Computational Brain & Behavior
https://doi.org/10.1007/s42113-019-00028-z

Neuronal Firing Rate As Code Length: a Hypothesis

Ning Qian[1] · Jun Zhang[2]

Abstract
Many theories assume that a sensory neuron's higher firing rate indicates a greater probability of its preferred stimulus. However, this contradicts (1) the adaptation phenomena where prolonged exposure to, and thus increased probability of, a stimulus reduces the firing rates of cells tuned to the stimulus; and (2) the observation that unexpected (low probability) stimuli capture attention and increase neuronal firing. Other theories posit that the brain builds predictive/efficient codes for reconstructing sensory inputs. However, they cannot explain that the brain preserves some information while discarding other. We propose that in sensory areas, projection neurons' firing rates are proportional to optimal code length (i.e., negative log estimated probability), and their spike patterns are the code, for useful features in inputs. This hypothesis explains adaptation-induced changes of V1 orientation tuning curves and bottom-up attention. We discuss how the modern minimum-description-length (MDL) principle may help understand neural codes. Because regularity extraction is relative to a model class (defined by cells) via its optimal universal code (OUC), MDL matches the brain's purposeful, hierarchical processing without input reconstruction. Such processing enables input compression/understanding even when model classes do not contain true models. Top-down attention modifies lower-level OUCs via feedback connections to enhance transmission of behaviorally relevant information. Although OUCs concern lossless data compression, we suggest possible extensions to lossy, prefix-free neural codes for prompt, online processing of most important aspects of stimuli while minimizing behaviorally relevant distortion. Finally, we discuss how neural networks might learn MDL's normalized maximum likelihood (NML) distributions from input data.

Keywords Encoding · Decoding · Bayesian universal code · Shannon information · Rate-distortion · Sparse coding · Image statistics

Introduction

What do neuronal activities mean? This fundamental question on the nature of neural codes has been pondered upon extensively since early recordings of nerve impulses (Adrian 1926). In this paper, we first review two major categories of theories for interpreting responses of sensory neurons. The first category views a sensory neuron's firing rate as indicating the probability that its preferred stimulus is present in the input. The second category contends that sensory neurons provide an efficient or predictive representation of input stimuli, with the goal of reconstructing the input stimuli. We evaluate these and other related theories and point out that they contradict some major experimental facts and sometimes contradict each other. To resolve these contradictions, we propose the new hypothesis that in sensory areas, firing rates of projection neurons are proportional to *optimal code lengths* for coding useful features in input stimuli. We show that this hypothesis, which implies that neurons' spike patterns are the actual codes, can naturally explain observed changes of V1 orientation tuning curves induced by orientation adaptation.

Core to our new framework for neural codes is the concept of optimal universal codes (OUCs) arising from modern minimum description length (MDL) principle (Grunwald 2007; Rissanen 2001); it differs from older prescriptions of MDL used in some previous neural models. OUCs balance data explanation and model complexity to avoid over-fitting. We argue that the MDL goals of maximizing regularity extraction for optimal data compression, prediction, and communication are consistent with the goals of neural processing and

✉ Ning Qian
nq6@columbia.edu

Jun Zhang
junz@umich.edu

[1] Department of Neuroscience, Zuckerman Institute, and Department of Physiology & Cellular Biophysics, Columbia University, New York, NY 10027, USA

[2] Department of Psychology and Department of Mathematics, University of Michigan, Ann Arbor, MI 48109, USA

transmission of input stimuli. Indeed, since compression must rely on regularities in the data, the degree of compression measures the degree of data understanding. Compared with previous theories of efficient and predictive coding, a distinctive feature of OUCs in modern MDL is that regularity extraction is relative to a model class (such as a family of cells indexed by their preferred stimulus properties). Consequently, OUCs match the brain's purposeful information processing, which cannot be achieved by reconstruction of input stimuli assumed in previous theories. Different areas along a sensory hierarchy may implement different model classes for understanding different levels of regularities in stimuli. To explain the brain's *selective* information processing, we discuss possible extensions of the standard MDL from lossless data compression to a lossy version, and to the inclusion of top-down modulation that prioritizes neural transmission of more behaviorally important information. We suggest that neural codes must be prefix free so that the next stage of processing can interpret incoming spikes online as soon as they are being received. We also discuss how neural networks might learn and tune a key OUC of MDL, namely the normalized maximum likelihood (NML) distribution, by sampling input stimuli.

Evaluations of Major Theories of Neuronal Coding

Firing-Rate-As-Probability Theories

An early notion of neural coding is that a sensory neuron's firing rate reflects the strength of stimulation (Adrian 1926), a higher rate indicating a stronger stimulation. In his neuron doctrine, Barlow (1972) casts this notion probabilistically by stating that "[h]igh impulse frequency in such neurons corresponds to high certainty that the trigger feature is present." The idea is made most explicit in the population-average method for decoding neuronal activities (Georgopoulos et al. 1986). For example, to decode a perceived orientation $\hat{\theta}$ from the firing rates, $\{r_i\}$, of a set of cells with preferred orientations $\{\theta_i\}$, the method assumes

$$\hat{\theta} = \frac{\sum_i r_i \theta_i}{\sum_i r_i} = \sum_i p_i \theta_i, \quad \text{where } p_i \equiv \frac{r_i}{\sum_j r_j} \tag{1}$$

implying that cell i's firing rate r_i, normalized by the sum of all cells' firing rates $\sum_j r_j$, is the probability p_i of its preferred orientation θ_i present in the input, and that the perceived orientation is the expectation of the probability distribution.

Other methods for interpreting neuronal responses have also been proposed. For instance, the maximum-likelihood method (Paradiso 1988) assumes that for a given stimulus orientation θ_s, the responses $\mathbf{r}$ of a set of orientation-tuned cells follow the distribution $p(\mathbf{r}|\,\theta_s)$. When a particular set of responses $\{r_i\}$ is observed, $p(\{r_i\}|\,\theta_s)$ can be viewed as a distribution function of θ_s (the likelihood function) parameterized by $\{r_i\}$, and the perceived orientation is assumed to be the θ_s that maximizes the likelihood:

$$\hat{\theta} = \arg\max_{\theta_s} p(r_i|\theta_s). \tag{2}$$

By definition, cell i's response r_i is more likely to be large when stimulus orientation θ_s is closer to the cell's preferred orientation θ_i. Then, within the response range, a large (or small) response r_i implies a large (or small) likelihood p that the stimulus orientation θ_s equals cell i's preferred orientation θ_i: $p(\text{large } r_i|\,\theta_s = \theta_i) > p(\text{small } r_i|\,\theta_s = \theta_i)$. In other words, the likelihood that a cell's preferred orientation is present in the input stimulus increases monotonically with the cell's response, similar to the population-average method (which posits the special case of a linear relationship). Correlations among different cells' responses do not change the conclusion because the correlations are significant (and positive) only among cells with similar preferences (Nowak et al. 1995; van Kan et al. 1985). One could simply group the cells with similar preferences and argue that a larger group response implies a larger likelihood that the group's mean preferred orientation is present in the stimulus.

If the prior probability distribution, $p(\theta_s)$, of stimulus orientation is known, then its product with the likelihood function determines the posterior distribution of θ_s given the responses $\{r_i\}$, according to the Bayes rule. The Bayesian method (Sanger 1996) posits that the perceived orientation is the θ_s that maximizes the posterior probability:

$$\hat{\theta} = \arg\max_{\theta_s} p(r_i|\theta_s)p(\theta_s). \tag{3}$$

Prior distributions are typically well behaved (smoothly varying) (Weiss et al. 2002; Yuille and Kersten 2006) and thus will not drastically change the aforementioned relationship between r_i and θ_s in the likelihood function. More importantly, although prior and likelihood are conceptually different, physiologically, the priors that the brain has learned must be reflected in relevant neuronal responses (Atick and Redlich 1990; Zhaoping 2014) and thus already included in the relationship between r_i and θ_s for the likelihood function. Short-term fluctuations of responses to temporary priors (e.g., adaptation to a particular θ_s) that are not yet learned by downstream neurons may distort the relationship between r_i and θ_s, but over longer time scales, these fluctuations and distortions average out. Therefore, Bayesian decoders must generally retain the property that large and small responses r_i indicate,

respectively, large and small probabilities that the stimulus orientation θ_s equals the cell's preferred orientation θ_i.

In sum, many neural-tuning-based theories, including the well-known population-average, maximum likelihood, and Bayesian decoders, assume that a cell's firing rate is monotonically related to the probability that its preferred stimulus is present in the input. For simplicity, we refer to this assumption as the *firing-rate-as-probability* assumption. In the population-average method, a cell's firing rate is directly proportional to the probability of its preferred stimulus. In maximum-likelihood and Bayesian methods, firing rates parameterize probability distributions of stimuli but a cell's higher firing rate still generally indicates a greater probability of its preferred stimulus.

Despite its intuitive appeal, the firing-rate-as-probability assumption contradicts two major classes of phenomena. First, adaptation to, say, vertical orientation, must increase the brain's estimated probability for vertical orientation; yet the cells tuned to vertical orientation reduce their firing rates to that orientation after the adaptation (Blakemore and Campbell 1969; Fang et al. 2005). (The cells' responses to other orientations may increase (Dragoi et al. 2000, Felsen et al. 2002, Teich and Qian 2003), an observation that we consider in the "Simulating Adaptation-Induced Changes of V1 Orientation Tuning Curves" section but does not affect the current discussion.) Second, salient stimuli capture our attention and increase neuronal firing rates (Gallant et al. 1998; Gottlieb et al. 1998; Itti and Koch 2001; Zhaoping 2002); yet these are low-probability stimuli such as sudden onset of light or sound, instead of high-probability stimuli such as constant background stimulation. Indeed, if a salient stimulus occurs frequently, it will gradually lose its saliency and evoke less response because the brain adapts to it. The firing-rate-as-probability assumption predicts the opposite.

Efficient/Predictive Coding Theories

A second prominent category of theories assumes that neurons in a visual area build an efficient or predictive code of input stimulus with the goal of reconstructing the retinal image according to some optimality criteria (Atick and Redlich 1990; Barlow and Foldiak 1989; Bell and Sejnowski 1997; Harpur and Prager 1996; Olshausen and Field 1996; Rao and Ballard 1999; Zhaoping 2014). Different theories optimize different cost functions which typically contain a reconstruction error term and a term encouraging a desired code property such as de-correlation, independence, or sparseness. The rationale is that by forcing the models to reconstruct retinal images through efficient representations, they can discover useful statistical regularities in the images.

Many efficient/predictive coding theories focus on reproducing important properties of receptive fields without explicitly specifying what neuronal activities represent. One of the theories does specify that activities of neurons projecting to the next stage represent the error between the actual input and the input predicted by the next stage (Rao and Ballard 1999). This assumption is consistent with the adaptation and bottom-up-attention phenomena mentioned above if it is further assumed that stimuli with larger and smaller probabilities are reconstructed/predicted more and less accurately, respectively. However, it is unclear how it may explain a variety of adaptation-induced tuning changes ("Simulating Adaptation-Induced Changes of V1 Orientation Tuning Curves" section). More importantly, by aiming to reconstruct input stimuli, these theories neglect the empirical fact that the brain processes inputs to extract behaviorally relevant information while ignoring irrelevant one; the best example is perhaps the change-blindness demonstrations (Pashler 1988): people are unaware of large, blatant changes between successively flashed images unless their attention is directed to the changes. Moreover, there is a well-known conundrum with the efficient/predictive coding theories: if, for example, the purpose of the visual system is to produce an efficient code that reconstructs retinal images, why, then, are there so many more cells in the visual cortices than on retina? In other words, how could such a great increase of the number of cells involved in coding the same information be called efficient?

