nature neuroscience

A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts

Xue-Xin Wei[1] & Alan A Stocker[1,2]

Bayesian observer models provide a principled account of the fact that our perception of the world rarely matches physical reality. The standard explanation is that our percepts are biased toward our prior beliefs. However, reported psychophysical data suggest that this view may be simplistic. We propose a new model formulation based on efficient coding that is fully specified for any given natural stimulus distribution. The model makes two new and seemingly anti-Bayesian predictions. First, it predicts that perception is often biased away from an observer's prior beliefs. Second, it predicts that stimulus uncertainty differentially affects perceptual bias depending on whether the uncertainty is induced by internal or external noise. We found that both model predictions match reported perceptual biases in perceived visual orientation and spatial frequency, and were able to explain data that have not been explained before. The model is general and should prove applicable to other perceptual variables and tasks.

Perception involves two important stages of processing: the representation of incoming sensory information, followed by the interpretation of that representation. Two prominent hypotheses have separately guided our understanding of these two processing stages, but each has limitations when considered alone. The efficient coding hypothesis argues that neural resource limitations lead to efficient sensory representations that are optimized with regard to the specific stimulus statistics of the natural environment[1,2]. This hypothesis can explain several key features of neural coding in early sensory areas (for example, see refs. 3–5), but it does not specify how these coding characteristics can give rise to important aspects of perceptual behavior such as perceptual biases. In contrast, the Bayesian hypothesis posits that perception is an act of unconscious inference that interprets the noisy sensory representation in the context of prior knowledge about the world[6–8]. This hypothesis provides a normative explanation for many aspects of perceptual and sensorimotor behavior (for example, see refs. 9–12), but it has been criticized for using somewhat arbitrary model specifications to explain psychophysical data[13,14]. We unified ideas of efficient coding and Bayesian inference into a new model of perceptual behavior. Specifically, we propose a Bayesian observer model that is constrained by assuming an efficient representation of the sensory input.

Two key components define a Bayesian observer: the prior distribution that reflects the observer's belief about how frequently a certain stimulus value occurs, and the likelihood function that captures the encoding accuracy in the sensory representation of the observer. Previous studies have proposed independent constraints on either the prior belief based on natural[15,16] or learned[9,12] stimulus statistics, or the likelihood function based on natural stimulus uncertainties[17,18] or neural physiological tuning characteristics[10], but not both. In contrast, our new model formulation jointly constrains the prior belief and the likelihood function by assuming that the sensory representation and the interpretation of the sensory evidence are optimized with regard to the stimulus statistics of its sensory environment. In particular, we formulated the efficient coding problem at the level of Fisher information, which, together with an assumption about the noise structure, allowed us to specify the likelihood function. Thus, we were able to precisely formulate a Bayesian observer model for any stimulus variable with known natural statistics.

We validated our framework by formulating observer models for two perceptual variables for which the natural statistics are known, visual orientation and spatial frequency. The models make a number of distinct and rather surprising predictions; for example, that percepts are frequently biased away from the peaks of the prior, a prediction that is seemingly anti-Bayesian[19]. We found that the predictions were well matched by data from several studies reporting measured biases in perceived visual orientation and spatial frequency under different levels and sources of uncertainty. Our results demonstrate that, by integrating the ideas of efficient coding with Bayesian decoding, it is possible to formulate well-constrained observer models that can account for perceptual behavior that has not been explained before. A preliminary version of this work has been presented previously[20].

RESULTS

We modeled perception as a probabilistic encoding-decoding process[10] (**Fig. 1a**). The presentation of a stimulus value θ elicits a noisy sensory measurement m (encoding), based on which the observer then generates an estimate $\hat{\theta}(m)$ that represents the perceived stimulus value (decoding). We combined two general assumptions to define our observer model. First, we assumed that encoding is efficient, that is, the sensory representation is optimally adapted to the natural stimulus

[1]Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [2]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, Pennsylvania, USA. Correspondence should be addressed to A.A.S. (astocker@sas.upenn.edu).

Received 3 June; accepted 11 August; published online 7 September 2015; doi:10.1038/nn.4105

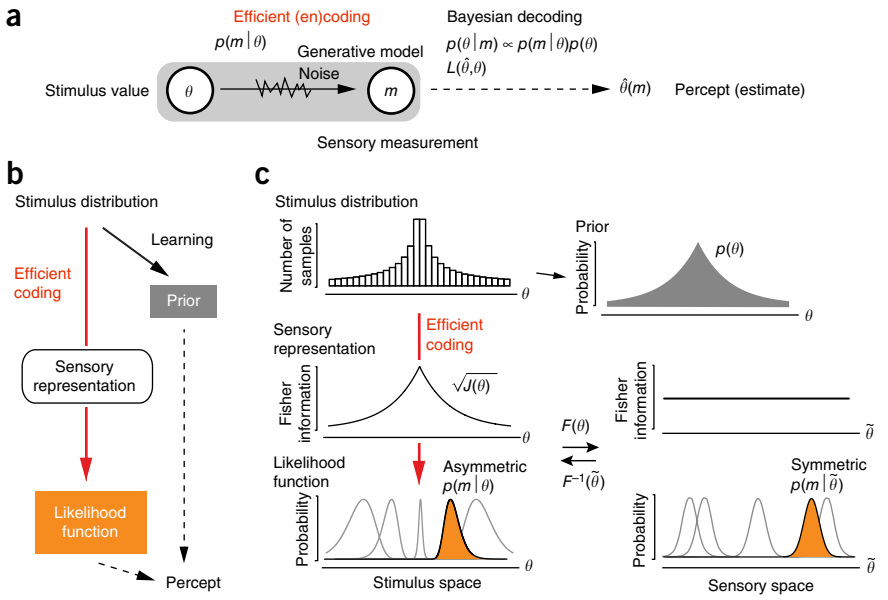

Figure 1 Bayesian observer model constrained by efficient coding. (**a**) We model perception as an encoding-decoding process. We assume encoding is governed by efficient coding and is characterized by the corresponding conditional probability distribution $p(m|\theta)$ of the sensory measurement m given a stimulus value θ. We also assume that decoding is Bayesian based on an accurate generative model of the sensory process. The percept $\hat{\theta}(m)$ is then specified based on the posterior distribution $p(\theta|m)$ and the loss function $L(\hat{\theta},\theta)$. (**b**) Our assumptions imply that the Bayesian observer is constrained by the natural stimulus distribution: the prior belief is assumed to directly match the stimulus distribution (for example, through learning), although the likelihood function is constrained by the stimulus distribution via efficient coding. (**c**) Example for an arbitrary stimulus distribution. An efficient coding principle that maximizes mutual information implies that the encoding accuracy (measured as the square-root of the Fisher information $J(\theta)$) matches the stimulus distribution. With some assumptions about the sensory noise characteristics, the likelihood function is fully constrained by the Fisher information. Likelihood functions for different sensory measurements are shown to illustrate their heterogeneity across the stimulus space. Technically, the likelihood functions can be computed by assuming a symmetric noise structure (that is, symmetric likelihood functions) in a space in which the Fisher information is uniform (sensory space, characterized by the mapping $F(\theta)$) and then transforming those symmetric likelihood functions back to the stimulus space.

distribution in the sense of maximizing mutual information between the stimulus and said representation. Second, we assumed that decoding is Bayesian and is based on an accurate (generative) model of the sensory process, that is, the observer's prior belief matches the true stimulus distribution and the likelihood function faithfully reflects the encoding characteristics. As a result, both the observer's prior belief and likelihood function are jointly constrained by the stimulus distribution (**Fig. 1b**). With the additional assumption about the observer's loss function (an important component of the Bayesian decoder that states how costly perceptual errors are for the observer), we can make quantitative predictions for the percept of a stimulus variable for which the natural stimulus distribution is known.

Efficient coding and the likelihood function

We adopted a definition of efficient coding that assumes that sensory encoding maximizes the mutual information $I[\theta, m]$ between the sensory measurement m and the stimulus variable θ with regard to the intrinsic uncertainty (internal noise) in the sensory representation[21]. This definition allowed us to establish a link between the probability distribution of the stimulus $p(\theta)$ and Fisher information $J(\theta)$ of the sensory representation using a bound on mutual information[22]. Assuming the bound is tight and mutual information is limited, we found that

$$p(\theta) \propto \sqrt{J(\theta)} \tag{1}$$

Fisher information is a measure of encoding accuracy and reflects the amount of coding resources that is dedicated to the representation of a certain stimulus value θ. Equation (1) provides an intuitive way of characterizing an efficient sensory representation: coding resources should be allocated according to the stimulus distribution, resulting in a more accurate representation of those stimulus values that occur more frequently. Note that in deriving equation (1) we relied on a formal constraint to limit the overall mutual information that may not be intuitive to interpret nor easy to validate in terms of neural data (equation (4), Online Methods). The derived relation between prior distribution and Fisher information (equation (1)), however, is supported by the results of previous studies that have maximized mutual information directly with regard to neural tuning parameters[22–24].

