient population code (bottom) using the solution from Ganguli and Simoncelli (2014).

A third implication is that psychophysical performance should correlate with coding fidelity. In particular, stimuli in high-probability regions of the stimulus space should be more discriminable, because more neural resources are allocated to those regions. This is true for orientation: cardinal orientations are more accurately discriminated than oblique orientations, a phenomenon known as the *oblique effect* (Appelle, 1972; Girshick et al., 2011).

2.4 *Fisher information as the fundamental unit of analysis*

The previous section made a set of parametric assumptions about the family of tuning functions, and then optimized these parameters. While this is a powerful approach that lends itself to experimental verification, one might wonder whether these assumptions are too restrictive. After all, the space of possible tuning functions is vast. An alternative approach is to abstract away from the details of specific tuning functions and characterize the neural representation in terms of more global properties.

Ganguli and Simoncelli (2014) found that the optimal Fisher information scales quadratically with stimulus probability: $J^*(s) \propto p(s)^2$. It turns out that this is a general property of efficient coding. Wei and Stocker (2015) obtained the same result by maximizing mutual

information subject to an upper bound on total Fisher information:

$$\int_s \sqrt{J(s)} ds \leq C. \tag{22}$$

Prat-Carrabin and Woodford (2021) analyzed a more general objective function:

$$f^* = \underset{f}{\operatorname{argmin}} \int_s \frac{p(s)^\alpha}{J(s)^{\beta/2}} ds \tag{23}$$

$$\text{subject to: } \int_s \sqrt{J(s)} ds \leq C. \tag{24}$$

In the limit $\beta \to 0$ and $\alpha = 1$, this recovers the efficient coding objective. When $\beta = 1$ and $\alpha = 2$, the objective function corresponds to maximizing expected reward in a discrimination task where the decision-maker receives a constant reward for a correct answer. The general form for the Fisher information under the optimal solution is given by:

$$J^*(s) \propto p(s)^{\frac{2\alpha}{\beta+1}}. \tag{25}$$

Thus, the general form is always a power function of the stimulus distribution. Inspecting this equation, we can see that the discrimination task leads again to $J^*(s) \propto p(s)^2$. Prat-Carrabin and Woodford also considered other tasks which lead to different powers. For example, the representation that minimizes squared estimation error (see next chapter) leads to $J^*(s) \propto p(s)^{2/3}$.

The take-away from this section is that it's possible develop a general theoretical account of optimal representation that abstracts away from specific assumptions about tuning functions. The cost of doing this, of course, is that the predictions about the precise form of tuning functions in the brain become weaker.

3 *Representations optimized for sparsity*

A number of experiments have demonstrated that only a small proportion of neurons are active at any given time. For example, Yoshida and Ohki (2020) measured the response of V1 neurons to natural images, finding that on average 2.5% of neurons were active for each image. Of these responsive neurons, only 5.4% of them exhibited overlap between pairs of images. In olfactory cortex, individual odors activate on average 10% of neurons (Poo and Isaacson, 2009). In auditory cortex, sounds activate on average 5% of neurons (Hromádka et al., 2008). In the medial temporal lobe, 40% of visually responsive neurons were selective for pictures of a single person, place, or object (Quiroga et al., 2005). These observations suggest a general principle of *sparse coding*.

One way to think about the logic underlying this principle is to start from the observation that our sensory data arise from many different causes, only a few of which are present at any given moment. For example, if your eyes scan a scene, you'll see a relatively small set of objects; the high-dimensional time series of retinal images arises from different glimpses of a slowly changing object set. As a consequence, the retinal images live on a low-dimensional subspace defined by the set of currently active causes (objects in this case). The problem of perceptual inference is identifying which causes are active at any given moment, while the problem of perceptual learning is identifying the stable mapping between causes and images.