To address this cell-number conundrum of the efficient/predictive coding theories, Olshausen and Field (1996) proposed that the brain needs a large number of cells to produce a sparse (and over-complete) representation. With appropriate total numbers of units in learning networks, sparse coding models have been successful in explaining some important receptive field properties (Olshausen and Field 2004). It has been argued that sparse coding with a large number of cells is more energy efficient (Balasubramanian et al. 2001; Olshausen and Field 2004), and sparsely firing neurons can be constructed from an integrate-and-fire mechanism (Yenduri et al. 2012). However, maintaining a large number of cells and their connections incur a great cost. We will argue that a large number of cells are needed for extracting various behaviorally relevant features from inputs, rather than for input reconstruction. Our MDL-based framework suggests that the brain attempts to minimize neuronal firing rates (i.e., code length, see "A New Framework for Neural Codes" section) and thus the number of cells firing at a given time, and in this sense, is consistent with the sparse coding theory.

Other Theories

An approach related to the efficient coding theories is to measure as many types of natural-image statistics as possible and use the measurements to explain and predict perceptual phenomena and neuronal responses (Field 1987; Geisler et al. 2001; Motoyoshi et al. 2007; Sigman et al. 2001; Simoncelli and Olshausen 2001; Yang and Purves 2003). For example,

the perception of a line segment is enhanced when it is smoothly aligned with neighboring segments (Li and Gilbert 2002). This is known as the Gestalt principle of good continuation and can be explained by the statistical result that nearby contour segments tend to form a smooth continuation in the real world (Geisler et al. 2001; Sigman et al. 2001). Although extremely powerful in accounting for many perceptual observations that would otherwise be puzzling, these studies either avoid specifying what neuronal responses represent or use the firing-rate-as-probability assumption and thus inherit its problems discussed above. Indeed, given the image-statistics-based explanation of the Gestalt principle of good continuation, it is unclear why many V1 cells *reduce* firing rates when a contour extends beyond their classical receptive fields (Bolz and Gilbert 1986; Hubel and Wiesel 1968; Li and Li 1994).

Normalization models were originally proposed to explain nonlinear response properties of V1 simple cells (Albrecht and Geisler 1991; Heeger 1992). These nonlinearities include contrast saturation and interactions among multiple stimuli. The models have since been applied to other visual areas (Simoncelli and Heeger 1998) and to attentional modulation of visual responses (Reynolds and Heeger 2009) and are regarded as a canonical module of neural computation (Carandini and Heeger 2012). The main assumption is that the actual response r_i of a cell is equal to its linear-filter response R_i normalized by a regularization constant σ plus the pooled linear-filter responses from all cells tuned to the full range of stimulus parameters:

$$r_i = r_0 \frac{R_i^n}{\sigma^n + \sum_j R_j^n} \tag{4}$$

The power index n introduces additional nonlinearity as suggested by typical contrast saturation curves. r_0 is a scaling constant. There is also a temporal version of the models (Carandini et al. 1997).

The rationale behind these models is that the normalization factor provides a gain control mechanism to allow a cell's limited dynamic range encode a broad range of stimulus intensity. Given their phenomenological nature and the small number of free parameters, these models are impressive in explaining neuronal responses across a broad range of systems and conditions (Carandini and Heeger 2012). However, in an extracellular recording test of the model, the constant model parameters for a given V1 cell have to be adjusted to fit data from different stimulus conditions (Carandini et al. 1997). Moreover, a circuit that implements normalization via divisive shunting inhibition (Carandini et al. 1997) is not supported by intracellular recording data (Anderson et al. 2000). Additionally, without modifications, normalization models cannot explain many interesting spatial interaction phenomena. For example, a V1 or MT cell's response to its preferred orientation/direction in the classical receptive field center is suppressed when the surround has the same orientation/direction, but the suppression becomes weaker, or even turns into facilitation, when the surround orientation/direction differs from that of the center (Allman et al. 1985; Levitt and Lund 1997; Li and Li 1994; Nelson and Frost 1978). Instead of pooling cross all cells, the normalization factor has to be tailored to select different subgroups for different situations. Other considerations, such as natural-image statistics, have to be used to justify such selection. Finally, normalization models focus on reproducing firing rates without specifying what they represent (probabilities, code lengths, or something else).

A set of studies aims to reproduce spiking statistics of real neurons. With the increasing availability of multi-single-unit recording data, much of this line of research focuses on how to capture second- and higher-order statistical relationships among multiple neurons (Ganmor et al. 2011; Schneidman et al. 2006; Shlens et al. 2006). While these studies provide useful hints on neural code, they do not in themselves address the nature of neural code. For example, knowing that two neurons have correlated responses to a stimulus does not immediately reveal the coding principle behind such correlation or what firing patterns represent.

There are inconsistencies among extant theories. For example, the firing-rate-as-probability hypothesis is incompatible with the efficient/predictive coding hypothesis: the former assumes that projection neurons transmit stimulus probability distributions (or their parameterizations) from one area to another to enable optimal inference based on products of the distributions, whereas the latter implies that the probability distributions be used to code stimuli efficiently for transmission and that projection neurons transmit reconstruction errors. As another example, a proposed implementation of optimal Bayesian inference using parameterized probability distributions (Ma et al. 2006) assumes that neurons sum up the firing rates they receive, contradicting nonlinear summation of real neurons emphasized by normalization models (Albrecht and Geisler 1991; Heeger 1992).

Stocker and Simoncelli (2006) noted that if adaptation to an orientation (adaptor) increases its prior probability, then a Bayesian framework predicts that a subsequently presented test orientation be attracted to the adaptor, contradicting the observed repulsive aftereffect (Gibson and Radner 1937; Meng and Qian 2005). They proposed that adaptation reduces noise in the likelihood function instead of increasing the prior probability of the adapting stimuli. However, the assumption that long exposure to a stimulus does not change its probability is at odds with frequentist probability definition. It also contradicts Bayesian probability definition as it asserts that subjective probability is never updated by prior experience. Moreover, if adaptation to a stimulus does not change its probability, then why should the brain adapt to natural-image statistics, an assumption used in numerous studies?

Additionally, the proposal does not save the firing-rate-as-probability assumption because it does not explain why the cells tuned to the adapting stimuli reduce their firing rates after the adaptation. To save the assumption, one would have to posit, unreasonably, that adaptation to a stimulus actually *reduced* its probability.

Although a typical, prospective Bayesian model incorrectly predicts attractive aftereffects, a recent study suggests that repulsive aftereffects could result from retrospective Bayesian decoding in working memory (Ding et al. 2017). According to this new framework, after all task-relevant features are encoded and enter working memory, the brain decodes more reliable, higher-level features first and uses them as priors to constrain the decoding of less reliable, lower-level features, producing repulsion in the process. In other words, although a prior from the adaptor may predict attraction, a different prior from high-level decoding could override it and generate a net repulsion.

It seems fair to summarize the state-of-the-art theories of neuronal coding as the story of the Blind Men and Elephant: each theory captures some important aspects of neural coding and appears plausible in some ways, but it is unclear how they fit together coherently.

A New Framework for Neural Codes

Understanding neural codes is an ambitious task that is unlikely to be accomplished in foreseeable future. Nevertheless, as a small step, we would like to outline a framework, based on the modern MDL principle, which aims to resolve the issues while retaining the strengths of the previous theories. In the following, we will first review the modern MDL principle briefly. We will then argue that when this principle is adopted for neural coding, it leads to our main hypothesis that firing rates of projection neurons are proportional to optimal lengths for coding useful features in the stimuli. This firing-rate-as-code-length hypothesis is fundamentally different from the firing-rate-as-probability or firing-rate-as-prediction-error hypotheses discussed above. We will apply this hypothesis to explain various changes of V1 orientation tuning curves induced by orientation adaptation. The hypothesis is also consistent with bottom-up attention because rare (low probability) stimuli should have a long code length, i.e., evoke high firing rate. We further suggest that the MDL framework could be modified to include top-down attention. Since the firing-rate-as-code-length hypothesis implies that spiking patterns are the actual code for useful features in the input, we will speculate on the nature of the code, particularly the prefix-free and lossy properties. Finally, we will discuss how a key distribution from the MDL principle could be learned and tuned as input stimuli are sampled.

An Overview of Modern MDL and OUC

We propose that the modern MDL principle (Barron et al. 1998; Grunwald 2007; Grunwald et al. 2005; Myung et al. 2006; Rissanen 1996; Rissanen 2001), built on the concept of OUC (in the form of NML distribution and related codes), provides a viable framework for understanding neural codes. This principle, different from some similarly or identically named theories, was developed for model-class selection, regression, and prediction by maximizing regularity extraction from data. In this section, we briefly review modern MDL.

Our overview of MDL follows Grunwald (2007). Intuitively, understanding a piece of data means extracting regularities in the data that enable prediction of other data drawn from the same source (generalization). And since regularity is redundancy, regularity extraction can be measured by data compression. Thus, to best understand a piece of data is to find a model (i.e., a probability distribution) that minimizes description length of the data. (A model expresses a relationship in the data, which can always be cast as a probability distribution by adding a proper noise distribution.) To avoid over-fitting, the model complexity should also be taken into account. The MDL principle provides a practical way of achieving these goals.

More formally, if the probability mass function $P(x)$ of data samples x's is known, then the expected code length is minimized when the code for x has a length (Shannon 1948):

$$L(x) = -\log P(x) \tag{5}$$

This is a consequence of the Kraft–McMillan inequality that relates code lengths and probability distributions and the information inequality

$$-\sum_x P(x)\log P(x) < -\sum_x P(x)\log Q(x) \tag{6}$$

for any probability mass function $Q(x) \neq P(x)$. Intuitively, Eq. 5 assigns short and long codes to frequent and rare data samples, respectively, thus minimizing the average code length. Since one can always find a code with length approaching that of Eq. 5, the terms "code" and "probability distribution" are often used interchangeably.

In reality, when a piece of data (e.g., a retinal image) is received, its probability is unspecified. The best one can do is to use any prior knowledge, experience, or belief about the data generation process to produce a model M such that according to M, data sample x has a probability $P(x|M)$. Then according to this model, the code for x should have a length $L(x|M) = -\log P(x|M)$. To take the model complexity into account, one may use the length $L(M)$ of coding M to represent its complexity and seek a model, among a class of models $\mathcal{M}$, that minimizes the total code length:

$$L = L(x|M) + L(M) \tag{7}$$

as the best description of the data. This is indeed an MDL principle Rissanen (1978) proposed first, now referred to as the old or crude two-part MDL (Grunwald 2007) and used by Rao and Ballard (1997, 1999). A major problem is that there is no objective way of assigning a probability to M (and all other models in the class $\mathcal{M}$). Consequently, one could assign a given M different probabilities and thus different code lengths, rendering Eq. 7 arbitrary. Although one could choose $L(M)$ sensibly for a given situation and obtain meaningful results with Eq. 7, this approach is ad hoc.