Although Fisher information constrains the likelihood function, it is not sufficient to fully specify its shape. An additional assumption about the noise structure is required. Let us consider a function $F(\theta)$ that maps the stimulus space to a new space in which Fisher information is uniform (**Fig. 1c**). We refer to this space as the 'sensory space' in reference to Gustav Fechner because discriminability, when measured in units of this space, is uniform[25]. The mapping $F(\theta)$ is defined as the cumulative of the stimulus distribution (equation (8), Online Methods). Uniform Fisher information implies that the noise and thus the likelihood function is homogeneous in this space. We make the additional assumption that the noise is such that the expected likelihood function (that is, averaged out over many trials) is symmetric around the stimulus value in the sensory space. Additive and symmetric noise is the simplest condition for which this assumption is true (for example, Gaussian as illustrated in **Fig. 1c**). Although the assumption seems parsimonious given the homogeneity of the space, the degree to which it is a valid assumption for real neural populations is unclear. In simulations of reasonably realistic neural population models, we found the assumption to approximately hold under a fairly large range of conditions (see Discussion).

With the likelihood defined in the sensory space, the likelihood function in the stimulus space can then be obtained by simply applying the inverse mapping $F^{-1}(\tilde{\theta})$. As a result, the likelihood functions when formulated in stimulus space are typically asymmetric, with a long tail away from the peak of the prior distribution.

Note that by formulating the efficient coding in terms of Fisher information we are able to specify the likelihood function without having to assume specific details about the tuning characteristics of the underlying neural representation. We deliberately chose such a formulation because it provided a more parsimonious, yet also more general, description of the Bayesian observer model. As we

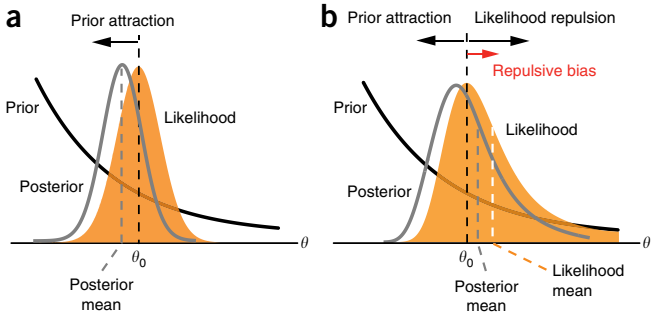

Figure 2 Prediction 1: Bayesian perception can be biased away from the prior peak. (**a**) A standard Bayesian observer model that a priori assumes a symmetric likelihood function typically predicts perceptual biases toward the peak of the prior. This bias towards the prior has been considered to be a fundamental characteristic of a Bayesian model. (**b**) In our new Bayesian observer model, efficient encoding promotes a nonlinear mapping between stimulus and sensory representation. Assuming that the sensory representation is affected by internal noise, the resulting likelihood function is asymmetric for any non-uniform prior distribution, with a long tail pointing away from the prior peak. As a result, the estimate can be biased away from the prior peak. Here, this is illustrated assuming a squared-error loss function. As a result of its asymmetry, the mean of the likelihood function is away from the peak of the prior relative to the true stimulus value θ_0 (likelihood repulsion). Although the prior still leads to an attractive shift of the posterior (prior attraction), the net bias can be repulsive. Note that the degree of asymmetry of the likelihood function, and thus the magnitude of the repulsive bias, depends directly on the steepness of the prior. Both examples are illustrated for the case of the median likelihood function (that is, the measurement m equals the stimulus value θ_0). The repulsive effect is further amplified because the distribution of the measurement also follows the same asymmetry.

demonstrate below, neural populations with quite different tuning characteristics, but equivalent distributions of Fisher information, can represent equivalent efficient sensory representations that lead to similar Bayesian decoding characteristics.

General predictions of the framework

The tight link between the stimulus distribution, the encoding accuracy and the shape of the likelihood function has important consequences for the resulting decoding characteristics of our Bayesian observer model. In particular, it makes two predictions with regard to perceptual bias that are surprising and counterintuitive from a standard Bayesian modeling point of view.

The first prediction is that perception can be biased away from the prior peak. A Bayesian modeling approach that assumes a priori a symmetric likelihood function predicts the percept to be biased toward the prior peak for relatively smooth prior distributions (**Fig. 2a**). The situation changes, however, if the likelihood function is asymmetric (**Fig. 2b**). Now, the asymmetry itself can lead to estimation biases (also see ref. 26). In our framework, the shape of the likelihood is asymmetric for any non-uniform stimulus distribution, with a heavier tail pointing away from the prior peak. This shape introduces a repulsive bias effect that we refer to as the likelihood repulsion. Although the effect depends on the chosen loss function, it is robust for commonly used choices (see below). The repulsive effect is further amplified when computing the expected bias over many measurements of a given stimulus value θ_0. The reason is that the distribution of these measurements in the stimulus space also follows the same asymmetry; that is, the noisy measurements and thus the position of the likelihood functions on each trial are, on average, also biased away from the true stimulus value θ_0. These observations suggest a nuanced account of perceptual biases as the net result of two bias effects, one introduced by the likelihood asymmetry and one by the prior distribution. Because of the above link, we can precisely predict perceptual biases for known natural stimulus distributions. We found that, under many conditions, the model predicted that perception was biased away from the peak of the stimulus distribution (that is, the prior belief). In particular, assuming small internal noise only and a squared-error loss function we were able to derive analytic solutions for the expected perceptual bias for arbitrary stimulus distributions (see Online Methods for details). We found that the bias was always repulsive if the prior distribution is well approximated by a monotonic function over the support of the likelihood function. This prediction is quite notable, as the 'bias towards the prior' has been considered to be a fundamental characteristic of Bayesian observer models.

The second prediction is that stimulus (external) and sensory (internal) noise differently affect perceptual bias. The difference emerges because our efficient coding assumption generally imposes an inhomogeneous sensory representation that has a different metric than the physical space. Thus, although both sources of uncertainties are ultimately jointly reflected in the noise of the sensory measurement m, their individual effects on the likelihood function are different because of the mapping function F (**Fig. 3a**). Consequently, noise added at the stimulus level leads to a different likelihood function than the equivalent noise added at the sensory level, which culminates in a different bias. Increasing sensory noise gives rise to a likelihood function that is more asymmetric in the stimulus space because the additional uncertainty is mapped from the sensory space (where it is symmetric; for example, Gaussian) to the stimulus space via the inverse mapping F^{-1}. Although the prior attraction increases as a result of the overall wider likelihood function, the increase in likelihood repulsion generally dominates, leading to a net increase in repulsive bias (**Fig. 3b**). Experimentally, we assume that sensory noise can be modulated by changing stimulus contrast or presentation time. In contrast, adding the same Gaussian noise at the stimulus level gives rise to an overall smoothed likelihood function that can be thought of as the result of convolving the original likelihood function with the Gaussian noise kernel. In this case, the asymmetry of the likelihood function does not change and the likelihood repulsion therefore remains constant. The prior attraction, however, becomes stronger because of the overall increase in likelihood width. Thus, the net bias is less repulsive and eventually becomes attractive for large noise magnitudes (**Fig. 3c**). The specific predictions, of course, depend on the shape of the stimulus noise. We focused on additive Gaussian noise because this choice allowed us to directly compare the predictions against data from psychophysical experiments that have used such noise. However, our observer model is not limited to any particular choice and it will be interesting to validate our model under asymmetric stimulus noise conditions[27,28]. **Figure 3d** summarizes the predicted noise dependencies. For comparison, we also included the predictions of a Bayesian observer model that assumes a symmetric likelihood function. The predicted bias is always attractive and grows with increasing stimulus or sensory noise. Note that systematic predictions are not possible for very large noise magnitudes. At those noise levels, the resulting biases depend on the overall shape of the prior.

In summary, our Bayesian observer model predicts that perception is often biased away from the peak of the prior. Furthermore, it predicts that internal and external noise can differentially modulate these biases: increasing internal noise increases repulsive bias, whereas increasing stimulus noise decreases repulsive bias, eventually leading