3.1 *Linear sparse coding*

Olshausen and Field (1996) formalized these ideas in terms of a linear model of stimuli:

$$s_n = \sum_d \phi_{dn} v_d + \epsilon_n, \tag{26}$$

where v_d is the activation of cause d, ϕ_{dn} is the contribution of cause d to stimulus component s_n (e.g., photoreceptor activity corresponding to retinotopic location n), and ϵ_n is a Gaussian error term. They also placed a prior on v that favors sparse activation of causes (i.e., most causes are inactive):

This is known as the Laplace distribution.

$$p(v_d) \propto \exp(-\lambda |v_d|), \tag{27}$$

where $\lambda > 0$ is a scaling parameter. Putting these together and taking logarithms leads to the following optimization problem:

The optimal parameters here correspond to the mode of the posterior distribution, $p(v, \phi | s)$.

$$v^*, \phi^* = \operatorname*{argmin}_{v,\phi} \sum_n \left[s_n - \sum_d \phi_{dn} v_d \right]^2 + \lambda \sum_d |v_d|. \tag{28}$$

Optimizing this cost function produces a sparse code $f(s) = v^*$ for each stimulus, as well as the optimal linear transformation ϕ^*.

The resulting tuning functions are remarkably similar to those of "simple" cells in V1: they are spatially localized, and tuned to edges of particular orientations and frequencies (Figure 9).

Tuning functions of this kind are sometimes known as *Gabor filters*, a classical model of V1 simple cells (Marcelja, 1980).

3.2 *The metabolic argument for sparsity*

So far, we have focused on the argument that sparsity is a good assumption about the distribution of stimuli, which in turn makes it a good constraint on neural representations. Here we will discuss how sparsity can also reduce the metabolic cost of information processing.

The two major contributors to energy consumption in the brain are spiking and synaptic transmission (Attwell and Laughlin, 2001).

The neocortex (the primary division of the cerebral cortex) consumes 44% of the brain's energy budget.

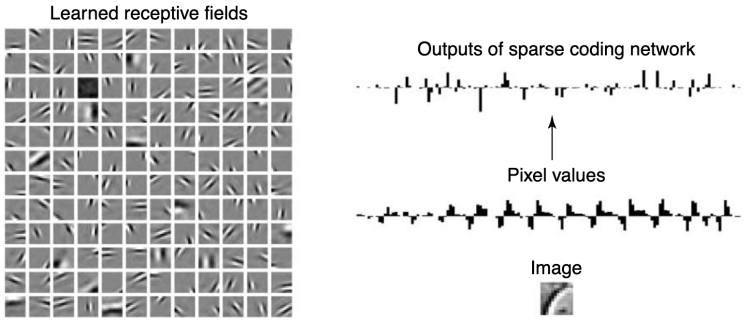

Figure 9: **Visual receptive fields from sparse coding**. (Left) Receptive fields learned from optimizing the sparse coding objective. (Right) Simulated neural responses show sparsification of the visual input. Reproduced from Olshausen and Field (2004).

Lennie (2003) estimated that a single spike in human cortex costs 2.4×10^9 molecules of ATP. Based on cortex-wide measurements of glucose metabolism, Lennie came to the conclusion that single cortical neurons would need to spike on average less than once per second in order to satisfy the energy budget. This is remarkably low given the fact that electrophysiology studies have reported spike rates of up to 100 Hz. In fact, both of these observations are present in Figure 6, which we saw earlier: neurons typically spike close to 1 Hz, but can infrequently achieve much higher spike rates.

Lennie then examined the case of a typical "strong" response to a stimulus, with a spike rate of 10 Hz over 200 ms. In this case, the energy budget could support concurrent spiking in 0.3% of neurons. Even allowing for large transient increases in glucose consumption during intense sensory stimulation, the average spike rate can only increase by a few spikes per second, yielding an estimate of 4% of concurrently active neurons spiking at 50 Hz. The point of these calculations is to show that metabolic constraints necessitate sparsity of neural activity.

Don't take these numbers too literally, they're mainly for the sake of argument.

Intriguingly, the glucose metabolism rate in human cortex is 3 times lower than in rats, despite the fact that individual spikes use 3.3 times *more* energy per spike. This suggests that the human brain has evolved a high degree of information processing efficiency. It is tempting to speculate that this is the dividend from a more powerful internal model: we can parsimoniously encode high-dimensional sensory signals into low-dimensional and sparse causes.