Rissanen (2001) then developed the modern or refined MDL to overcome this arbitrariness in Eq. 7. Consider a model class $\mathcal{M}$ consisting of a finite number of models parameterized by the parameter set θ. For a given piece of data x, each model in the class prescribes it a probability $P(x|\theta)$ and thus a code with length $-\log P(x|\theta)$. The model $\hat{\theta}(x)$ that compresses the data x most is the one giving the data maximum likelihood $P\left[x|\hat{\theta}(x)\right]$, with code the length $L\left[x|\hat{\theta}(x)\right] = -\log P\left[x|\hat{\theta}(x)\right]$. However, this degree of compression is unattainable because in this scheme, different inputs would be encoded by different probability distributions (i.e., different M's in the model class $\mathcal{M}$), and the next stage could not consistently use or interpret the encoded message. The solution relies on the concept of a universal code: a single probability distribution $\overline{P}(x)$ defined for a model class $\mathcal{M}$ such that for any data x, the code for x is almost as short as $L\left[x|\hat{\theta}(x)\right]$, with the difference (termed regret) bounded in some way. The two-part code defined by Eq. 7 is actually a universal code because one can use a uniform distribution to code every model in $\mathcal{M}$ with equal probability $1/m$ so that the regret is bounded by $\log m$ where $m = |\mathcal{M}|$ is the number of models in $\mathcal{M}$. However, there are other, better universal codes. In particular, there is an optimal universal code (OUC) that minimizes the worst-case regret and avoids assigning an arbitrary distribution to $\mathcal{M}$. This so-called minimax optimal solution is the NML distribution:

$$P_{NML}(x) = \frac{P\left[x|\hat{\theta}(x)\right]}{\sum_y P\left[y|\hat{\theta}(y)\right]} \tag{8}$$

where the summation is over the data sample space (Fig. 1). With this distribution, the regret is the same for all data sample x and is given by:

$$\begin{aligned}\text{regret}_{NML} &\equiv -\log P_{NML}(x) + \log P\left[x|\hat{\theta}(x)\right] \\ &= \log \sum_y P\left[y|\hat{\theta}(y)\right]\end{aligned} \tag{9}$$

which is the log of the denominator in Eq. 8. Importantly, this expression also provides a natural definition of the model-class complexity:

$$\text{COMP}(M) \equiv \log \sum_y P\left[y|\hat{\theta}(y)\right] \tag{10}$$

because the summation indicates how many different data samples can be well explained by the model class. The more data samples the model class can explain well (i.e., large $P\left[y|\hat{\theta}(y)\right]$ for many data y's), the more complex the model class is. Thus, the numerator and denominator of the NML distribution in Eq. 8 represent how well the model class fits a specific piece of data and how complex the model class is, respectively.

There are other universal codes, one of which is Bayesian universal code with Jeffery's prior which approximates NML. In the following, we often use NML to represent OUC for simplicity but using other related codes will not change our conclusion.

The optimal universal code in the form of NML establishes the modern MDL principle for model-class selection: given a piece of data and multiple, competing model-classes, the one that produces the maximum NML probability explains the data best (Grunwald et al. 2005; Myung et al. 2006). This MDL principle has been extended to cases where the number of models in a class is not finite and COMP($\mathcal{M}$) and NML may not be defined (Grunwald 2007). We will not discuss those extensions because the number of cells in a brain area, and thus the number of models in a model class, is always finite. In this case, COMP($\mathcal{M}$) and NML are well defined even when input sample space is continuous (e.g., orientation). In fact, the sum (or integration for continuous input spaces) in COMP($\mathcal{M}$) is always smaller than or equal to the number of models in the class (see Fig. 1 caption):

$$\sum_y P\left[y|\hat{\theta}(y)\right] = m - \sum_{y,j} P\left[y|\theta_j \neq \hat{\theta}(y)\right] \tag{11}$$

We finally note that because the NML distribution is defined for a model class, regularity extraction and data compression in the MDL framework are relative to a model class. The true model that produces the data does not have to be a member of a model class in order for the model class to extract useful regularities. Different model classes extract different regularities. We will return to this point later.

An MDL-Based Framework for Neural Coding

Using the MDL concepts reviewed above, we start by assuming that each processing level of the brain implements many model classes, each class in the form of a set of cells tuned to a range of input properties (Fig. 2). For example, in area V1, the set of cells tuned to different orientations (Hubel and Wiesel 1968) can be viewed as forming a model class parameterized by the cells' preferred orientation. Different model classes are concerned with different properties of the input. Since some cells are simultaneously tuned to multiple properties (e.g.,

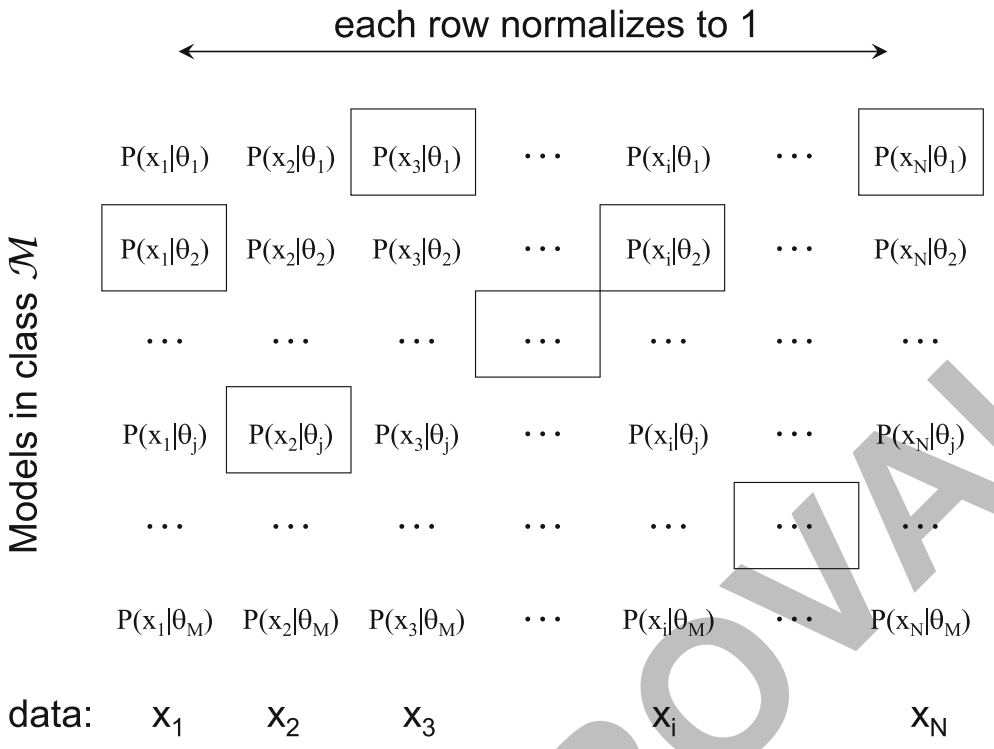

Fig. 1 The normalized maximum likelihood (NML) distribution for a model class $\mathcal{M}$. The models in the class, $P(\cdot|\theta)$, are parameterized by the parameter set θ. The xs in the bottom row represents all possible data samples. Each of the other rows represents the probability mass function of a given model (a fixed θ) for all data, and thus sums to 1 (this remains true for probability density functions of continuous data). Each column represents different probabilities (likelihoods) assigned to a given piece of data x_i by different models (different θ's). The model that gives the maximum likelihood is indicated by a box, and it is $\theta \equiv \hat{\theta}(x_i)$ by definition. The maximum likelihoods (the terms in the boxes) may not sum to 1 because they are from different models. However, they can be normalized by the sum to produce a proper probability mass function, which is the NML distribution in Eq. 8. To understand Eq. 11, note that the three terms of the equation are, respectively, the sums of the boxed terms, the sum of all terms, and the sum of the non-boxed terms

orientation, disparity, and motion direction), there are overlaps among cells in different model classes.

Each processing level strives to extract regularities from the input and thus should use the MDL principle to balance input explanation and model-class complexity. Different model classes at a processing level extract different (possibly overlapping) regularities that are behaviorally relevant. For example, motion-selective and color-selective cells in V1 form two model classes. If the motion and color of a stimulus are both relevant to the current behavioral task (e.g., catching a flying, red ball), then V1 needs to use both model classes simultaneously. (This is different from traditional applications of MDL to model-class selection, which pick only one model class with the largest NML probability.) Along a processing hierarchy, higher-level areas extract more complex regularities based on simpler regularities extracted at lower levels, suggesting that the MDL principle should be applied hierarchically. For instance, V1 cells may use oriented segments in retinal image to compress data, and the face cells in IT may compress inputs further by exploring regular face configuration and view-independent representation of face identity.

Regularity extraction in the MDL framework is relative to a model class, and as such, can be viewed as processing, rather than reconstructing, inputs. Consider a class of cells sensitive to various contrast ranges, each cell responding to input images according to the difference between the luminance levels in the center and surround of its receptive field. These center-surround cells can extract the useful regularity that luminance contrasts likely delineate object boundaries under changing lighting conditions. However, they would be poor at reconstructing the center and surround luminance values separately because their responses depend only on the difference of the values. Similarly, disparity-selective cells form a model class that codes the displacement between an object's left and right retinal images while largely discounting many other aspects of the images (such as the difference in contrast magnitude) (Qian 1994). This model class focuses on the useful relationship between an object's disparity and its distance from the fixation point (Qian 1997) but would have difficulty reconstructing other aspects of the two images. Generally, regularity extraction by a class of cells emphasizes certain relevant input dimensions for specificity while ignoring other, irrelevant dimensions for invariance. In this sense, it can be better viewed as behaviorally relevant processing than accurate input reconstruction. Thus, the large number of cells in the cortex is needed to process, not reconstruct, the inputs. This

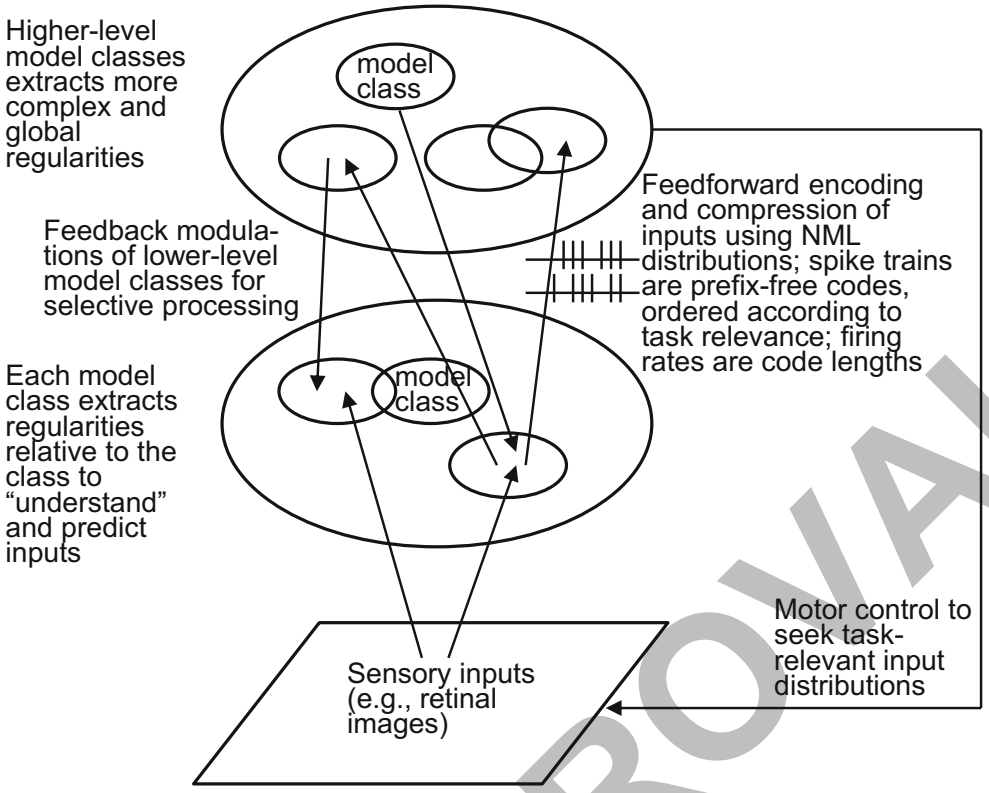

Fig. 2 MDL-based framework for neural coding. Large ovals represent brain areas along a processing hierarchy; only two processing levels are shown. Each small oval represents a model class devoted to extracting a certain stimulus regularity; for example, a model class can be a set of V1 cells parameterized by their preferred orientations. Core distinctions between our framework and many other existing ones in interpreting physiology and anatomy include (i) firing rates of projection neurons represent the code lengths of inputs, instead of the probability distributions (or their parameterization) of inputs; (ii) each model class can predict inputs based on the regularity it extracts, instead of relying on predictions from a higher-level area; (iii) feedback connections from higher-level areas modify lower-level model classes to selectively process inputs according to the current task or goal; and (iv) spike trains of projection neurons are a prefix-free code based on an NML distribution. We hypothesize that the process of regularity extraction (as measured by data compression) through the hierarchy *is* the process of "understanding" the input

avoids the cell-number conundrum of previous efficient/predictive coding theories.