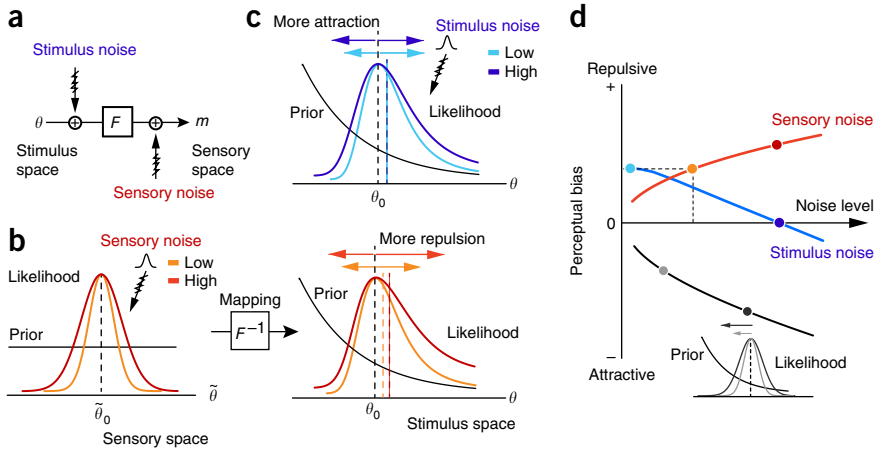

Figure 3 Prediction 2: stimulus (external) and sensory (internal) noise differentially affect perceptual bias. (**a**) Stimulus noise directly affects stimulus uncertainty and thus the likelihood function (formulated in stimulus space). The uncertainty introduced by sensory noise, however, is transformed back through the inverse of the mapping function F (equation (8), Online Methods) between sensory and stimulus space; the very reason the likelihood function is asymmetric in the first place. (**b**) Increasing the (symmetric) noise at the level of the sensory representation leads to a more asymmetric likelihood function (formulated in the stimulus space) and thus increases likelihood repulsion (dashed lines). As a result, the increase in prior attraction resulting from the increase in likelihood width is smaller than the increase in likelihood repulsion, leading to an overall net increase in repulsive bias. (**c**) In contrast, adding (symmetric) stimulus noise does not affect the asymmetry of the likelihood function (dashed lines) because the added noise essentially convolves the likelihood function with the noise kernel. The likelihood repulsion remains the same while the prior attraction grows because the overall width of the likelihood increases. As a result, the perceptual bias becomes more attractive. (**d**) Summary plot illustrating how perceptual biases depend on stimulus and sensory noise. We assumed additive Gaussian noise and a squared-error loss function. Dots correspond to the conditions shown in **b**. In general, the perceptual bias is repulsive and grows with increasing sensory noise. However, increasing stimulus noise reduces the repulsive bias, eventually leading to attractive biases for large noise levels. Note that this differential dependency on the different noise sources is a direct consequence of the inhomogeneous sensory representation imposed by efficient coding. For comparison, the black curve illustrates the expected biases for a Bayesian observer model that a priori assumes a symmetric likelihood function.

to attractive perceptual biases. These predictions are surprising and at odds with predictions of standard Bayesian observer models.

Model validation against human psychophysical data

We validated the model predictions against measured perceptual biases for two visual stimulus variables with known natural stimulus distributions, local orientation θ and spatial frequency ξ.

Orientation perception. Several studies have measured the distribution of visual orientations in natural environments by carefully analyzing natural image data. The extracted distributions are fairly robust with regard to the specifics of the analysis (that is, amplitude spectrum[29] versus dominant orientation[16,30]) and the image content (for example, indoor versus outdoor scenes[29,30]). These studies consistently reported multimodal distributions with peaks at each of the two cardinal orientations. We used a parametric approximation of the measured distribution by Girshick and colleagues[16] (**Fig. 4a**). **Figure 4b,c** shows the predicted mean biases as a function of stimulus orientation θ for different levels of sensory and stimulus noise. The predicted biases are typically repulsive and thus toward the nearest

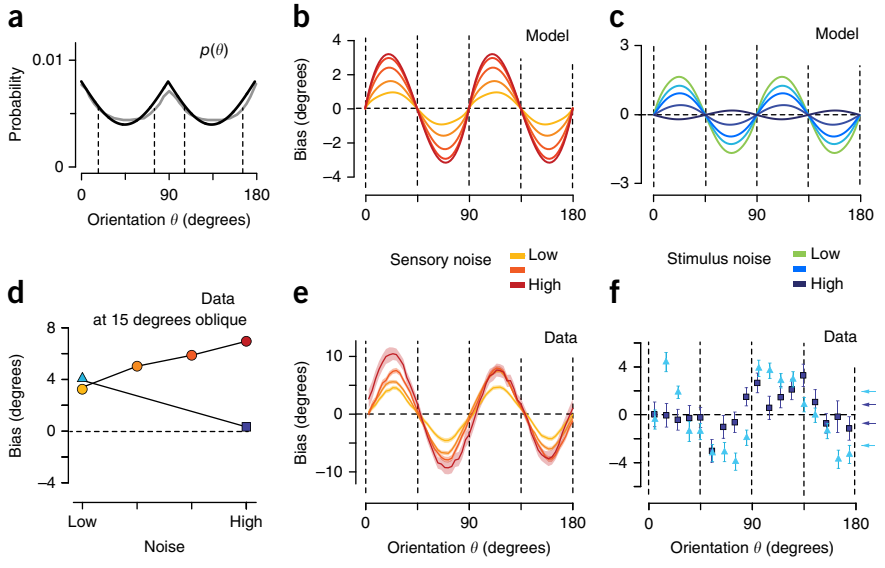

Figure 4 Biases in perceived orientation. (**a**) Measured distribution of local visual orientation in natural images (gray line, replotted from ref. 16), superimposed with the parametric description we used for the model predictions (black line: $p(\theta) = c_0(2 - |\sin \theta|)$ where c_0 is a normalization constant). (**b**) Predicted mean biases as a function of stimulus orientation θ and different levels of sensory noise; biases are generally repulsive, that is, away from the nearest cardinal orientation, with larger biases for larger noise magnitudes. (**c**) Predictions are presented as in **b** but for different levels of stimulus noise; here the repulsive biases are smaller for larger noise magnitudes, eventually becoming attractive for very large levels. Curves in **b** and **c** represent the expected bias values over the full measurement distributions. (**d**) Measured biases at 15 degrees oblique orientation[15,32] (average over all four orientations indicated by dashed lines in **a**). The biases match the predicted behavior shown in **Figure 3d** well. (**e**) Measured biases as a function of sensory noise (±1 s.e.m.). Sensory noise was modulated by different stimulus presentation times (low to high: 1,000 ms, 160 ms, 80 ms, 40 ms). Data from ref. 32 were reanalyzed. (**f**) Measured biases for two levels of additive Gaussian stimulus noise ($N = 5$ subjects, mean ± 2 s.e.m.). Arrows indicate the mean bias over all orientations in each of two corresponding quadrants (for example, top dark blue arrow: mean bias for high stimulus noise computed over the range (0,45)∪(90,135) degrees). The overall biases were clearly repulsive and were reduced for larger stimulus noise. Data are replotted from ref. 15.

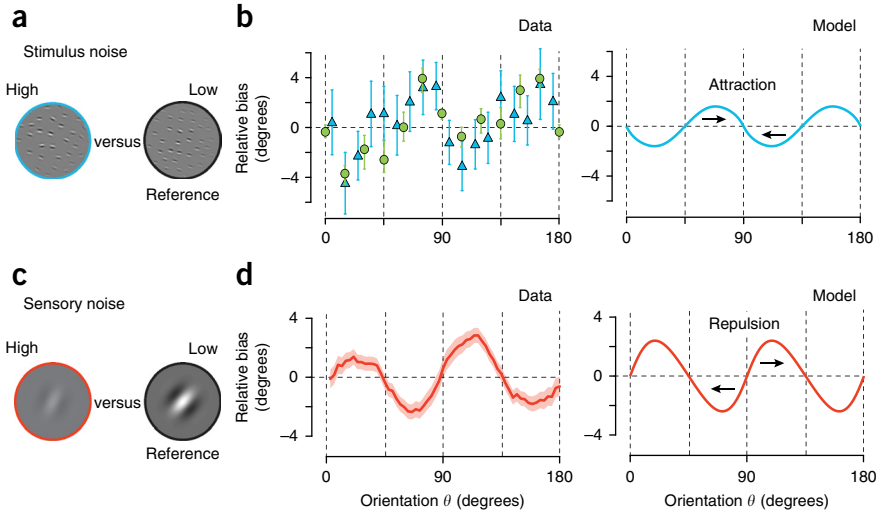

Figure 5 Relative biases in perceived orientation. Relative bias is the difference in perceived orientation between a high-noise and a low-noise stimulus (reference). (**a**) Two orientation stimuli with different levels of stimulus noise. Each stimulus consists of an array of Gabor elements, and the width of the distribution from which the orientations of the elements were sampled controls the noise level. (**b**) Measured relative biases as a function of stimulus orientation using the stimuli shown in **a**; data are replotted from ref. 15 (blue; $N = 5$ subjects, mean ± 2 s.e.m.) and ref. 16 (green; $N = 5$ subjects, mean and 95% quantile). The relative bias is attractive because the repulsive bias is smaller for the high-noise stimulus (see **Fig. 4f**). (**c**) Two orientation stimuli associated with different levels of sensory noise. Sensory noise can be modulated by stimulus contrast (this example) or presentation time[32] with lower contrast/shorter presentation time corresponding to higher sensory noise. (**d**) Measured relative bias (±1 s.e.m.) between the percepts of two stimuli with different sensory noise as a function of stimulus orientation[32]. Relative bias is repulsive because the repulsive bias is larger for larger sensory noise. Unlike previous models[15,16], the new model accounts for both relative bias patterns.

oblique orientation. Biases are zero for the cardinal and oblique orientations, yet reach their maximum for orientations that lie in between. These oblique biases have been reported as early as in the late 19th century[31]. The shape of the bias curves as a function of stimulus orientation is similar for both noise types. However, we predict that the bias amplitude grows with increasing sensory noise (**Fig. 4b**) and it decreases for increasing stimulus noise (**Fig. 4c**). Psychophysical data from two recent studies match those predictions[15,32]. **Figure 4d–f** show the measured perceptual bias for stimulus orientations at 15 degrees oblique as well as for the entire range of orientations as a function of stimulus and sensory noise.