4 *Representations optimized for prediction*

As Shakespeare wrote in The Tempest, "what's past is prologue"—the future can, at least partially, be predicted from the past. A perceptual system that can exploit this predictability can reap several benefits.

For example, our high-acuity (foveal) vision is limited to a small

The fovea is a structure at the center of the retina containing tightly packed cone photoreceptors, the main source of visual information under well-lit conditions.

portion of the visual field (about 2 degrees, approximately twice the width of your thumbnail held at arm's length). One reason that we perceive much more than the central 2 degrees is that our eyes are making frequent saccades—ballistic, high-velocity movements to salient regions of the visual field. Saccades to unpredictable stimuli usually take around 200 ms. In contrast, saccades to predictable stimuli can be initiated *even before the stimulus appears* (Stark et al., 1962). This is useful in a fast-changing but predictable world, where predictive saccades can increase the rate of information flow.

A similar story can be told about smooth pursuit, where the eyes stay fixated on a moving object. With extended experience tracking an object that follows a repeating path (e.g., sinusoidal motion), smooth pursuit improves and can even make anticipatory changes in direction (Dodge et al., 1930).

See Chapter 6 for more on the mechanisms underlying cued attentional allocation.

Even without eye movements, prediction can improve perception. A centrally presented cue, indicating the likely future location of a target, speeds detection of the target when it appears in the cued location, and slows down detection when it appears unexpectedly in an uncued location (Posner, 1980).

With these observations as backdrop, let's return to the question of perceptual representation. We will formalize a principle of predictive optimality for neural encoding, and then examine to what extent this principle fits with empirical data.

4.1 The predictive information bottleneck principle

Let s_{past} denote the history of stimuli, and s_{future} denote future stimuli that haven't been observed yet. The prediction problem is illustrated in Figure 10. A population of neurons encodes the stimulus history into its spiking activity. If this population carries predictive information, it should be able to predict the future trajectory of the stimulus over some timescale.

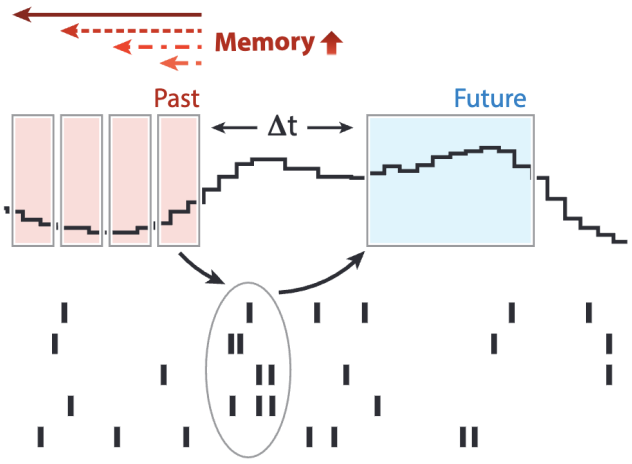

Figure 10: **The prediction problem**. (Top) A one-dimensional time-varying stimulus. (Bottom) A spike raster showing the activity of a neural population. Reproduced from Rust and Palmer (2021).

An optimal predictive representation $x = f(s_{\text{past}})$ should maximize predictability of the future subject to a constraint on memory of the past (Bialek et al., 2001):

$$f^* = \underset{f}{\operatorname{argmax}}\, \mathcal{I}[x; s_{\text{future}}] \tag{29}$$

$$\text{subject to } \mathcal{I}[s_{\text{past}}; x] \leq C. \tag{30}$$

For different choices of the capacity parameter C, we can chart an optimality frontier (Figure 11). This tells us the highest achievable predictive information for a given constraint on memory capacity. We can then use this as a quantitative benchmark for assessing predictive information in neural populations.