Where do model classes in the brain (i.e., sensory cells with various response properties) come from? We assume that the response properties are learned via evolutionary and developmental processes and tuned by experiences to serve functions of the brain and to increase survival. Although low-level visual responses can be explained by image statistics, we suspect that an understanding of neuronal responses across the visual hierarchy must take behavioral tasks into account. This is consistent with recent comparisons between layers in deep neural networks and stages along the visual hierarchy: networks with better performances (for classification tasks) also explain visual responses better (Khaligh-Razavi and Kriegeskorte 2014; Yamins et al. 2014). It is possible that a model class and its NML distribution are learned together (see "NML and Learning" section).

Consistent with the MDL philosophy, a model class does not have to contain the "true" generative model of the environmental stimuli in order to be useful. For example, the brain does not need to know the exact optics of image formation to see, or the exact Newtonian mechanics to move. In fact, it is well known that exact knowledge of optics or Newtonian mechanics is insufficient to see or move because vision and motor-control problems faced by the brain are ill-posed mathematically (Flash and Hogan 1985; Poggio et al. 1985; Tanaka et al. 2006) and the brain has to make additional assumptions (in the form of regularities to be extracted by model classes, according to MDL) to solve the problems. An OUC does not have to be (and usually *is not*) a member of the model class. The brain merely approximates the "rules" underlying environmental stimuli through an optimal encoding strategy relative to a model class.

Regularity extraction by a model class is essential not only for input processing but also for input compression to afford efficient information transmission from one level to the next. The MDL principle solves these problems together using OUCs, and the solution is the NML distribution (or related distributions) for a model class (Eq. 8). It is

natural to assume that the brain uses an OUC (of a model class) to encode information for transmission because it minimizes the worst-case code length for both efficiency and robustness. However, unlike previous efficient/predictive coding theories that aim to reconstruct the input, here, efficiency is relative to a model class serving a function of the brain. As we show in the "Firing-Rate-As-Code-Length Hypothesis, Adaptation, and Bottom-up Attention" section, this difference leads to completely different interpretations of projection neurons' firing. Finally, input explanation and model-class complexity are balanced in NML (its numerator and denominator, respectively) to extract regularity and avoid over-fitting. This is critical for input understanding, prediction, and generalization.

Firing-Rate-As-Code-Length Hypothesis, Adaptation, and Bottom-up Attention

The above formulation suggests that in each brain area, the pyramidal cells that project to the next level should spike according to the NML distribution (or a related OUC) to efficiently code useful features in inputs. *Thus, projection neurons' firing rates (spikes per unit time) is proportional to the code length, equal to the negative log probability of the distribution.* The code-length minimization then becomes firing-rate minimization. Since firing rates of a set of cells are related to the number of cells firing at a given time (analogous to ergodic assumption that time average equals ensemble average), the firing-rate minimization is consistent with sparse coding (Olshausen and Field 1996). A set of projection cells, instead of a single cell, is involved in coding an input for two reasons. First, a set of cells can transmit the most important aspect of the input instantly using their spike pattern at a given time whereas a single cell would need more time to transmit the same information as a sequence of spikes. (Each cell does fire a sequence of spikes, but as we will discuss in the "Lossy MDL and Prefix-Free Neural Code" section, we suggest that latter spikes encode less important aspect of the input instead of a temporal code of the most important aspect of the input.) Second, neurons are noisy and may become dysfunctional; using a set of cells improves the reliability and robustness of transmission.

The firing-rate-as-code-length hypothesis naturally accommodates neural adaptation and bottom-up attention phenomena. For adaptation, prolonged exposure to a stimulus (adaptor) transiently increases its probability in the corresponding NML distribution. For example, adaptation to stimulus x increases $P\left[x|\hat{\theta}(x)\right]$ in the numerator and its appearance in the sum of the denominator of Eq. 8, with the net effect of increasing $P_{\mathrm{NML}}(x)$ (while decreasing NML probability for other stimuli y). Consequently, the code length (firing rate) for the adapting stimulus decreases. Indeed, Eq. 5 suggests that firing-rate (code-length) change equals negative relative probability change:

$$\Delta L(x) = -\frac{\Delta P(x)}{P(x)} \tag{12}$$

We provide a more detailed analysis in the "Simulating Adaptation-Induced Changes of V1 Orientation Tuning Curves" section for orientation adaptation. For attention-grabbing salient stimuli, because they are unexpected, low-probability events, the code length (firing rate) is large.

We emphasize that our *firing-rate-as-code-length* assumption only applies to projection neurons which transmit information from one brain area to the next. The common *firing-rate-as-probability* assumption may apply to local interneurons or alternatively, a more implicit probability representation may be learned ("NML and learning" section). Once a probability distribution is computed in an area, whether it is the NML distribution of the MDL framework or the posterior distribution of the Bayesian framework, it should be used to minimize code length for efficient information transmission according to Eq. 5. We therefore suggest the following framework for conceptualizing neural processing: when sensory stimuli are processed along a hierarchy, each brain area receives inputs from the lower-level areas, provides new processing by using its own model classes to compute the corresponding NML probabilities of the inputs, and use these probabilities to encode and transmit the inputs to the next level. This encoding process is the process of understanding the inputs because it maximizes regularity extraction from, and compression of, the inputs, according to the model classes in the area.

Simulating Adaptation-Induced Changes of V1 Orientation Tuning Curves

Our formulation readily explains the observed response reduction for cells tuned to the adapted stimulus (Eq. 12). However, it is known that orientation adaptation produces additional changes to V1 orientation tuning curves (Dragoi et al. 2001; Dragoi et al. 2000; Felsen et al. 2002; Teich and Qian 2003). Some experimental data from Dragoi et al. (2000, 2001) are shown in Fig. 3. Define the two sides of a cell's pre-adaptation tuning curve as the near and far sides according to whether the side includes the adapted orientation or not (e.g., the left and right sides of the red tuning curve in Fig. 3b are the far and near sides, respectively, because the right side contains the adapted orientation indicated by the green arrow). Then, the adaptation-induced changes of orientation tuning curves can be summarized as follows. (1) Responses on the near side of a tuning curve decrease (Fig. 3a, b). (2) Responses on the far side of the tuning curves increase (Fig. 3a, b). (3) For cells whose preferred orientations are around the adapted

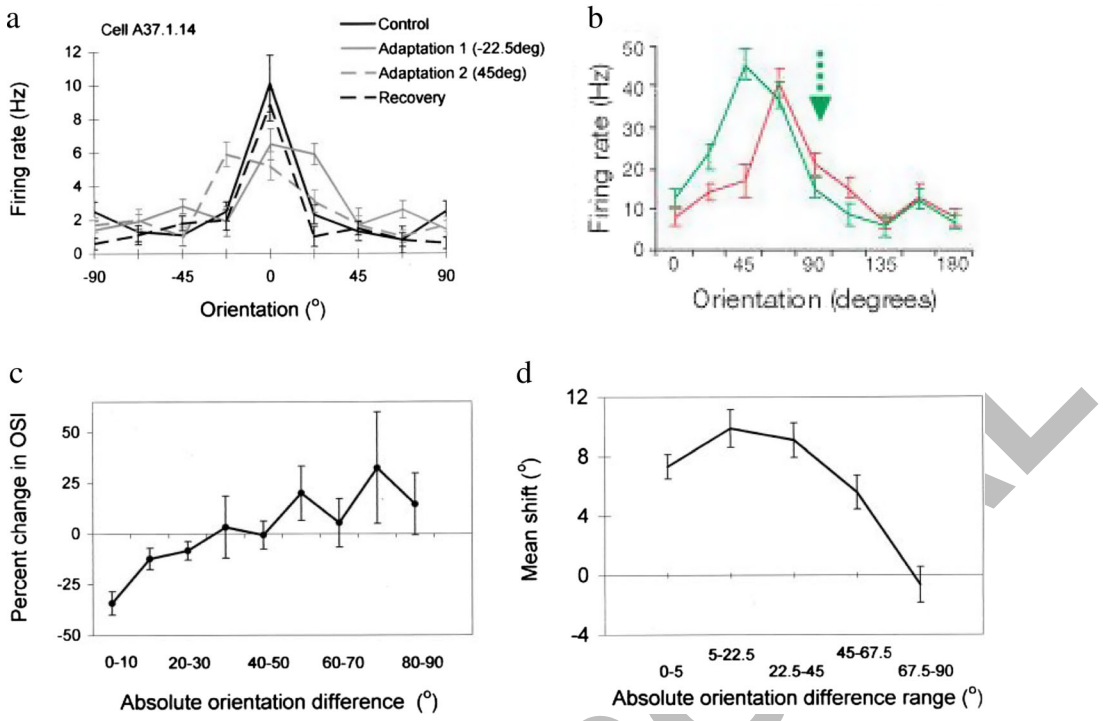

Fig. 3 Observed changes of V1 orientation tuning curves induced by adaptation. **a** The solid black curve represents the pre-adaptation tuning curve with the preferred orientation centered at 0°. The solid and dashed gray curves are the same cell's tuning curves after adaptation at − 22.5° and 45°, respectively. **b** The red and green curves represent a cell's pre- and post-adaptation tuning curves, respectively. The adapted orientation is indicated by the green arrow. The peak response after adaptation was even larger than before adaptation. **c** Adaptation-induced change of orientation selectivity index (OSI, see text for definition) as a function of the difference between the pre-adaptation-preferred orientation and the adapted orientation. Negative and positive OSI changes indicate increase and decrease of tuning width, respectively. **d** Adaptation-induced peak shift of tuning curves as a function of the orientation difference between the pre-adaptation-preferred orientation and the adapted orientation. Note that the orientation difference ranges for the first two data points are different from each other and from the remaining three points. Panels *a*, *c*, and *d* are from Dragoi et al. (2000) and panel *b* from Dragoi et al. (2001), with permissions

orientation, the peaks of their tuning curves shift away from the adapted orientation (Fig. 3a, b, d). (4) Also for cells whose preferred orientations are around the adapted orientation, their tuning widths become broader (Fig. 3a–c). (5) For cells whose preferred orientations are far away from the adapted orientation, their tuning widths become narrower (Fig. 3c). In Fig. 3c, cells' tuning widths are quantified by orientation selectivity index (OSI) defined as $\mathrm{OSI} = \sqrt{(\alpha^2 + \beta^2)}/\gamma$ where $\alpha = \sum_x r(x) \cos(2x)$, $\beta = \sum_x r(x) \sin(2x)$, $r(x)$ is the firing rate at the sampled stimulus orientation x, and $\gamma = \mathrm{mean}[r(x)]$. Large and small OSI indicate narrow and broad tuning widths, respectively.