Note that Bayesian observer models for the perception of visual orientation have been proposed before[15,16]. These models were validated against psychophysical measurements of relative bias between two stimuli with different levels of stimulus (external) noise. Specifically, both studies used the type of array stimuli shown in **Figure 5a** and measured the difference in perceived orientation between a stimulus with a high versus low stimulus noise. Although the percept of each of the two stimuli is biased toward the oblique orientations, it is less repulsive for the high-noise stimulus (**Fig. 4f**). Thus, the relative bias is indeed attractive and therefore can be accounted for by these models (**Fig. 5b**). However, such standard models cannot explain the repulsive biases and their differential noise dependencies (**Fig. 4**), nor can they account for the relative bias between a stimulus with high versus low sensory noise (**Fig. 5c**). This relative bias is again repulsive because high sensory noise leads to larger repulsive biases (**Fig. 4e**). Thus, we predict that if Girshick and colleagues had fit their Bayesian model[16] to data collected with stimuli of different sensory rather than different stimulus noise (**Fig. 5d**), their fit prior distribution would not have matched the natural stimulus distribution (**Fig. 4a**) and would have shown peaks at the oblique orientations instead. Our results suggest that the notion that perceived orientation is biased toward the cardinal axes because of a prior belief that favors cardinal orientations is simplistic.

Spatial frequency perception. Natural visual scenes are dominated by low spatial frequency content. Specifically, the empirically computed

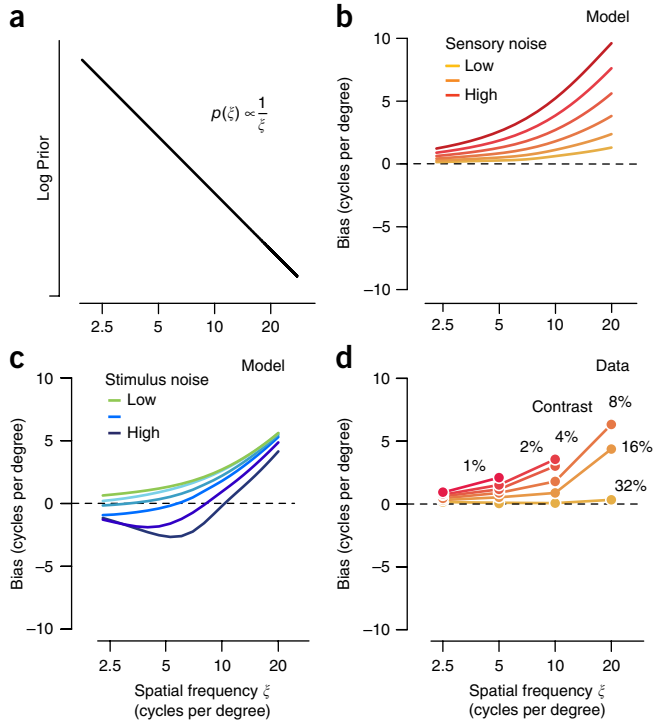

Figure 6 Biases in perceived spatial frequency. (**a**) The amplitude spectrum of spatial frequency in natural images approximately follows a power-law function of the form $p(\xi) \propto 1/\xi^{\alpha}$, with reported values for α around 1 (refs. 33,34). We assumed the spectrum to be a good proxy for the prior distribution over spatial frequency (we set $\alpha = 1$). (**b**) The predicted biases as a function of spatial frequency for different levels of sensory (internal) noise. (**c**) Predicted biases for different levels of stimulus (external) noise. (**d**) Biases in perceived spatial frequency measured for different levels of sensory noise ($N = 3$; mean). Data are replotted from ref. 36. The experiments used different levels of stimulus contrast (1, 2, 4, 8, 16 and 32%) to modulate sensory noise. Stimuli consisted of a Gabor patch with different spatial frequency. The predicted biases for stimulus noise in **c** have not been validated yet. Note that, at very low and very high spatial frequencies, the amplitude spectrum is no longer well described by a single power-law function[33]. As a result, our predictions here are limited to the intermediate frequency range.

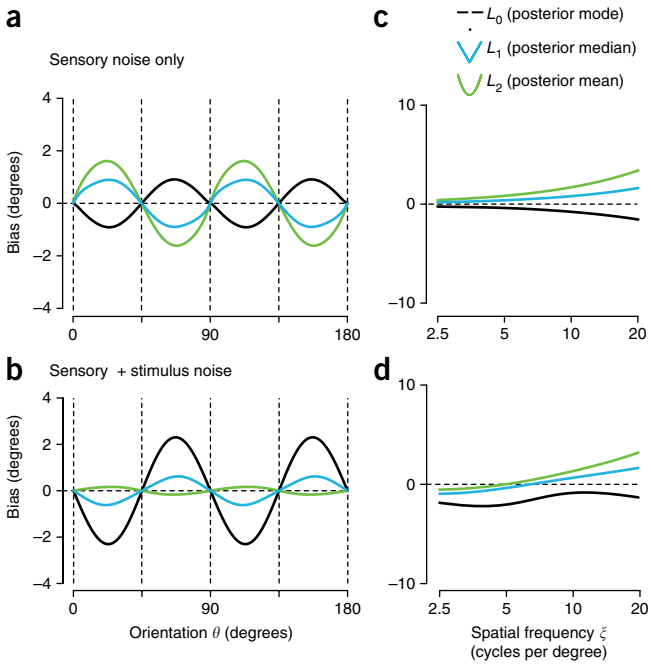

Figure 7 Predicted biases for different loss functions. (**a**) Predicted biases in perceived orientation for the observer model with L_0 norm (posterior mode, black), L_1 norm (posterior median, blue) or L_2 norm (posterior mean, green) loss function. Both the L_1 and L_2 norm predict repulsive biases, whereas the L_0 norm always leads to attractive biases. Sensory noise is fixed and identical for all three models. (**b**) Adding stimulus noise reduces the likelihood asymmetry and thus increases the attractive influence of the prior. The influence of the likelihood asymmetry is weaker with the L_1 loss than the L_2 loss, explaining the transition to an attractive bias curve. (**c**,**d**) A similar pattern is predicted for the perceptual biases in spatial frequency.

amplitude spectra of natural images robustly follow the power-law function $p(\xi) \propto 1/\xi^{\alpha}$ over a relatively large frequency range with values for α around 1 (refs. 33,34). Given that the human visual system can simultaneously sense a broad range of spatial frequencies at any given spatial location (spatial frequency channels[35]), we assumed that the empirically measured amplitude spectrum of natural images is a good proxy of the total distribution of spatial frequencies ξ an observer is exposed to in natural visual environments (**Fig. 6a**). We chose $\alpha = 1$ for simplicity, but verified that our results were robust with regard to other values in the reported range. Because $p(\xi)$ is monotonically decreasing, we predict that, in the absence of stimulus noise, perceived spatial frequency is biased toward higher frequencies across the entire frequency range. We also predict that increasing sensory noise (by for example, reducing stimulus contrast) biases the percept toward even higher frequency values (**Fig. 6b**), whereas increasing stimulus noise leads to a decrease in repulsive bias that eventually can turn into an attractive bias at low frequencies (**Fig. 6c**). Our predictions are consistent with psychophysically measured biases in perceived spatial frequency as a function of stimulus contrast[36] (**Fig. 6d**). Biases for different levels of stimulus noise have not been reported yet, but could probably be measured using synthesized stimuli with different spectral bandwidths[28].

Specifying the loss function

The proposed Bayesian observer model is fully specified for known natural stimulus distributions, with the exception of the loss function. The loss function is an integral part of any optimal Bayesian observer