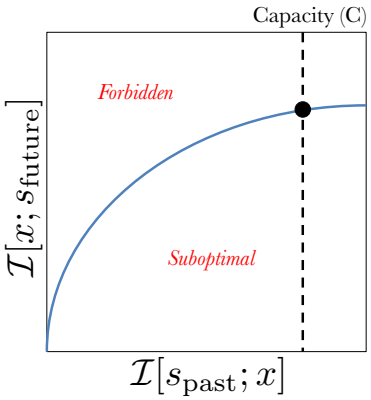

Figure 11: **The predictive information bottleneck**. The blue curve shows the optimality frontier. Representations that lie above the frontier are forbidden; representations that lie below the frontier are suboptimal. The dashed line shows a hypothetical capacity parameter.

4.2 *Predictive information in a retinal population*

The retina is one of the earliest stages of vision in which signatures of prediction are present. For example, Berry et al. (1999) found that a moving bar evokes a wave of activation in retinal ganglion cells that tracks the leading edge of the bar. This is remarkable given that firing latency of retinal ganglion cells to unpredictable flashes is around 50 ms. The population apparently learns to compensate for this delay by anticipating the bar position.

Applying the predictive information bottleneck to the retinal encoding of a moving bar, Palmer et al. (2015) found that the encoding was very close to optimal: given the amount of information carried by the retinal ganglion cells about past stimuli, information about future stimuli was almost perfectly on the optimality frontier (Figure 12). This result could not be explained by a conventional model of

See also Liu et al. (2021) for converging evidence with other motion patterns.

retinal ganglion cells based on linear filtering of the spatiotemporal stimulus followed by a spiking non-linearity; fitting this model to the neural data produced predictive information far below the optimality frontier.

4.3 *Predictive information in sensory cortex*

Moving into cortex, we can find even more impressive feats of prediction. For example, a moving spot evokes a sequential pattern of activity in V1; after repeated stimulus presentations, this sequential pattern is reproduced autonomously by a stationary spot of light at the starting point of the motion path (Xu et al., 2012; Ekman et al., 2017). Similarly, V1 neurons can "complete" sequences of stimuli with missing elements (Gavornik and Bear, 2014). Another study found that V1 activity during spatial navigation becomes increasingly predictive of upcoming stimuli (Fiser et al., 2016).

Singer et al. (2018) asked whether known properties of receptive fields in sensory cortex could be explained as the result of representations optimized for prediction. They trained simple neural networks to predict future auditory or visual stimuli based on recently presented stimuli. For visual stimuli, they fed the network the 7 most recent frames and predicted the next frame. For auditory stimuli, they fed the network the cochleagrams covering the last 200 ms and predicted the next 15 ms. Singer et al. compared the learned receptive fields to those of V1 and primary auditory cortex (A1), finding that the model reproduced many of the receptive field types observed experimentally, with a quantitative match superior to non-predictive models. They also found that the predictive ability of the trained networks was correlated with their match to neural data. These findings support the claim that representations in sensory cortex are optimized for prediction.

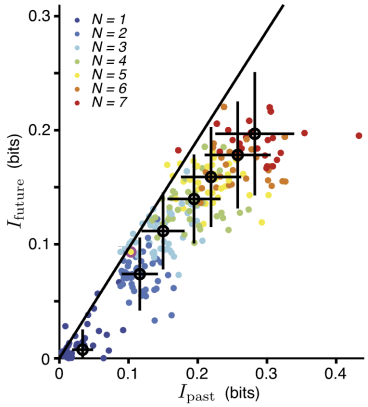

Figure 12: **Predictive information in a retinal population**. Color denote groups of cells of different sizes (N). The black line shows the optimality frontier. Adapted from Palmer et al. (2015).

A cochleagram is a time-frequency representation of auditory input that is matched to the filtering properties of auditory nerve fibers carrying sound information from the cochlea.