We now demonstrate that the firing-rate-as-code-length hypothesis can explain all of these observed physiological changes. Consider a set of V1 cells (60 in our simulations) whose preferred orientations uniformly sample the full 180° range. Let cell i's preferred orientation be x_i and its firing rate in response to stimulus orientation x be $r(x, x_i)$. According to our firing-rate-as-code-length hypothesis, $r(x, x_i)$ should be proportional to the length $L(x)$ for coding x (Eq. 5). Additionally, the cell has an intrinsic orientation tuning function $t(x, x_i)$ according to the feedforward inputs it receives (Hubel and Wiesel 1968; Reid and Alonso 1995; Teich and Qian 2006). We therefore assume that the observed response $r(x, x_i)$ is a product of the code length and the tuning function:

$$r(x, x_i) = L(x)t(x, x_i) \tag{13}$$

Before adaptation, all orientations over the full range of π are equally probable so that $P(x) = P_0 = 1/\pi$ in Eq. 5, indicated by the flat red line in Fig. 4a (we neglect prior orientation bias here because it is irrelevant to this discussion). Then, $L(x)$ is a constant, and Eq. 13 implies that $r(x, x_i) \propto t(x, x_i)$. That is, before adaptation, the observed tuning curve has the shape of the cell's intrinsic tuning function, which is peaked at preferred orientation x_i and typically bell-shaped (Schiller et al. 1976; Webster and De Valois 1985). For convenience, we used the following periodic function (Teich and Qian 2003) for $t(x, x_i)$:

$$t(x, x_i) = c\{\cos[2(x - x_i)] + 1\}^k + b, \tag{14}$$

where b and c determine the baseline and peak firing rates in Eq. 14, respectively. The exponent k controls the tuning width (larger k produces narrower width). Examples of pre-

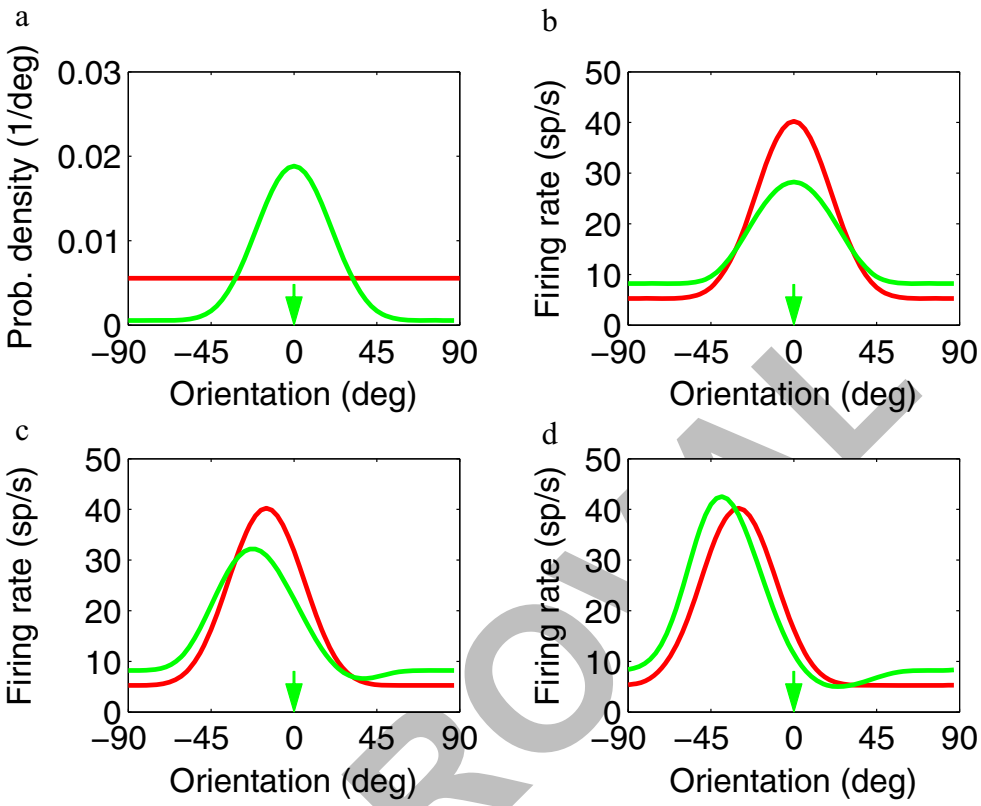

Fig. 4 The firing-rate-as-code-length hypothesis explains the adaptation-induced changes of orientation tuning curves. The adapted orientation is assumed to be 0° indicated by the green arrow in each panel. **a** The orientation probability distributions before (red) and after (green) the adaptation. **b**–**d** Comparison of tuning curves before (red) and after (green) the adaptation for cells whose preferred orientations are 0, 15, and 30° away from the adapted orientation, respectively

adaptation tuning curves (i.e., $r(x, x_i)$ as a function of x for fixed x_i) with $k = 4$ are shown in red in Figs. 4 and 5, panels b and c.

Now assume that there is adaptation at 0° orientation, and after adaptation, $P(x) = P_a(x)$. Although we do not yet know exactly how the brain updates $P(x)$ represented by interneurons (see "NML and Learning" section), $P_a(x)$ should have increased values at and around the adapted orientation and decreased values at other orientations, as we argued in connection with Eq. 12. We therefore used the following expression:

$$P_a(x) = P_0 + A\{z_+[\cos(2x) + 1]^m - z_-[\cos(2x) + 1]^n\}, \quad (15)$$

where the two cosine terms determine the increase and decrease of probabilities at different orientations, respectively. z_+ and z_- are not free parameters but normalize the two cosine terms so that $P_a(x)$ is normalized. m and n together control the orientation ranges of the probability increase and decrease, and A determines the strength of adaptation. When $n = 0$, Eq. 15 reduces to:

$$P_a(x) = P_0 + A\{z_+[\cos(2x) + 1]^m - P_0\}, \quad (16)$$

and an example with $m = 4$ and $A = 0.9$ is shown as the green curve in Fig. 4a. Relative to the constant baseline $P_0(x)$ (flat red line in Fig. 4a), this $P_a(x)$ has increased values at and around the adapted orientation and uniformly decreased values at other orientations. When $n > 0$, $P_a(x)$ has non-uniformly decreased values at the other orientations and an example with $n = 0.2$, $m = 4$, and $A = 0.9$ is shown as the green curve in Fig. 5a. This could occur if the updating of $P(x)$ during adaptation depends on the so-called Mexican-hat connectivity profile among cells tuned to different orientations (Teich and Qian 2006; Teich and Qian 2010). The broad peaks of $P_a(x)$ in Figs. 4a and 5a reflect the fact that the brain's estimation of an individual orientation is poor (Ding et al. 2017).

Plugging post-adaptation $P(x) = P_a(x)$ into Eqs. 5 and 13, we can then determine the tuning curves that reflect the adaptation-induced change of code length. Figures 4 and 5, panels c and d, compare the pre-adaptation (red) and post-adaptation (green) tuning curves for cells whose preferred orientations are 0°, 15°, and 30° away from the adapted orientation at 0° (green arrow). These simulation results explain all the adaptation-induced tuning changes listed above.

Dragoi et al. (2000) measured the adaptation-induced percent change in OSI and shift of tuning peak (Fig. 3, panels c and d). The corresponding simulations using the two different $P_a(x)$ in Figs. 4a and 5a are shown in Fig. 6. Results similar to the simulations in Figs. 4, 5, and 6 can be obtained with many other parameter sets.

We conclude that the observed tuning changes induced by adaptation may reflect the brain's adjustment of code lengths for different orientations after adaptation. Since at

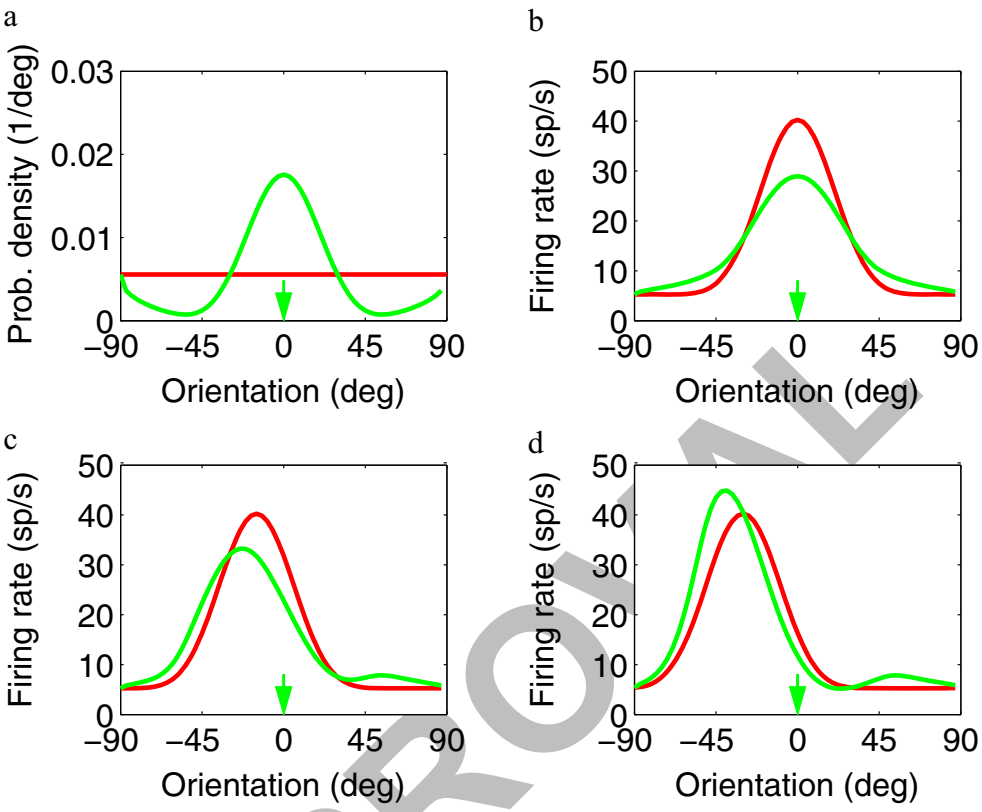

Fig. 5 The firing-rate-as-code-length hypothesis explains the adaptation-induced changes of orientation tuning curves. The same simulations as in Fig. 4 except that a Mexican-hat-shaped post-adaptation probability density function is used. The presentation format is identical to that of Fig. 4

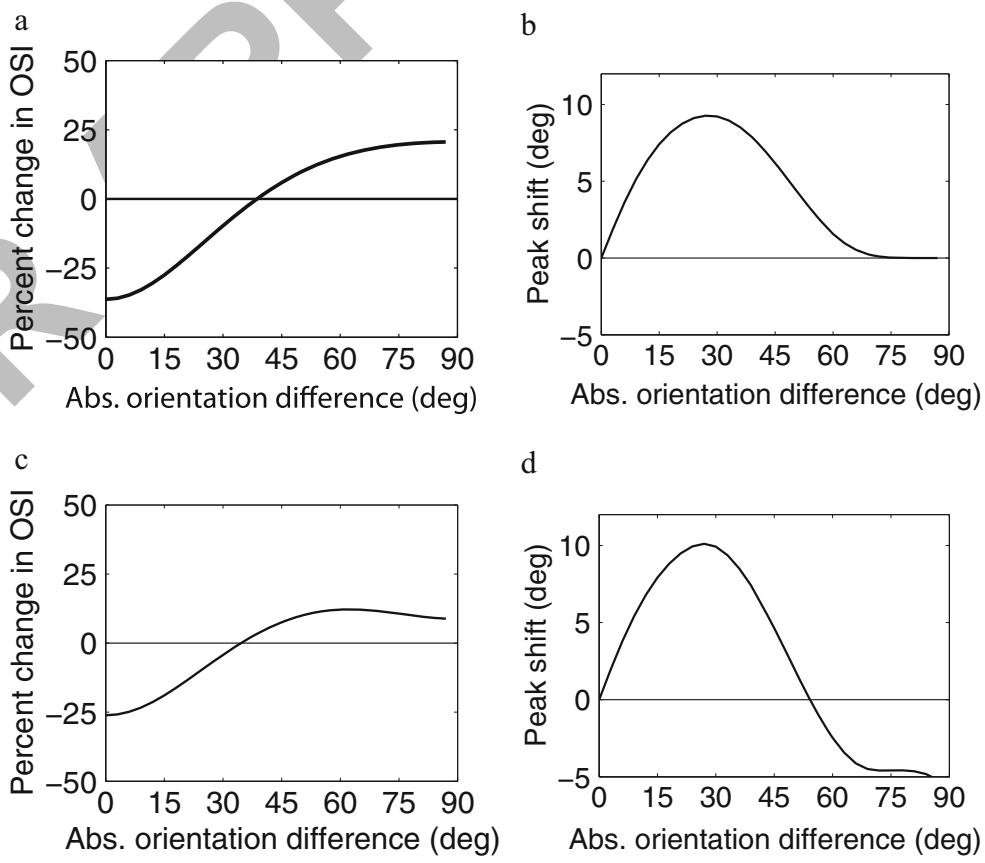

Fig. 6 The simulated percent change in OSI (**a**, **c**) and peak shift (**b**, **d**) according to the firing-rate-as-code-length hypothesis. Simulations using the post-adaptation probability density in Fig. 4a are shown in panels a and b, and those using the post-adaptation probability density in Fig. 5a are shown in panels c and d

the circuit level, the tuning changes can be explained by modifying recurrent connections among cells (Teich and Qian 2003; Teich and Qian 2010), the recurrent connection plasticity could be a physiological mechanism for online code-length minimization.