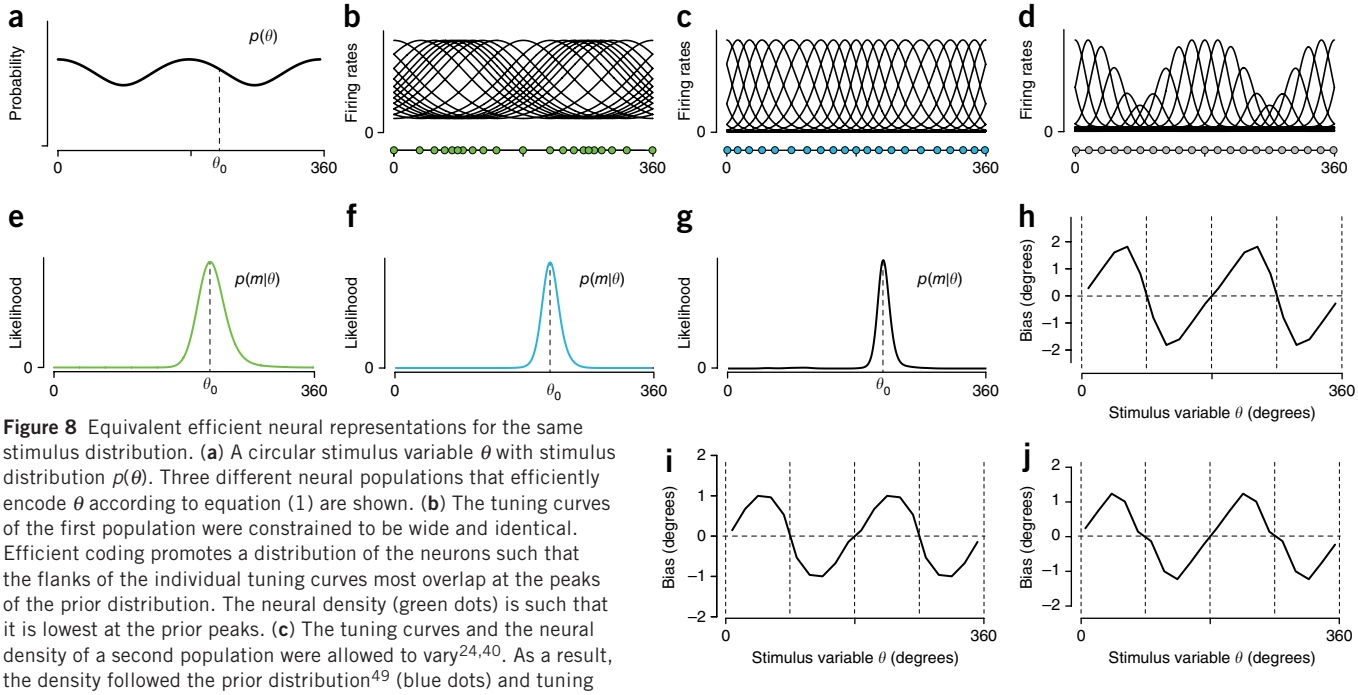

Figure 8 Equivalent efficient neural representations for the same stimulus distribution. (**a**) A circular stimulus variable θ with stimulus distribution $p(\theta)$. Three different neural populations that efficiently encode θ according to equation (1) are shown. (**b**) The tuning curves of the first population were constrained to be wide and identical. Efficient coding promotes a distribution of the neurons such that the flanks of the individual tuning curves most overlap at the peaks of the prior distribution. The neural density (green dots) is such that it is lowest at the prior peaks. (**c**) The tuning curves and the neural density of a second population were allowed to vary[24,40]. As a result, the density followed the prior distribution[49] (blue dots) and tuning curves were narrowest at the prior peaks. (**d**) The tuning curve shapes as well as the neural density (gray dots) of the third population were constrained to be identical/homogeneous. Only the gain was allowed to vary. As a result, neurons at the peak of the prior had highest gain. (**e**–**g**) Population likelihoods for all three populations (averaged over 400 presentations of the same stimulus value θ_0, assuming independent Poisson spike count variability) were similar (up to a scale factor) and showed the predicted asymmetry with the heavy tail pointing away from the nearest peak of the prior. (**h**–**j**) As a result, Bayesian decoding with prior $p(\theta)$ of all three populations resulted in similar repulsive biases. Biases were computed over 10,000 samples of the neural population response.

model, specifying how costly some perceptual errors are for the observer. The assumption is that the observer chooses an estimate (percept) that minimizes the expected loss (Online Methods). Unfortunately, it is difficult to determine the actual loss function of human observers when performing low-level perceptual tasks. Our predictions thus far assume a squared-error loss function (or L_2 norm), which is equivalent to computing the posterior mean. To explore the degree to which the model predictions depend on the specific choice of the loss function, we compared them to predictions based on two other commonly used loss functions from the L_p family: the L_0 loss (equivalent to the posterior mode) and the L_1 loss (equivalent to the posterior median).

Overall, the predictions for the L_1 and L_2 loss are qualitatively similar although the bias magnitudes are smaller for the L_1 loss (**Fig. 7a**,**c**). This is expected, as the median of the posterior is less repulsed than the mean. The effect translates to the case of added stimulus noise (**Fig. 7b**,**d**). Predictions for the L_0 loss, however, are distinctly different in that the bias is attractive in these examples. The L_0 loss is unique in the sense that it does not take into account the shape of the posterior distribution. This intuitively explains why the predicted bias is attractive: the repulsive influence of the likelihood asymmetry is masked by the unique shape of the L_0 loss function, and thus prior attraction dominates. Any symmetric loss function that employs the full information contained in the posterior distribution, however, preserves the repulsive influence of the likelihood asymmetry on the percept. Thus, the qualitative predictions of our observer model are fairly robust with regard to the specific choice of the loss function. The comparison also suggests that humans' perception of low-level stimuli is unlikely guided by a L_0 loss function, which supports previous findings[12,37].

DISCUSSION

We have investigated the idea that the natural stimulus statistics not only determine how sensory information is represented, but also how this representation is interpreted to form a percept. Specifically, we introduced a new Bayesian observer model that is constrained by efficient coding. As a result, the likelihood function and prior belief of our model are linked and jointly constrained by the natural stimulus statistics. The observer model makes two surprising and, at first sight, counterintuitive general predictions. It predicts that perceptual biases are repulsive, that is, away from the peak of the prior distribution, under many conditions. It also predicts that sensory (internal) and stimulus (external) noise differentially affect perceptual bias when the stimulus distributions are non-uniform. We confirmed these predictions using reported perceptual biases for visual orientation and spatial frequency, two perceptual variables for which the natural stimulus distributions are known. The model accounts for biases measured over a wide range of noise and experimental conditions. In particular, the model provides a theoretical explanation for repulsive biases that previously proposed Bayesian observer models have failed to account for.

Our formulation of the Bayesian observer model is based on certain assumptions. For example, we considered a particular efficient coding scheme (maximizing mutual information), although other formulations are also possible, such as minimizing redundancy[2] or reconstruction error[38], or formulations that take into account the coding requirements of downstream (motor) representations and actions[39]. Some evidence suggests that maximizing mutual information may not be the optimal encoding strategy for a decoder that minimizes mean squared-error[38,40]. In general, the choice of the encoding criterion depends on many constraints, not least on the task for which the encoded sensory information is used[41]. However, for low-level stimulus variables (such as local visual orientation and spatial frequency) that are likely to represent the sensory information for many different perceptual tasks (including the estimation tasks studied here, but also fine/coarse discrimination or categorization tasks), optimizing for a more generic information criterion may represent a good encoding strategy of the visual system[2,42]. As we show, the chosen formulation is well supported by the data, yet other choices of the efficient coding scheme and the loss function may not lead to the same predictions (see **Fig. 7**). Modeling more cognitive stimulus variables or tasks may require a modification of our model formulation.

Similarly, although we only considered scalar stimulus variables, there is behavioral evidence that the brain can rapidly learn to efficiently encode also more complex stimulus variables (for example, sound frequency spectra[43]). Efficient coding solutions for these more complex variables, however, may also differ from the solution presented here. Because different efficient coding schemes impose different constraints on the shape of the likelihood, this may lead to different predictions for perceptual biases in all these cases. Such avenues would be interesting to explore in the future, although potential model predictions might be difficult to validate experimentally.

Although not explicitly specified, we implicitly assumed a (quasi-) stationary perceptual environment, and thus stationary stimulus distributions. This assumption is probably valid for low-level stimulus variables (such as for example, spatial frequency or visual orientation), yet is likely violated in experiments that require subjects to rapidly learn a particular stimulus distribution[12,44], or during instances of perceptual adaptation. We previously proposed that the characteristic repulsive adaptation aftereffects can be explained by asymmetric likelihood functions that result from an efficient re-distribution of sensory resources according to changes in the recent stimulus history[26]. Our proposed observer model uses a more mathematically rigid formulation and, in addition, imposes a tight link between prior belief, likelihood function and stimulus distribution. It will be interesting to determine the degree to which our proposed observer model can account for adaptation aftereffects when formulated for stimulus distributions over shorter timescales.

Our model formulation does not specify how the sensory measurement was extracted from low-level sensory signals, such as generating a measurement of local visual orientation based on the high-dimensional retinal image signal. Understanding this feature extraction process is important for characterizing what form of uncertainty and ambiguity is induced at the stimulus level under natural conditions[17]. We focused on simple stimulus noise models that are sufficient to capture the typical noise characteristics of the artificial stimulus displays used in psychophysical experiments (**Fig. 4**). However, there is no principled reason why the framework could not be extended to incorporate more complex uncertainty structures.

It is worth considering the implications for a potential physiological instantiation of the proposed perceptual encoding-decoding process. We purposefully used a formulation of efficient coding that is not based on detailed assumptions about the tuning characteristics of the underlying neural representation of the sensory information. It has the advantage of being sufficiently specific to define the likelihood function and to therefore permit clear predictions of perceptual behavior, yet is general in that it is not tied to any particular neural implementation (in contrast with our initial formulation[20]). Consider, for example, a stimulus variable θ (with distribution shown in **Fig. 8a**) that is encoded in three different neural populations. Each population consists of the same number of independent Poisson neurons, but has

quite different tuning characteristics in terms of neural density, tuning curve shape and response gain (**Fig. 8b–d**). However, all three populations constitute equivalent efficient representations according to our definition because each population's Fisher information satisfies equation (1). As we would expect, the average likelihood functions computed for each population's response are similar (assuming that the gain is sufficiently large such that our assumption about noise symmetry is met) and show the predicted asymmetries (**Fig. 8e–g**). As a result, Bayesian decoding of each of the three neural population leads to similar, repulsive bias curves (**Fig. 8h–j**). We believe that physiological constraints determine the specific neural coding solution. For example, wiring constraints could limit the amount by which tuning curve widths can vary in a population, which would favor the solutions shown in **Figure 8b**,**d** over the solution shown in **Figure 8c** for a highly non-uniform stimulus distribution. Notably, this may explain the observed differences in tuning characteristics between neurons in area V1 encoding orientation and neurons in area MSTd encoding heading direction, respectively. Perceptually, both stimulus variables exhibit similar repulsive biases away from the cardinal orientations[15,32], respectively, from heading directions straight ahead or backwards[45,46]. The measured neural tuning characteristics, however, are quite different. Although the orientation tuning density and widths of neurons in V1 are loosely in agreement with the population shown in **Figure 8c**[24,47], neurons in area MSTd rather resemble the population shown in **Figure 8b**, with the majority of neurons preferentially tuned to left- and rightward directions[48]. Our findings suggest that both the V1 and the MSTd population efficiently represent a stimulus variable with similar natural distributions, leading to similar perceptual biases, yet may be subject to additional constraints at the level of implementation. However, we currently do not have a good estimate of the natural stimulus distribution for heading direction, which would allow us to confirm this conjecture.