5 *Predictiveness vs. efficiency*

The attentive reader may have noticed a tension between the solution to the predictive information bottleneck problem (Section 4) and the predictive coding solution to the efficient coding problem (Section 2). Although both solutions involve predictions, what they do with these predictions is quite different. In the predictive information bottleneck, only the predictively useful information is kept; in contrast, predictive coding discards predictive information by only encoding prediction errors. Since sparse coding can arise from efficient coding (as we saw in the case of the retina), predictiveness may also sometimes be at odds with sparsity.

Although these different objectives seem fundamentally incompatible, it's important to keep in mind that efficient coding in fact requires that probability information is retained in some form—though not necessarily in the spiking activity. For example, the approach of Ganguli and Simoncelli (2014) assumes that the stimulus distribution is encoded in the set of tuning functions. In the approach of Srinivasan et al. (1982), the stimulus distribution is encoded in the inhibitory connections impinging on retinal ganglion cells. Furthermore, a downstream decoder reading out this information would need to have access to this information in a computationally useful format. Ganguli and Simoncelli (2014) show how a biologically plausible decoder can be constructed that reports the posterior mean.

Another point of view is that these different objectives may arise as special cases of a single unifying objective. Chalk et al. (2018) generalized the predictive information bottleneck problem by allowing prediction at some future time point $t + \Delta$ (where t is the current time point) to depend on neural activity in a window of length τ. When $\Delta < 0$, the goal is to reconstruct the stimulus history based on the neural activity $x_{t-\tau:t}$; the optimal solution is sparse, with highly selective tuning functions (e.g., to particular motion directions when trained on moving stimuli). When $\Delta > 0$, the goal is to predict future stimuli, and the code becomes distributed, with relatively non-selective tuning functions. Underlying this transition from sparse to distributed is the constraint that prediction (particularly for short Δ) requires rapid processing of stimuli. This is not possible with sparse codes, where each neuron only glimpses a slice of the stimulus. Thus, prediction latency imposes a strong constraint on sparsity.

Chalk et al. (2018) also discuss how the framework predicts different effects depending on τ, the capacity parameter C, and the statistical structure of the stimulus. In the interest of brevity, we do not discuss those aspects here.

6 *Conclusion*

While no single principle can explain all the relevant empirical phenomena, we have seen that a small set of principles has a remarkably wide scope. All of these principles are closely related to one another, even when they make qualitatively different predictions. Weaving through the different formalisms is the unifying idea that representations should be optimized to encode information that is useful for certain tasks (reconstruction, inference, prediction). In the following chapters, we will explore in greater detail how the brain carries out these tasks.

Study questions

1. Efficiency, sparsity, and prediction principles are partly complementary but sometimes contradictory. How might these principles be reconciled into a single unifying framework of perceptual representation? To what extent are they incompatible?
2. Representations are conceptualized here in terms of tuning functions. What are the limitations of this conceptualization when compared with a more dynamical view of neural computation?
3. The energy efficiency of the brain is remarkable (its power usage is comparable to a dim light bulb). What might we learn about the design of energy-efficient artificial systems from studying the brain?

References

Appelle, S. (1972). Perception and discrimination as a function of stimulus orientation: the "oblique effect" in man and animals. *Psychological Bulletin*, 78:266–278.

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, 3:213–251.

Atick, J. J. and Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, 4:196–210.

Attwell, D. and Laughlin, S. B. (2001). An energy budget for signaling in the grey matter of the brain. *Journal of Cerebral Blood Flow & Metabolism*, 21:1133–1145.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith, W., editor, *Sensory Communication*, pages 217–233. MIT Press.

Berry, M. J., Brivanlou, I. H., Jordan, T. A., and Meister, M. (1999). Anticipation of moving stimuli by the retina. *Nature*, 398:334–338.

Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463.

Chalk, M., Marre, O., and Tkačik, G. (2018). Toward a unified theory of efficient, predictive, and sparse coding. *Proceedings of the National Academy of Sciences*, 115:186–191.

Dodge, R., Travis, R., and Fox, J. (1930). Optic nystagmus: III. Characteristics of the slow phase. *Archives of Neurology & Psychiatry*, 24:21–34.

Ekman, M., Kok, P., and de Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, 8:15276.