Our firing-rate-as-code-length hypothesis calls for a re-interpretation of neuronal tuning curves. Consider, for example, V1 cells tuned to vertical orientation. The traditional view is that when they fire, they signal the presence of vertical orientation on retina. According to the MDL framework, these cells' firing not only signals the presence of vertical orientation. In addition, their firing rates are modulated up or down according to whether vertical orientation is less or more probable than other orientations. This interpretation is also consistent with the observation that natural images usually evoke weaker neural responses than isolated patches of natural images or artificial stimuli (Gallant et al. 1998) because the former, with its large context, is more probable than the latter.

Top-down Attention, NML with Data Prior, and Feedback Connections

In addition to the adaptation and bottom-up attention discussed above, top-down attention can also be incorporated into the MDL framework. In the case of bottom-up attention, salient stimuli, because of their small probabilities reflected in the NML distributions, have longer code lengths and drive cells to higher firing rates. For top-down attention, on the other hand, the brain seeks a specific type of information based on its current functional needs. Such information-seeking could be realized, in the MDL framework, by a top-down modulation of the NML probabilities in lower levels. For example, area V1 may normally assign horizontal orientation a certain probability, and the corresponding firing rate, based on actual frequencies of orientations in the input. Now if horizontal orientation becomes subjectively more important (e.g., when searching for a horizontal key slot), then higher-level visual areas could use top-down, feedback connections to V1 to reduce the estimated probability of, and thus increase the firing rate to, horizontal orientation. In other words, since rare stimuli are bottom-up salient, the top-down process could instruct lower-level areas to treat a task-relevant stimulus as if it were rare, to boost its saliency. Thus, we must modify the MDL principle to take into account task relevance or subjectivity of information content, an aspect not encompassed by previous efficient/predictive coding theories.

Zhang (2011) introduced a positive data prior function, $s(x)$, to modify the NML distribution as:

$$P_{NML}(x) = \frac{s(x)P\left[x|\hat{\theta}(x)\right]}{\sum_y s(y)P\left[y|\hat{\theta}(y)\right]} \tag{17}$$

This is precisely what we need for modeling top-down attention. The data prior function $s(x)$ emphasizes certain inputs, at the expense of other inputs, according to the current, task-relevant need of the brain. Specifically, when a certain x is task-relevant, top-down attention will reduce its $s(x)$, increasing the code length (firing rate) for it. Alternatively, $s(x)$ can be viewed as modifying the models' likelihood functions in Eq. 17. In fact, there can be a dual relationship between data prior and model prior (Zhang 2011), which produce so-called informative versions of MDL (Grunwald 2007).

Thus, according to the MDL framework, a major role of top-down, feedback connections in the brain, is for higher levels to modify the lower-level model classes in order to increase transmission of behaviorally relevant information. The framework is consistent with the fact that top-down attention is slower than bottom-up attention because it takes time for high-level areas to send spikes down the feedback connections to modify NML distributions of lower-levels. This is fundamentally different from Rao and Ballard's proposal that feedback connections send higher-level predictions of inputs to the lower level for subtraction (Rao and Ballard 1999). The difference reflects different aims of the two approaches. Rao and Ballard's model, as are typical of most efficient/predictive coding models, aims to reconstruct retinal inputs. Therefore, a high-level sends its input prediction to the lower level, which subtracts this prediction and sends the error to the higher level for improvement. In contrast, our MDL framework focuses on regularity extraction to serve the brain's needs of sensory processing without input reconstruction. Although regularity extraction is the basis for both efficient coding and prediction, in the MDL framework, there is no input prediction coming from higher levels for lower-levels to subtract. Instead, NML unifies prediction, regularity extraction, and efficient coding at each level of processing.

Top-down processes may also direct motor outputs (including eye movements) to actively seek relevant information in the world.

Comparison with Existing Models

Our firing-rate-as-code-length hypothesis differs significantly from previous theories. We already mentioned some differences above. Here, we recapitulate the discussions and make some further comparisons. Although negative log probability is frequently used in the literature for computational convenience or for linkage to MDL concepts, to our knowledge, the firing-rate-as-code-length hypothesis for interpreting sensory neurons' responses has not been proposed.

Predictive Coding Models

Rao and Ballard (1999) used a two-part version of MDL (Rissanen 1978; Rissanen 1983) in their predictive coding

model, which, like other efficient/predictive coding models, aims to reconstruct the retinal image. Our NML-based MDL framework is very different in that it uses regularity extraction to serve the brain's functional needs rather than to reconstruct retinal images, and consequently, interprets neuronal responses and connections differently. In particular, Rao and Ballard's model and our framework interpret projection neurons' responses as representing errors of input reconstruction and coding useful features in the input, respectively. Additionally, while they assume that feedback connections carry the higher-level's prediction of the lower-level input, we assume that feedback connections modify the lower-level's model classes to transmit task-relevant information in the input.

Firing-Rate-As-Probability Theories

Firing-rate-as-probability theories, including a proposed implementation of Bayesian inference (Ma et al. 2006), posit that projection neurons transmit probability distributions of input features (or parameterizations of the distributions), whereas we suggest that the probability distributions computed in an area are not transmitted but are used to code input features efficiently and that probability distributions computed in different areas are relative to different model classes and concern different regularities of the inputs. As we noted in the "Evaluations of Major Theories of Neuronal Coding" section, firing-rate-as-probability theories are not consistent with adaptation and bottom-up-attention phenomena while our framework is. Note that we are not arguing against Bayesian inference, only the firing-rate-as-probability assumption used in many models including those that have been called Bayesian inference models. In fact, the Bayesian universal code with Jeffery's prior asymptotically achieves the minimax optimal regret of the NML code and may well be used by the brain because of its prequential property which is useful for prediction without a pre-specified time horizon (Grunwald 2007).

Saliency Models

Zhaoping (2002) proposed that V1 constructs a bottom-up saliency map such that, for a given visual scene, firing rate of V1 output neurons increases monotonically with the saliency values of the visual input in the classical receptive fields. There are no separate feature maps for creating such a bottom-up saliency map. Neuronal responses encode universal values of saliency that govern subsequent actions (e.g., saccades). In our framework, neuronal responses are also related to saliency. However, this is realized via neurons' firing rates being proportional to the code lengths for coding useful features. The code lengths are determined by the features' probabilities, which, in turn, are related to the saliency values.

Han and Vasconcelos (2010) presented another saliency model for object recognition in biological systems. Motivated by the observation that stimulus features with high bottom-up saliency have a low probability of occurrence, they proposed a top-down saliency measure using log likelihood ratio of Gabor-filter responses to target and non-target objects and demonstrated that this computation can be realized by a selective normalization procedure. In contrast, we assume that the top-down attention modifies lower-level NML distributions for coding-relevant stimulus features. More importantly, they eventually let cells' firing rates represent the posterior probability of target object via a nonlinear function of the log likelihood ratio so their model follows the traditional firing-rate-as-probability assumption. Instead, we assume that firing rates represent code length, not probability.

Normalization Models

On first glance, the NML distribution (Eq. 8) resembles the normalization models for sensory responses (Eq. 4), and the NML distribution with a data prior (Eq. 17) resembles the normalization models for attentional modulation (Reynolds and Heeger 2009). However, the normalization factors (denominators) in NML and in normalization models are very different. In NML, the denominator sums the maximum likelihood of a model class across all input data samples. In normalization models, the denominator is a constant plus the summed responses of all cells with a range of tuning (i.e., all cells in a model class) to the current input sample.

A key motivation for the normalization models is to fit V1 cells' contrast response curves. Indeed, the form of the normalization models mimics contrast saturation functions. The MDL framework offers an alternative, computational-level explanation of contrast responses, namely that high contrast occurs less frequently than low contrast in the real world; this reflects the fact that the world consists of coherent surfaces of objects and high contrast typically occurs at relatively rare object boundaries whereas low contrast typically occurs at relatively abundant object interiors. Indeed, Ruderman (1994) measured contrast distribution of natural images and his result can be approximated by:

$$P(c) = a - b\log(1 + c) \tag{18}$$

where c is the contrast and a and b are the positive constants; the probability decreases with contrast. If, as we postulated earlier, the brain learns this statistical regularity based on the MDL principle, then the corresponding NML distribution for encoding stimulus contrast should reflect the statistics. The contrast responses of projection neurons covering different ranges of contrast should then have the envelope -log *P(c)*, a curve resembling saturation. Thus, contrast response may not result from shunting inhibition of pooled responses to a given

stimulus; rather, it may reflect code-length optimization by a circuit that sample contrast statistics from many stimuli.

The MDL framework may also account for phenomena that the normalization models fail to explain. For example, we mentioned above that end-stopped V1 cells fire less when a contour extends beyond their classical receptive fields (Bolz and Gilbert 1986; Hubel and Wiesel 1968). More generally, V1 or MT cells' responses to their preferred orientation/direction within the classical receptive fields are suppressed when the surround has the same orientation/direction, but the suppression becomes weaker, or even turns into facilitation, when the surround orientation/direction differs greatly (Allman et al. 1985; Levitt and Lund 1997; Nelson and Frost 1978). The normalization models cannot explain these results because the normalization factor is untuned. Of course, one could modify the normalization models by making the normalization factor follow the observed results; however, this means that the normalization models have to be adjusted for each specific situation. The MDL framework may be able to explain these experimental findings because when the classical receptive field and its surround have similar (different) stimuli, the presence of the surround stimuli increases (decreases) the probability of the stimuli in the classical receptive field, and consequently, a shorter (longer) code length, in the form of a lower (higher) firing rate, is needed to transmit the information. Similarly, when a contour extends smoothly beyond an end-stopped cell's classical receptive field, the probability of the segment inside the receptive field is increased, leading to a shorter code length (reduced firing) of the cell.