Several neural implementations of Bayesian inference have been proposed, which use decoding mechanisms that are similar to the population vector read-out[40,49,50]. The implementations all rely on neural populations whose tuning densities match the prior distribution. Note that the population shown in **Figure 8c** has these tuning characteristics and could be readily decoded with such a population vector read-out, thereby providing a neural implementation of our observer model. Whether other, equivalent efficient encoding solutions (see for example, **Fig. 8b**,**d**) also allow for simple and physiologically plausible decoding mechanisms is an interesting question for the future.

An obvious question is how our proposed Bayesian observer model and its predictions are consistent with the results of previous studies that showed the characteristic 'biases toward the prior' behavior. First, it is important to note that the new observer model does not exclusively predict repulsive biases. For example, as stimulus noise gets large, biases can become attractive (**Fig. 4c**). The same applies when considering stimuli that are in a range where the prior is not monotonic over the support of the likelihood (for example, stimuli near the peak of a uni-modal prior distribution). In addition, measured percepts depend on the specifics of the experimental setup, and thus what looks like an attractive bias might be, for example, a relative difference between repulsive biases (**Fig. 5**). Finally, some previous results may have relied on incorrect assumptions about the stimulus distribution, again with the result that biases that appear to be attractive may actually be repulsive. The formulation of our new Bayesian observer model is general and we think it will allow us to explain perceptual biases far beyond the examples presented here, including biases that currently cannot be explained. The problem we see for such future investigations is to obtain good estimates of the relevant stimulus distributions, which is often difficult (for example, distributions of visual speed[34]). But even if this information is not available or too difficult to obtain, the proposed observer model is better constrained, allowing improved fits to psychophysical data with fewer free parameters compared with previous Bayesian modeling approaches.

Last but not least, our work addresses the common criticism that Bayesian observer models are not well constrained and thus can explain essentially any data with the appropriate *post hoc* choice of prior belief and likelihood function[13,14]. We have shared this concern to some degree, as we have expressed in the past[10]. However, we think we have addressed this criticism in a constructive way by introducing a better constrained Bayesian observer model that, at the same time, also can explain perceptual data that were previously unaccounted for. We think that Bayesian models with arbitrarily chosen parametric descriptions have served their purpose, providing an intuitive understanding of how prior beliefs may affect perception. Although the focus on prior beliefs was important, our results demonstrate that it can lead to a rather simplistic understanding of the Bayesian modeling approach, which also fails to capture various interesting aspects of perceptual behavior (such as the repulsive biases). Our new observer model is a next step in elaborating the Bayesian hypothesis, putting the focus on a more principled definition of the likelihood function and the way different noise sources affect perceptual processing.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank V. DeGardelle, S. Kouider and J. Sackur for providing their data. We also would like to express our gratitude to E. Salinas, J. Gold and D. Swingley for providing valuable feedback on previous versions of the manuscript. The work was supported by the Office of Naval Research (grant N000141110744) and the University of Pennsylvania (including a Benjamin Franklin fellowship to X.-X.W.).

AUTHOR CONTRIBUTIONS

Both authors jointly designed and performed the research, and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Attneave, F. Some informational aspects of visual perception. *Psychol. Rev.* **61**, 183 (1954).
2. Barlow, H.B. Possible principles underlying the transformation of sensory messages. in *Sensory Communication* (ed. Rosenblith, W.A.) 217–234 (MIT Press, 1961).
3. Olshausen, B.A. & Field, D.J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
4. Dan, Y., Atick, J.J. & Reid, R.C. Efficient coding of natural scenes in the lateral geniculate nucleus: Experimental test of a computational theory. *J. Neurosci.* **16**, 3351–3362 (1996).
5. Lewicki, M.S. Efficient coding of natural sounds. *Nat. Neurosci.* **5**, 356–363 (2002).
6. Helmholtz, H. *Treatise on Physiological Optics (transl.)* (Thoemmes Press, Bristol, UK, 2000).
7. Curry, R.E. A Bayesian model for visual space perception. in *Seventh Annual Conference on Manual Control NASA SP-281*, 187ff (NASA, 1972).
8. Knill, D.C. & Richards, W. *Perception as Bayesian Inference* (Cambridge University Press, 1996).
9. Körding, K.P. & Wolpert, D. Bayesian integration in sensorimotor learning. *Nature* **427**, 244–247 (2004).
10. Stocker, A.A. & Simoncelli, E.P. Noise characteristics and prior expectations in human visual speed perception. *Nat. Neurosci.* **9**, 578–585 (2006).
11. van den Berg, R., Vogel, M., Josic, K. & Ma, W.J. Optimal inference of sameness. *Proc. Natl. Acad. Sci. USA* **109**, 3178–3183 (2012).

12. Jazayeri, M. & Shadlen, M.N. Temporal context calibrates interval timing. *Nat. Neurosci.* **13**, 1020–1026 (2010).
13. Jones, M. & Love, B.C. Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behav. Brain Sci.* **34**, 169–188 (2011).
14. Bowers, J.S. & Davis, C.J. Bayesian just-so stories in psychology and neuroscience. *Psychol. Bull.* **138**, 389 (2012).
15. Tomassini, A., Morgan, M.J. & Solomon, J.A. Orientation uncertainty reduces perceived obliquity. *Vision Res.* **50**, 541–547 (2010).
16. Girshick, A.R., Landy, M.S. & Simoncelli, E.P. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nat. Neurosci.* **14**, 926–932 (2011).
17. Geisler, W.S., Najemnik, J. & Ing, A.D. Optimal stimulus encoders for natural tasks. *J. Vis.* **9**, 17.1–17.16 (2009).
18. Burge, J. & Geisler, W.S. Optimal defocus estimation in individual natural images. *Proc. Natl. Acad. Sci. USA* **108**, 16849–16854 (2011).
19. Brayanov, J.B. & Smith, M.A. Bayesian and "Anti-Bayesian" biases in sensory integration for action and perception in the size-weight illusion. *J. Neurophysiol.* **103**, 1518–1531 (2010).
20. Wei, X.-X. & Stocker, A.A. Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. *Adv. Neural Inf. Process. Syst.* **25**, 1313–1321 (2012).
21. Linsker, R. Self-organization in a perceptual network. *Computer* **21**, 105–117 (1988).
22. Brunel, N. & Nadal, J.-P. Mutual information, Fisher information, and population coding. *Neural Comput.* **10**, 1731–1757 (1998).
23. McDonnell, M.D. & Stocks, N.G. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Phys. Rev. Lett.* **101**, 058103 (2008).
24. Ganguli, D. & Simoncelli, E.P. Implicit encoding of prior probabilities in optimal neural populations. *Adv. Neural Inf. Process. Syst.* **23**, 658–666 (2010).
25. Fechner, G.T. *Elemente der Psychophysik* (Breitkopf und Haertel, Leipzig, 1860).
26. Stocker, A.A. & Simoncelli, E.P. Sensory adaptation within a Bayesian framework for perception. *Adv. Neural Inf. Process. Syst.* **18**, 1289 (2006).
27. Webb, B.S., Ledgeway, T. & McGraw, P.V. Relating spatial and temporal orientation pooling to population decoding solutions in human vision. *Vision Res.* **50**, 2274–2283 (2010).
28. Putzeys, T., Bethge, M., Wichmann, F., Wagemans, J. & Goris, R. A new perceptual bias reveals supoptimal population decoding of sensory responses. *PLoS Comput. Biol.* **8**, e1002453 (2012).
29. Switkes, E., Mayer, M.J. & Sloan, J.A. Spatial frequency analysis of the visual environment: anisotropy and the carpentered environment hypothesis. *Vision Res.* **18**, 1393–1399 (1978).
30. Coppola, D.M., Purves, H.R., McCoy, A.N. & Purves, D. The distribution of oriented contours in the real world. *Proc. Natl. Acad. Sci. USA* **95**, 4002–4006 (1998).
31. Jastrow, J. Studies from the University of Wisconsin: on the judgment of angles and positions of lines. *Am. J. Psychol.* **5**, 214–248 (1892).
32. de Gardelle, V., Kouider, S. & Sackur, J. An oblique illusion modulated by visibility: non-monotonic sensory integration in orientation processing. *J. Vis.* **10**, 6 (2010).
33. Ruderman, D.L. The statistics of natural images. *Network* **5**, 517–548 (1994).
34. Dong, D.W. & Atick, J.J. Statistics of natural time-varying images. *Network* **6**, 345–358 (1995).
35. Campbell, F.W. & Robson, J.G. Application of Fourier analysis to the visibility of gratings. *J. Physiol. (Lond.)* **197**, 551 (1968).
36. Georgeson, M.A. & Ruddock, K.H. Spatial frequency analysis in early visual processing. *Phil. Trans. R. Soc. Lond. B* [and discussion] **290**, 11–22 (1980).
37. Körding, K.P. & Wolpert, D. The loss function of sensorimotor learning. *Proc. Natl. Acad. Sci. USA* **101**, 9839–9842 (2004).
38. Wang, Z., Stocker, A.A. & Lee, D.D. Optimal neural tuning curves for arbitrary stimulus distributions: Discrimax, Infomax and minimum Lp loss. *Adv. Neural Inf. Process. Syst.* **25**, 2177–2185 (2012).
39. Salinas, E. How behavioral constraints may determine optimal sensory representations. *PLoS Biol.* **4**, e387 (2006).
40. Ganguli, D. & Simoncelli, E.P. Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. *Neural Comput.* **26**, 2103–2134 (2014).
41. Simoncelli, E.P. & Olshausen, B.A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24**, 1193–1216 (2001).
42. Laughlin, S.B. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch. C.* **36**, 910–912 (1981).
43. Stilp, C.E. & Kluender, R.K. Efficient coding and statistically optimal weighting of covarinace among acoustic attributes in novel sounds. *PLoS ONE* **7**, e30845 (2012).
44. Chalk, M., Seitz, A.R. & Series, P. Rapidly learned stimulus expectations alter perception of motion. *J. Vis.* **10**, 2 (2010).
45. Crane, B.T. Direction specific biases in human visual and vestibular heading perception. *PLoS ONE* **7**, e51383 (2012).
46. Cuturi, L.F. & MacNeilage, P.R. Systematic biases in human heading estimation. *PLoS ONE* **8**, e56862 (2013).
47. Rose, D. & Blakemore, C. An analysis of orientation selectivity in the cat's visual cortex. *Exp. Brain Res.* **20**, 1–17 (1974).
48. Gu, Y., Fetsch, C.R., Adeyemo, B., DeAngelis, G.C. & Angelaki, D.E. Decoding of MSTd population activity accounts for variations in the precision of heading perception. *Neuron* **66**, 596–609 (2010).
49. Fischer, B.J. Bayesian estimates from heterogeneous population codes. *Proc. Int. Jt. Conf. Neural Netw.*, 1–7 (IEEE, 2010).
50. Wei, X.-X. & Stocker, A.A. Bayesian inference with efficient neural population codes. in *Artificial Neural Networks and Machine Learning–ICANN 2012* (eds. Villa, A., Duch, W., Erdi, P., Masulli, F. & Palm, G.) 523–530 (Springer, 2012).