Fiser, A., Mahringer, D., Oyibo, H. K., Petersen, A. V., Leinweber, M., and Keller, G. B. (2016). Experience-dependent spatial expectations in mouse visual cortex. *Nature Neuroscience*, 19:1658–1664.

Ganguli, D. and Simoncelli, E. (2010). Implicit encoding of prior probabilities in optimal neural populations. *Advances in Neural Information Processing Systems*, 23.

Ganguli, D. and Simoncelli, E. P. (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Computation*, 26:2103–2134.

Gavornik, J. P. and Bear, M. F. (2014). Learned spatiotemporal sequence recognition and prediction in primary visual cortex. *Nature Neuroscience*, 17:732–737.

Girshick, A. R., Landy, M. S., and Simoncelli, E. P. (2011). Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14:926–932.

Hromádka, T., DeWeese, M. R., and Zador, A. M. (2008). Sparse representation of sounds in the unanesthetized auditory cortex. *PLoS Biology*, 6:e16.

Laughlin, S. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung c*, 36:910–912.

Lennie, P. (2003). The cost of cortical computation. *Current Biology*, 13:493–497.

Lillywhite, P. and Laughlin, S. (1979). Transducer noise in a photoreceptor. *Nature*, 277:569–572.

Liu, B., Hong, A., Rieke, F., and Manookin, M. B. (2021). Predictive encoding of motion begins in the primate retina. *Nature Neuroscience*, 24:1280–1291.

Mansfield, R. (1974). Neural basis of orientation perception in primate vision. *Science*, 186:1133–1135.

Marcelja, S. (1980). Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America*, 70:1297–1300.

Nadal, J.-P. and Parga, N. (1994). Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in Neural Systems*, 5:565–581.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., and Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature*, 411:698–701.

Olshausen, B. A. and Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, 14:481–487.

Palmer, S. E., Marre, O., Berry, M. J., and Bialek, W. (2015). Predictive information in a sensory population. *Proceedings of the National Academy of Sciences*, 112:6908–6913.

Pitkow, X. and Meister, M. (2012). Decorrelation and efficient coding by retinal ganglion cells. *Nature Neuroscience*, 15:628–635.

Poo, C. and Isaacson, J. S. (2009). Odor representations in olfactory cortex: "sparse" coding, global inhibition, and oscillations. *Neuron*, 62:850–861.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25.

Prat-Carrabin, A. and Woodford, M. (2021). Bias and variance of the bayesian-mean decoder. *Advances in Neural Information Processing Systems*, 34:23793–23805.

Quiroga, R., Reddy, L., Kreiman, G., Koch, C., and Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, 435:1102–1107.

Rust, N. C. and Palmer, S. E. (2021). Remembering the past to see the future. *Annual Review of Vision Science*, 7:349–365.

Schuster, P. and Sigmund, K. (1983). Replicator dynamics. *Journal of Theoretical Biology*, 100:533–538.

Singer, Y., Teramoto, Y., Willmore, B. D., Schnupp, J. W., King, A. J., and Harper, N. S. (2018). Sensory cortex is optimized for prediction of future input. *elife*, 7:e31557.

Srinivasan, M. V., Laughlin, S. B., and Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 216:427–459.

Stark, L., Vossius, G., and Young, L. (1962). Predictive control of eye tracking movements. *IRE Transactions on Human Factors in Electronics*, (2):52–57.

Taylor, P. D. and Jonker, L. B. (1978). Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40:145–156.

van Hateren, J. H. (1992). Theoretical predictions of spatiotemporal receptive fields of fly LMCs, and experimental validation. *Journal of Comparative Physiology A*, 171:157–170.

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

Wei, X.-X. and Stocker, A. A. (2016). Mutual information, Fisher information, and efficient coding. *Neural Computation*, 28:305–326.

Xu, S., Jiang, W., Poo, M.-m., and Dan, Y. (2012). Activity recall in a visual cortical ensemble. *Nature Neuroscience*, 15:449–455.

Yoshida, T. and Ohki, K. (2020). Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nature Communications*, 11:872.