Lossy MDL and Prefix-Free Neural Code

The standard MDL uses the terms "code" and "probability distribution" interchangeably because once a probability distribution is specified, one can always design a lossless, prefix-free code (a.k.a., prefix code) that saturates the Kraft–McMillan inequality such that the code length is equal to negative log probability (Grunwald 2007). In contrast, phenomena such as change blindness (Pashler 1988) suggest that the brain uses a lossy code to transmit behaviorally relevant information and discard irrelevant details of the input. We will therefore speculate on a lossy MDL code as a candidate for neural code. To motivate our proposal, consider the example of seeing something moving in a jungle. The most survival-relevant information may be whether the moving thing is a predator or a pray. If it is a predator, the next most relevant information may be whether it is the type that one could fight against (e.g., a wolf) or better flee from (e.g., a tiger). To optimize survival, the brain should use its visual neurons' first few spikes to transmit the most relevant information, and the next few spikes to transmit the second most relevant information, and so on. Only crude aspects of low-level features that are sufficient for building relevant, high-level categorical decisions should be transmitted quickly. It would be a huge mistake to waste the precious first several spikes on transmitting, for example, the precise orientation of a stripe on the animal's fur. On the other hand, the brain is certainly able to judge the orientation when one is asked to do so in a safe setting.

These considerations suggest that a partially transmitted code should be meaningful so that a brain area can start processing inputs immediately after receiving spikes from lower areas, that the code should be as short (efficient) as possible and carry pieces of information ordered according to their behavioral relevance/urgency, and that higher-level areas should instruct lower-level areas on what and how much details to transmit depending on the situation. Therefore, the brain might use entropic, prefix-free codes (based on NML distributions) with earlier spikes carrying more behaviorally important information.

Consider the toy example in Table 1 of coding four symbols (column 1) with known probabilities (column 2). Code 1 is fixed length and inefficient (the length of 2 bits/symbol is greater than the entropy of 1.75 bits/symbol). Code 2 is the Huffman code, which is entropic (average length 1.75 bits/symbol) and prefix free (no code word is a prefix of another code word). Code 3 reverses the bit order of each code word of code 2. It is entropic but not prefix free. Although code 3, like the other two codes, is uniquely decodable (after receiving a whole message, the bit string can be reversed and decoded according to code 2), a partial message is meaningless. In contrast, a Huffman-coded string can be decoded online as each bit is received without the need to wait for a whole message or a whole code word. For example, the first bit divides choices into A vs. (B, C, D). Because each bit of a code word divides the remaining choices into two sets with equal probabilities, the bits are ordered from the most to least informative. (Although the Huffman code is a symbol code, similar arguments could be made with the entropic, arithmetic coding for blocks of arbitrary lengths.)

We propose that the brain might use a Huffman-like code (or arithmetic-like coding) based on NML distributions. Such a code is attractive because of the efficiency, the bit ordering from the most to least informative, and the prefix-free property

Table 1 Three codes for the four symbols with the given probabilities. Codes 1 and 2 are prefix free. Codes 2 and 3 are entropic. Code 2 (Huffman) is both prefix free and entropic

Symbol	Probability	Code 1	Code 2	Code 3
A	1/2	00	0	0
B	1/4	01	10	01
C	1/8	10	110	011
D	1/8	11	111	111

allowing immediate decoding as each bit comes in. We suggest that neural codes should be similar in that the first spikes of a neuronal population carry the most task/situation-relevant information so that the brain can take most pressing actions at the earliest possible time. The later spikes carry less relevant details that may be truncated by top-down instructions or by a change of inputs (e.g., a saccade to a different part of the world or a changing world), resulting in a lossy code. Experiments that present stimuli for only several to tens of milliseconds provide indirect evidence for the prefix-free and lossy nature of neural codes: subjects could identify global or high-level, categorical features better than local or low-level details (Chen 1982; Navon 1977; Thorpe et al. 1996), suggesting that truncated visual transmission is meaningful and that the transmission leading to high-level categorization, which is more behaviorally relevant than low-level details, is prioritized.

In information theory, the rate-distortion curve is a standard tool for studying lossy transmission (Blahut 1972). Each point of the curve specifies the minimum input information that has to be transmitted to the output (i.e., the rate) in order to keep the average distortion under a given value. Equivalently, each point specifies the minimum average distortion for a given rate. The distortion for each input/output pair is pre-defined. (The rate is similar to channel capacity except that the former is the mutual information minimized against the channel transition probabilities whereas the latter is the mutual information maximized against input distribution. We will not distinguish the two terms in the following for simplicity.) The rate-distortion curve has been used as a computational-level theory for understanding discrimination vs. generalization in perception (Sims 2018). The main idea is that when a system transmits inputs whose information (i.e., entropy) exceeds the system's channel capacity, the output will have distortion which determines discrimination between, or generalization across, different inputs. The information bottleneck theory (Tishby et al. 2000) is a version of the rate-distortion theory in which the distortion for each input/output pair is not pre-defined, but determined according to how much information the output carries about the input's assigned label (e.g., the label "cat" for an input image). The truncated, lossy code discussed above could be viewed as a possible neural implementation of the rate-distortion function. Specifically, because of the limited rate or channel capacity, projection neurons cannot transmit all input information as stimuli stream in, and truncated transmission leads to distortion. If the spikes of a neural code are arranged from the most to least relevance to current behavior, then the distortion with respect to the behavior "label" is minimized for a given rate of transmission.

The firing-rate-as-code-length hypothesis implies that the channel capacity (firing rates) of projection neurons is dynamic, being greater for lower-probability stimuli which require longer codes. This ensures that unexpected, salient stimuli are not truncated more than common stimuli.

Encoding Vs. Decoding

Coding can be divided into encoding and decoding. The engineering notion of encoding and decoding is well defined: When a signal needs to be transmitted over a noisy communication channel of limited capacity (e.g., a phone line), one should encode the signal to compress it (while allowing error correction), transmit it, and then decode it to recover the original signal on the other end. It is widely assumed that the brain does similar encoding and decoding. Our MDL framework suggests that the brain encodes input stimuli into neuronal responses but does not decode the responses to recover the original inputs. The main reason is that, unlike a phone line that has to reproduce the input voice at the other end, the brain never needs to reconstruct the raw sensory inputs it receives. Rather, as we already emphasized, the brain attempts to understand the sensory inputs by processing them. For example, the brain processes the retinal images to reveal objects and their relationships but hardly needs to reconstruct the retinal images because retina is part of the brain and no homunculus exists in the cortex to look at the reconstructed images. More generally, it is hard to imagine that one brain area needs to accurately reconstruct neural firing patterns (spike trains) of another area; rather, a brain area should extract additional regularity from, and thus achieve further understanding of, the input. If the firing patterns of a sensory area are needed, for instance, to guide a certain motor response, then the motor area of the brain should use the firing patterns directly, instead of encoding, transmitting, and decoding. For example, in the unlikely scenario that raw retinal image was needed, the brain would have evolved to use the retinal image instead of decoding a poorer version of it from, say, LGN or V1 responses.

One may reasonably identify the brain's logic of relating neuronal responses to subjective perception as decoding. Note, however, that this decoding is fundamentally different from the engineering notion of decoding. Specifically, neuronal responses along hierarchical stages of sensory pathways extract and encode progressively more complex statistical regularities in the input stimuli. A small subset of these responses presumably gives rise to our subjective perception of useful features in inputs without any need of reconstructing the raw inputs. We therefore suggest that neural decoding should be viewed as the link from neuronal responses to perceptual estimation of useful stimulus features, but not as input reconstruction. Also, note that encoding and decoding are often related; for example, the population-average method of Eq. 1 is a *decoding* model but it implies that firing rates *encode* the probabilities of preferred stimuli.

A related question is whether sensory decoding follows the same low-to-high-level hierarchy of sensory encoding. Many studies assume, often implicitly, that the answer is affirmative. However, a recent study shows that this assumption fails to explain a simple psychophysical experiment and suggests that

visual decoding progresses from high-to-low-level features in working memory, with higher-level features constraining the decoding of lower-level features (Ding et al. 2017). Since higher-level features have greater functional significance than lower-level features, this decoding scheme is consistent with the above notion that the brain should prioritize transmission of behaviorally relevant information.

NML and Learning

Given the importance of the NML distribution (or a related OUC as its approximation) in the MDL framework, a relevant question is how can a brain area produce such a distribution particularly when the input data space is high dimensional? Variational methods in machine learning provide a potential answer as they have demonstrated that neural networks can learn complex probability distributions via gradient decent (Dayan et al. 1995; Kingma and Welling 2013) or even a local plasticity rule (Hinton et al. 1995). To outline the approach for the NML distribution (Eq. 8), we define the "energy" of a data sample x relative to a model class parameterized by θ as:

$$E(x) = -\log P\left[x|\hat{\theta}(x)\right] \tag{19}$$

(i.e., the code length according to the model in the class that maximizes the likelihood of the sample) and rewrite Eq. 8 in the form of a Boltzmann distribution (with $\beta = 1$):

$$P_{NML}(x) = \frac{\exp[-E(x)]}{\sum_y \exp[-E(y)]} \tag{20}$$

The numerator is known as the partition function $Z \equiv \sum_y \exp[-E(y)]$, and the regret and complexity measure in Eqs. 9 and 10 become $\log Z$. Use the standard definition of Helmholtz free energy as the mean energy minus entropy:

$$A = \sum_x P(x)E(x) + \sum_x P(x)\log P(x) \tag{21}$$

for any probability distribution $P(x)$. Then (Dayan et al. 1995),

$$A = -\log Z + KL\left[P(x)\|P_{NML}(x)\right] \tag{22}$$

where KL is the Kullback–Leibler divergence. Since KL is non-negative and minimized to 0 when the two distributions are equal, A reaches the minimum value of $-\log Z$ when $P(x) = P_{NML}(x)$. (The physical analogy is that Helmholtz free energy A approaches the minimum $-\log Z$ when any non-equilibrium distribution $P(x)$ approaches the equilibrium, Boltzmann distribution $P_{NML}(x)$.) Therefore, if $P(x; \phi)$ is a family of probability distributions parameterized by weights ϕ of a neural network, then the network could be trained to approximate $P_{NML}(x)$ by minimizing the cost A in Eq. 21 against ϕ as input data are sampled, and $-A$ is a lower bound for $\log Z$ (Dayan et al. 1995; Friston 2010; Hinton et al. 1995; Kingma and Welling 2013). If data statistics are changed, the neural network's approximation of $P_{NML}(x)$ would change accordingly (as we assumed in the example of orientation adaptation in "Simulating Adaptation-Induced Changes of V1 Orientation Tuning Curves" section).

Moreover, $E(x)$ and $\log Z$ depend on the model-class parameterization θ, which can also be implemented as weights of a neural network. One could thus, for example, adjust the *equilibrium* Helmholtz free energy ($A = -\log Z$) by modifying θ in order to control the NML regret or complexity ($\log Z$). Since the model-class complexity and the code length for an input (i.e., neuronal firing rate) may be related to coding sparsity, this could be a mechanism for adjusting the degree of sparsity. Finally, the learning of the ϕ and θ parameters could be interleaved.

Discussion

Understanding the nature of neural code is of fundamental importance. Although extant theories have been successful in revealing many properties of neural coding, they are not always consistent with major empirical observations or with each other. Our efforts in this project focus on proposing a novel, modern MDL-based framework for characterizing neural code. The framework aims to integrate the strengths of extant theories, explain (or at least be consistent with) more empirical observations, and unify sensory processing and attention. The framework leads to the specific proposal that neural firing rates are proportional to code lengths given by negative log NML probability distributions (or closely related OUCs) for stimulus features. We showed via simulations that this firing-rate-as-code-length hypothesis can explain all the observed changes of V1 orientation tuning curves induced by orientation adaptation.

Our framework contains five essential elements, the combination of which, to our knowledge, has never been suggested before.

1. The firing rates of sensory projection neurons are proportional to code length, not the probability or its parameterization, of stimulus features. Indeed, for efficient transmission of inputs, a system should use a proper probability distribution to encode/compress the inputs instead of transmitting the probability distribution itself.
2. The code length is based on an OUC (such as NML distribution) of a given model class which maximizes regularity extraction, predictive ability, and data compression to achieve input understanding by balancing data fitting and model-class complexity. Parameters specifying a

model class and its NML distribution might be learned or tuned together.