ONLINE METHODS

Efficient encoding. We assumed an efficient coding constraint that maximizes the mutual information between a scalar stimulus variable θ and its sensory representation m (refs. 2,21). Fisher information $J(\theta)$ defined as

$$J(\theta) = \int \left(\frac{\partial \ln p(m|\theta)}{\partial \theta} \right)^2 p(m|\theta) dm \tag{2}$$

can be used to specify a bound on mutual information in the asymptotic limit of vanishing noise[22]. Assuming the bound is tight, mutual information can be expressed as[23]

$$I[\theta, m] = \frac{1}{2} \ln \left(\frac{S^2}{2\pi e} \right) - KL\left(p(\theta) \,||\, \frac{\sqrt{J(\theta)}}{S} \right) \tag{3}$$

where $S = \int_\theta \sqrt{J(\theta)} d\theta$. S can be intuitively understood as the total amount of coding resource available. $KL(||)$ represents the Kullback-Leibler (KL) divergence[51], which is always non-negative.

The goal is to choose $J(\theta)$ to maximize $I[\theta, m]$ for a fixed $p(\theta)$. Technically, this requires us to impose an additional constraint because $I[\theta, m]$ is not limited otherwise. To bound $I[\theta, m]$ and create a well-posed optimization problem, we assume

$$S = \int_\theta \sqrt{J(\theta)} d\theta \le C \tag{4}$$

With this constraint, maximizing mutual information requires the above KL divergence term to be zero. This is equivalent to

$$p(\theta) \propto \sqrt{J(\theta)} \tag{5}$$

Because the mutual information $I[\theta, m]$ is invariant with respect to any re-parameterization of θ, it is desirable that the constraint also shares this property. The chosen constraint (equation (4)) is invariant whereas constraints using other power functions of $J(\theta)$, for example, $\int_\theta J(\theta) d\theta$, are not.

Bayesian decoding. Bayesian decoding consists of defining an estimate $\hat{\theta}(m)$ of the stimulus value given measurement m such that the expected loss according to a loss function $L(\hat{\theta}(m), \theta)$ is minimal, that is,

$$\arg\min_{\hat{\theta}} \int p(\theta|m) L(\hat{\theta}(m), \theta) d\theta \tag{6}$$

The key quantity here is $p(\theta|m)$, which represents the posterior probability distribution over θ for a given sensory measurement m. According to Bayes' rule[52], the posterior can be computed as $p(\theta|m) \propto p(\theta)p(m|\theta)$, where $p(\theta)$ is the prior belief and $p(m|\theta)$ represents the likelihood function on θ. For the specific L_0, L_1 and L_2 loss functions considered in this paper, the optimal estimator $\hat{\theta}(m)$ is the posterior mode, median, and mean, respectively.

Perceptual bias for L_2 loss (sensory noise only). We can analytically derive the expected bias $b(\theta)$ of our Bayesian observer model in the case of a squared error loss function (L_2 loss) assuming no stimulus noise. The posterior mean can be computed in terms of the likelihood function and the prior belief as following

$$\hat{\theta}_{L_2}(m) = \frac{\int \theta p(m|\theta) p(\theta) d\theta}{\int p(m|\theta) p(\theta) d\theta} \tag{7}$$

With the efficient coding assumption above, equation (5), we can now express the bias as a function of the prior belief. First, we define a one-to-one mapping $F(\theta)$ that transforms the stimulus space to a *sensory space* with units $\tilde{\theta} = F(\theta)$ for which the Fisher information (as well as the stimulus distribution) is uniform[25,42]. The mapping is defined as

$$F(\theta) = \int_{-\infty}^{\theta} p(\chi) d\chi \tag{8}$$

which is the cumulative of the prior distribution $p(\theta)$.

We then re-write the estimate equation (7) by replacing θ with the inverse of the mapping, that is, $\theta = F^{-1}(\tilde{\theta})$. Given a sensory measurement m, we can write the estimator as

$$\hat{\theta}_{L_2}(m) = \frac{\int F^{-1}(\tilde{\theta}) p(m|F^{-1}(\tilde{\theta})) p(F^{-1}(\tilde{\theta})) dF^{-1}(\tilde{\theta})}{\int p(m|F^{-1}(\tilde{\theta})) p(F^{-1}(\tilde{\theta})) dF^{-1}(\tilde{\theta})} = \frac{\int F^{-1}(\tilde{\theta}) p(m|F^{-1}(\tilde{\theta})) d\tilde{\theta}}{\int p(m|F^{-1}(\tilde{\theta})) d\tilde{\theta}} \tag{9}$$

with

$$K(m, \tilde{\theta}) = \frac{p(m|F^{-1}(\tilde{\theta}))}{\int p(m|F^{-1}(\tilde{\theta})) d\tilde{\theta}} \tag{10}$$

we can further simplify the notation and get

$$\hat{\theta}_{L_2}(m) = \int F^{-1}(\tilde{\theta}) K(m, \tilde{\theta}) d\tilde{\theta} \tag{11}$$

In order to get the expected value of the estimate $\langle \hat{\theta}_{L_2} \rangle$ for a particular stimulus value θ_0, we marginalize equation (11) over the measurement space M for θ_0, thus

$$\begin{aligned} \langle \hat{\theta}_{L_2} \rangle_{\theta_0} &= \int\int_M p(m|\theta_0) F^{-1}(\tilde{\theta}) K(m, \tilde{\theta}) dm d\tilde{\theta} \\ &= \int F^{-1}(\tilde{\theta}) \int_M p(m|\theta_0) K(\boldsymbol{m}, \tilde{\theta}) dm d\tilde{\theta} \\ &= \int F^{-1}(\tilde{\theta}) \mathcal{L}_{\theta_0}(\tilde{\theta}) d\tilde{\theta}, \end{aligned} \tag{12}$$

where we define

$$\mathcal{L}_{\theta_0}(\tilde{\theta}) = \int_M p(m|\theta_0) K(m, \tilde{\theta}) dm \tag{13}$$

Therefore, $\mathcal{L}_{\theta_0}(\tilde{\theta})$ is the expected normalized likelihood function expressed in the sensory space given a particular stimulus value θ_0. We assume that $\mathcal{L}_{\theta_0}(\tilde{\theta})$ is symmetric around the true stimulus value $\tilde{\theta}_0$ in this space. Thus, with equation (11) we then can compute the expected bias at θ_0 as

$$b(\theta_0) = \int F^{-1}(\tilde{\theta}) \mathcal{L}_{\theta_0}(\tilde{\theta}) d\tilde{\theta} - F^{-1}(\tilde{\theta}_0) \tag{14}$$