3. The actual code in the temporal firing pattern of a neuronal population is Huffman-like such that it has minimal firing rates, is prefix free, and the order of information transmission is from the most relevant to the least relevant according to the current task or goal. In this way, a partially transmitted message is meaningful and can be processed immediately by the next stage, the system could respond to the most relevant aspect of input with the shortest delay, and a truncated, lossy transmission would minimize behaviorally relevant distortion.
4. The brain does not really face a decoding problem in the form of input reconstruction because the input representation is already in the brain. Rather, the brain extracts useful stimulus features during efficient encoding, without the need to reconstruct the original input signal. The brain processes input hierarchically to extract progressively more complex and global regularities to serve various perceptual and motor functions.
5. Top-down signals are sent to modulate lower-level model classes, direct eyes to relevant regions, and set prior expectations of data statistics, to allow selective processing of informative and relevant inputs according to the current task demand.

Needless to say, any theory is only a crude approximation of reality but we hope our MDL framework will provide a fresh perspective for characterizing neural code. Future empirical data may be able to evaluate our specific, firing-rate-as-code-length hypothesis and our speculations on the nature of neural codes in sensory firing patterns.

Acknowledgments We thank Dr. Jay Myung for encouraging us to explore this topic and for his many helpful comments and suggestions.

Funding Information This work was supported by AFOSR FA9550-15-1-0439 and NSF 1754211.

References

Adrian, E. D. (1926). The impulses produced by sensory nerve endings. *The Journal of Physiology, 61*, 49–72.

Albrecht, D. G., & Geisler, W. S. (1991). Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Visual Neuroscience, 7*, 531–546.

Allman, J., Miezin, F., & McGuinness, E. (1985). Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons. *Annual Review of Neuroscience, 8*, 407–430.

Anderson, J. S., Carandini, M., & Ferster, D. (2000). Orientation tuning of input conductance, excitation, and inhibition in cat primary visual cortex. *Journal of Neurophysiology, 84*, 909–926.

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation, 2*, 308–320.

Balasubramanian, V., Berry, M. J., & Kimber, D. (2001). Metabolically efficient information processing. *Neural Computation, 13*, 799–816.

Barlow, H. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception, 1*(4), 371–394.

Barlow, H., & Foldiak, P. (1989). Adaptation and decorrelation in the cortex. In R. Durbin, C. Miall, & G. Mitchinson (Eds.), *The computing neuron* (pp. 54–72). New York: Addison-Wesley.

Barron, A., Rissanen, J., & Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory, 44*, 2743–2760.

Bell, A. J., & Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Research, 37*, 3327–3338.

Blahut, R. E. (1972). Computation of channel capacity and rate-distortion functions. *IEEE Transactions on Information Theory, 18*, 460–473.

Blakemore, C., & Campbell, F. W. (1969). Adaptation to spatial stimuli. *The Journal of Physiology, 200*, 11P–13P.

Bolz, J., & Gilbert, C. D. (1986). Generation of end-inhibition in the visual cortex via interlaminar connections. *Nature, 320*, 362–365.

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews. Neuroscience, 13*, 51–62.

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience, 17*, 8621–8644.

Chen, L. (1982). Topological structure in visual perception. *Science, 218*, 699–700.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The Helmholtz machine. *Neural Computation, 7*, 889–904.

Ding, S., Cueva, C. J., Tsodyks, M., & Qian, N. (2017). Visual perception as retrospective Bayesian decoding from high- to low-level features. *Proceedings of the National Academy of Sciences, 114*, *E9115–E9124*.

Dragoi, V., Sharma, J., & Sur, M. (2000). Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron, 28*, 287–298.

Dragoi, V., Rivadulla, C., & Sur, M. (2001). Foci of orientation plasticity in visual cortex. *Nature, 411*, 80–86.

Fang, F., Murray, S. O., Kersten, D., & He, S. (2005). Orientation-tuned fMRI adaptation in human visual cortex. *Journal of Neurophysiology, 94*, 4188–4195.

Felsen, G., Shen, Y. S., Yao, H., Spor, G., Li, C., & Dan, Y. (2002). Dynamic modification of cortical orientation tuning mediated by recurrent connections. *Neuron, 36*, 945–954.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical-cells. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision, 4*, 2379–2394.

Flash, T., & Hogan, N. (1985). The coordination of arm movements - an experimentally confirmed mathematical-model. *The Journal of Neuroscience, 5*, 1688–1703.

Friston (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138.

Gallant, J. L., Connor, C. E., & Van Essen, D. C. (1998). Neural activity in areas V1, V2 and V4 during free viewing of natural scenes compared to controlled viewing. *Neuroreport, 9*, 2153–2158.

Ganmor, E., Segev, R., & Schneidman, E. (2011). The architecture of functional interaction networks in the retina. *The Journal of Neuroscience, 31*, 3044–3054.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*, 711–724.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*, 1416–1419.

Gibson, J. J., & Radner, M. (1937). Adaptation, after-effect and contrast in the perception of tilted lines. I. Quantitative studies. *Journal of Experimental Psychology, 20*, 453–467.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature, 391*, 481–484.

Grunwald, P. D. (2007). *The minimum description length principle*. Cambridge: MIT Press.

Grunwald, P. D., Myung, I. J., & Pitt, M. A. (2005). *Advances in minimum description length: theory and applications*. Cambridge: MIT Press.

Han, S., & Vasconcelos, N. (2010). Biologically plausible saliency mechanisms improve feedforward object recognition. *Vision Research, 50*, 2295–2307.

Harpur, G. F., & Prager, R. W. (1996). Development of low entropy coding in a recurrent network*. *Network: Computation in Neural Systems, 7*, 277–284.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience, 9*, 181–197.

Hinton, G., Dayan, P., Frey, B., & Neal, R. (1995). The "wake-sleep" algorithm for unsupervised neural networks. *Science, 268*, 1158–1161.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology, 195*, 215–243.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews. Neuroscience, 2*, 194–203.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology, 10*, e1003915.

Kingma DP, Welling M. (2013). Auto-encodrwing variational bayes. *arXiv preprint arXiv:*1312.6114.

Levitt, J. B., & Lund, J. S. (1997). Contrast dependence of contextual effects in primate visual cortex. *Nature, 387*, 73–76.

Li, W., & Gilbert, C. D. (2002). Global contour saliency and local colinear interactions. *Journal of Neurophysiology, 88*, 2846–2856.

Li, C.-Y., & Li, W. (1994). Extenstive integration field beyond the classical receptive field of cat's striate cortical neurons - classification and tuning properties. *Vision Research, 34*, 2337–2355.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience, 9*, 1432–1438.

Meng, X., & Qian, N. (2005). The oblique effect depends on perceived, rather than physical, orientation and direction. *Vision Research, 45*, 3402–3413.

Motoyoshi, I., Sy, N., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*, 206–209.

Myung, J. I., Navarro, D. J., & Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology, 50*, 167–179.

Navon, D. (1977). Forest before trees: the precedence of global features in visual perception. *Cognitive Psychology, 9*, 353–383.

Nelson, J. I., & Frost, B. J. (1978). Orientation-selective inhibition from beyond the classic visual receptive field. *Brain Research, 139*, 359–365.

Nowak, L. G., Munk, M. H., Nelson, J. I., James, A. C., & Bullier, J. (1995). Structural basis of cortical synchronization. I. Three types of interhemispheric coupling. *Journal of Neurophysiology, 74*, 2379–2400.

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

Paradiso, M. A. (1988). A theory for the use of visual orientation information which exploits the columnar structure of striate cortex. *Biological Cybernetics, 58*, 35–49.

Pashler, H. (1988). Familiarity and visual change detection. *Attention, Perception, & Psychophysics, 44*, 369–378.

Poggio, T., Torre, V., & Koch, C. (1985). Computational vision and regularization theory. *Nature, 317*, 314–319.

Qian, N. (1994). Computing stereo disparity and motion with known binocular cell properties. *Neural Computation, 6*, 390–404.

Qian, N. (1997). Binocular disparity and the perception of depth. *Neuron, 18*, 359–368.

Rao, R. P. N., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation, 9*, 721–763.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87.

Reid, R. C., & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature, 378*, 281–284.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron, 61*, 168–185.

Rissanen, J. (1978). Modeling by the shortest data description. *Automata, 14*, 465–471.

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics, 11*, 416–431.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42*, 40–47.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Information Theory, 47*, 1712–1717.

Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems, 5*, 517–548.

Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *Journal of Neurophysiology, 76*, 2790–2793.

Schiller, P. H., Finlay, B. L., & Volman, S. F. (1976). Quantitative studies of single-cell properties in monkey striate cortex. II. Orientation specificity and ocular dominance. *Journal of Neurophysiology, 39*(6), 1320–1333.

Schneidman, E., Berry, M. J., Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature, 440*, 1007–1012.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379–423.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., & Chichilnisky, E. J. (2006). The structure of multi-neuron firing patterns in primate retina. *The Journal of Neuroscience, 26*, 8254–8266.

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: natural scenes and gestalt rules. *Proceedings of the National Academy of Sciences of the United States of America, 98*, 1935–1940.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38*, 743–761.

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience, 24*, 1193–1216.

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science, 360*, 652–656.

Stocker, A. A., & Simoncelli, E. P. (2006). Sensory adaptation within a Bayesian framework for perception. In Y. Weiss, B. Scholkopf, & J. Platt (Eds.), *Advances in neural information processing systems*. Cambridge: MIT Press.

Tanaka, H., Krakauer, J. W., & Qian, N. (2006). An optimization principle for determining movement duration. *Journal of Neurophysiology, 95*, 3875–3886.

Teich, A. F., & Qian, N. (2003). Learning and adaptation in a recurrent model of V1 orientation selectivity. *Journal of Neurophysiology, 89*, 2086–2100.

Teich, A. F., & Qian, N. (2006). Comparison among some models of orientation selectivity. *Journal of Neurophysiology, 96*, 404–419.

Teich, A. F., & Qian, N. (2010). V1 orientation plasticity is explained by broadly tuned feedforward inputs and intracortical sharpening. *Visual Neuroscience, 27*, 57–73.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Tishby N, Pereira FC, Bialek W. (2000). The information bottleneck method. *arXiv:physics/0004057.*

van Kan, P. L. E., Scobey, R. P., & Gabor, A. J. (1985). Response covariance in cat visual cortex. *Experimental Brain Research, 60*, 559–563.

Webster, M. A., & De Valois, R. L. (1985). Relationship between spatial-frequency and orientation tuning of striate-cortex cells. *Journal of the Optical Society of America A. Optics and Image Science, 2*, 1124–1132.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*, 598–604.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, 111*, 8619–8624.

Yang, Z., & Purves, D. (2003). A statistical explanation of visual space. *Nature Neuroscience, 6*, 632–640.

Yenduri PK, Zhang J, Gilbert A. (2012) *Proceedings of the Third International Conference on Intelligent Control and Information Processing, ICICIP 20122012*.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends in Cognitive Sciences, 10*, 301–308.

Zhang, J. (2011). Model selection with informative normalized maximal likelihood: data prior and model prior. In E. N. Dzhafarov & L. Perry (Eds.), *Descriptive and normative approaches to human behavior* (pp. 303–319). Hackensack: World Scientific.

Zhaoping, L. (2002). A saliency map in primary visual cortex. *Trends in Cognitive Sciences, 6*, 9–16.

Zhaoping, L. (2014). *Understanding vision: theory, models, and data*. London: Oxford University Press.