Assuming the prior density to be smooth, we expand F^{-1} in the neighborhood $(\tilde{\theta}_0 - h,\ \tilde{\theta}_0 + h)$, which covers the support of the likelihood function. Using a first-order Taylor expansion with mean-value form of the remainder, we get

$$F^{-1}(\tilde{\theta}) = F^{-1}(\tilde{\theta}_0) + F^{-1}(\tilde{\theta}_0)'(\tilde{\theta} - \tilde{\theta}_0) + \frac{1}{2} F^{-1}(\tilde{\theta}_x)''(\tilde{\theta} - \tilde{\theta}_0)^2 \tag{15}$$

with $\tilde{\theta}_x$ guaranteed to exist in between $\tilde{\theta}_0$ and $\tilde{\theta}$. By re-writing equation (14) in terms of this expansion, we find that

$$\begin{aligned} b(\theta_0) &= \int_{\tilde{\theta}_0 - h}^{\tilde{\theta}_0 + h} \frac{1}{2} F^{-1}(\tilde{\theta}_x)''_{\tilde{\theta}} (\tilde{\theta} - \tilde{\theta}_0)^2 \mathcal{L}_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{2} \int_{\tilde{\theta}_0 - h}^{\tilde{\theta}_0 + h} \left(\frac{1}{p(F^{-1}(\tilde{\theta}_x))} \right)'_{\tilde{\theta}} (\tilde{\theta} - \tilde{\theta}_0)^2 \mathcal{L}_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{2} \int_{\tilde{\theta}_0 - h}^{\tilde{\theta}_0 + h} - \left(\frac{p(\theta_x)'_\theta}{p(\theta_x)^3} \right) (\tilde{\theta} - \tilde{\theta}_0)^2 \mathcal{L}_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \\ &= \frac{1}{4} \int_{\tilde{\theta}_0 - h}^{\tilde{\theta}_0 + h} \left(\frac{1}{p(\theta_x)^2} \right)'_\theta (\tilde{\theta} - \tilde{\theta}_0)^2 \mathcal{L}_{\theta_0}(\tilde{\theta}) d\tilde{\theta} \end{aligned} \tag{16}$$

In general, there is no simple rule to judge the sign of $b(\theta_0)$, because θ_x varies with θ and the sign of $\left(\frac{1}{p(\theta_x)^2} \right)'$ thus may change. However, if the prior is monotonic on the interval $F^{-1}((\tilde{\theta}_0 - h, \tilde{\theta}_0 + h))$ then the sign of $\left(\frac{1}{p(\theta_x)^2} \right)'$ is always the same as the sign of $\left(\frac{1}{p(\theta_0)^2} \right)'$, and therefore, the sign

doi:10.1038/nn.4105

of $b(\theta_0)$ is the same as the sign of $\left(\frac{1}{p(\theta_0)^2}\right)'$. This means that the bias and the local slope of the prior have opposite signs. It implies that the bias is repulsive, that is, away from the peak of the prior.

Additionally, in the small noise regime where the likelihood is sufficiently narrow, the prior can always be approximated as being monotonic over the support of the likelihood function. Due to the continuity of $\left(\frac{1}{p(\theta)^2}\right)'$, we can approximate $\left(\frac{1}{p(\theta_x)^2}\right)'$ by $\left(\frac{1}{p(\theta_0)^2}\right)'$ and thus write the bias as

$$b(\theta_0) \approx C\left(\frac{1}{p(\theta_0)^2}\right)' \tag{17}$$

where C is a positive constant.

The key assumption we made in the above derivation is that the average likelihood function $\mathcal{L}_{\theta_0}(\tilde{\theta})$ in the sensory space ($\tilde{\theta}$) is symmetric. The dimensionality of the measurement m is not important, that is, m can be a scalar or a vector (for example, response vector of a neural population), as long as the assumption that $\mathcal{L}_{\theta_0}(\tilde{\theta})$ is symmetric is approximately true.

Perceptual bias under more general conditions. Under more general conditions that include stimulus noise and/or different loss functions, the expected perceptual bias can no longer be computed analytically. However, numerical solutions can be computed for general conditions according to the encoding-decoding cascade description of the proposed Bayesian observer model. In particular, we can distinguish the effect of stimulus versus sensory noise (**Figs. 4** and **6**) by modeling the sensory measurement m as

$$m = F(\theta + \delta_s) + \delta_n \tag{18}$$

where δ_s represents the stimulus noise (expressed in stimulus space) and δ_n the sensory noise (expressed in sensory space). We assume the sensory noise to be Gaussian (respectively, vonMises) distributed, and the stimulus noise to follow the actual noise distributions used in the psychophysical experiments we modeled (often Gaussian/vonMises distributed as well). The transformation F that imposes the Efficient coding constraint determines how the stimulus noise is mapped to the sensory space (equation (8)). For any stimulus value θ_0 the conditional probability $p(m|\theta_0)$ can be computed according to equation (18) and the specific noise distributions. For each m, we can numerically compute the Bayesian estimator $\hat{\theta}(m)$ according to a specific loss function (L_0, L_1, L_2) using equation (6). Finally, for any given stimulus value θ_0, the expected bias $b(\theta_0)$ can be computed by marginalizing the estimate $\hat{\theta}(m)$ over the measurement distribution $p(m|\theta_0)$ and then subtracting the true value θ_0, thus

$$b(\theta_0) = \int \hat{\theta}(m) p(m|\theta_0) dm - \theta_0 \tag{19}$$

Neural simulation. We applied a little trick in order to generate three neural populations that have different tuning characteristics yet match in their Fisher information $J(\theta)$ (up to a scaling factor) and satisfy the efficiency constraint equation (5). We first generated the population in **Figure 8b** by assuming that it consists of $N = 20$ neurons with wide and uniform tuning curves (von Mises distribution) whose preferred tuning follow an arbitrary density distribution $d(\theta) \propto 1.2 - |\cos\theta|$. We then computed the population Fisher information assuming independent Poisson noise, and with equation (5) derived the stimulus distribution (i.e., the prior belief) $p(\theta)$ (**Fig. 8a**). The tuning curves of the second population (**Fig. 8c**) were obtained by re-parameterizing a set of homogeneous tuning curves through the cumulative prior $F(\theta)$ as previously proposed[24,40]. To create the third neural population in (**Fig. 8d**), we started from a homogeneous set of tuning curves with relatively narrow tuning widths and adjusted the gain of individual neurons such that the square-root of the population Fisher information matched the prior distribution. Numerically, this is done via a non-negative least-squares fit. These procedures guaranteed that all three populations have identical Fisher information (up to a scale factor) and thus are efficient representations of the prior distribution. The likelihoods shown in **Figure 8e–g** represent the average likelihoods computed over 400 samples of the population responses for a fixed stimulus value θ_0. The biases (**Fig. 8h–j**) are computed by drawing 10,000 samples assuming independent Poisson-spiking neuron models, and calculating the average bias of the Bayes' least-squares estimator over the samples while exploiting the symmetry in the stimulus distribution.

Data re-analysis. The bias curves shown in **Figure 4e** were obtained by re-analyzing the data set presented by DeGardelle and colleagues[32]. In their experiments, stimulus orientation was randomly sampled over the entire range (i.e., [0, 180] degrees). Bias was computed by averaging the trials over a sliding window (3 degrees size). The resulting bias $b(\theta)$ was then further smoothed with a boxcar filter with width w = 45 degrees. We performed this analysis for four stimulus conditions corresponding to stimulus presentation times of 40, 80, 160, and 1,000 ms. For these conditions, the shape and amplitudes of the bias curves were robust with regard to the chosen bin size and the width of the smoothing kernel. A fifth stimulus condition corresponding to a presentation time of 20 ms was excluded in our analysis because the amplitude of the bias curve was dependent on the bin size, making it impossible to reliably determine the magnitude of the bias. The error bars for individual orientations θ_0 in **Figure 4e** represent the circular standard error, which was estimated based on the data samples within the window $[\theta_0 \pm 22.5]$ degrees. Relative bias shown in **Figure 5d** was calculated by taking the difference between the biases corresponding to the 160 ms and 1,000 ms stimulus presentation conditions reported in **Figure 4e**. The error bar in **Figure 5d** was calculated as the square root of the sum of the squared s.e.m. in **Figure 4e** (160-ms and 1,000-ms conditions).

The bias curves shown in **Figures 4f**, **5b** and **6d** were obtained by extracting the data points from the original document files[15,16,36] using a dedicated software tool (GraphClick, http://www.arizona-software.ch/). Note that the biases in **Figure 6d** were originally reported as relative change (percentage increase). We transformed these relative values into absolute values with units of cycles per degree.

Code availability. The simulation results shown in **Figures 4–8** were generated by code written in R (free statistical software package, Free Software Foundation). The scripts are freely available on request.

A **Supplementary Methods Checklist** is available.

51. Kullback, S. & Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86 (1951).
52. Mr. Bayes & Mr. Price. An essay towards solving a problem in the doctrine of chances, by the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S. *Philos. Trans.* **53**, 370–418 (1763).