Cognition 232 (2023) 105334

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Efficient sensory encoding predicts robust averaging

Long Ni, Alan A. Stocker *

Department of Psychology, University of Pennsylvania, USA

ARTICLE INFO

Keywords:
Ensemble perception
Efficient coding
Hierarchical inference

ABSTRACT

Not every item in a stimulus ensemble equally contributes to the perceived ensemble average. Rather, items with feature values close to the ensemble mean (inlying items) contribute stronger compared to those items whose feature values are further away from the mean (outlying items). This nonuniform weighting process, named robust averaging, has been interpreted as evidence against an optimal integration of sensory information. Here, however, we show that robust averaging naturally emerges from an optimal integration process when sensory encoding is efficiently adapted to the ensemble statistics in the experiment. We demonstrate that such a model can accurately fit several existing datasets showing robust perceptual averaging in discriminating low-level stimulus features such as orientation. Across various feature domains, our model accurately predicts subjects' decision accuracy and nonuniform weighting profile, and both their dependency on the specific stimulus distribution in the experiments. Our results suggest that the human visual system forms efficient sensory representations on short time-scales to improve overall decision performance.

1. Introduction

Ensemble perception refers to a perceptual integration process that extracts certain summary statistics from an array of similar stimulus items. Empirical studies often focus on the ensemble average of a particular stimulus feature shared among these items. Numerous experiments have demonstrated that observers are able to rapidly estimate the average value of both low- and high-level visual features of an stimulus ensemble (for a review, see Whitney & Yamanashi-Leib, 2018). However, individual items typically do not equally contribute to these estimates. For instance in spatial averaging, items that are more salient or attended to are more heavily weighted than those that are less salient or unattended to (De Fockert & Marchant, 2008; Im, Park, & Chong, 2015; Vandormael, Castañón, Balaguer, Li, & Summerfield, 2017). Furthermore, when averaging over time observers typically assign larger weights to the most recent or earliest items in the sequence (Hubert-Wallander & Boynton, 2015; Tong, Dubé, & Sekuler, 2019). Larger weights are also given to more informative items in the ensemble (i.e., associated with lower variances) in sequential averaging of spatial locations (Juni, Gureckis, & Maloney, 2012). Similarly, earlier work has shown that items that are statistical outliers in an ensemble seem to contribute less to subjects' estimates of the average feature value of the ensemble (Anderson, 1968; Epstein, Quilty-Dunn, Mandelbaum, & Emmanouil, 2020; Rosenbaum, de Gardelle, & Usher, 2021; Spencer, 1963).

Regression analyses can help to quantitatively characterize the weighting profiles in human percepts of ensemble averages. The results of such analyses, however, are difficult to interpret when applied to subjects' direct reports of their perceived ensemble averages (e.g. by the method of adjustment) because potential response biases not related to the sensory weighting of the ensemble stimuli may strongly affect the regression weights. When applied to subjects' reports in binary discrimination tasks, however, the influence of non-perceptual biases is minimized, thus providing a reliable quantitative characterization of how subjects weight individual items of a stimulus ensemble. In a series of recent discrimination experiments, subjects were presented with visual stimulus ensembles that consisted of a set of items varying in terms of orientation (Li, Castañón, Solomon, Vandormael, & Summerfield, 2017) or both shape and color (de Gardelle & Summerfield, 2011). Subjects' task was to discriminate the average feature value of the ensemble relative to a reference value. Across all three feature domains (color, shape and orientation), regression analyses revealed that items with values close to the set mean (inlying elements) contributed more to the subjects' decisions than items away from the set mean (outlying elements), an averaging process named "robust averaging" (de Gardelle & Summerfield, 2011; Li et al., 2017). Robust averaging seems difficult to justify from an optimal performance perspective as it obviously introduces systematic errors in the percept of the ensemble average. However, downweighting outliers can be beneficial if ensemble perception is prone with late decision noise. Thus, robust average has been considered to represent a purposeful trade-off between accuracy and robustness (Li et al., 2017; Vandormael et al., 2017).

* Correspondence to: Computational Perception and Cognition Laboratory, Goddard laboratories, Rm 421, 3710 Hamilton walk, Philadelphia, PA 19104, USA.
E-mail address: astocker@psych.upenn.edu (A.A. Stocker).

https://doi.org/10.1016/j.cognition.2022.105334
Received 24 June 2022; Received in revised form 16 November 2022; Accepted 20 November 2022
Available online 5 December 2022

In this paper, we present an alternative interpretation. We demonstrate that robust averaging naturally emerges from an efficient sensory representation of the stimulus ensemble ("efficient coding" (Attneave, 1954; Barlow, 1961)) in a perceptual process that optimally integrates sensory information over the ensemble. Intuitively, efficient coding postulates that the limited sensory bandwidth of a perceptual system is optimally allocated according to the statistical regularities of the sensory input such that stimuli that occur frequently are more accurately represented than those occurring rarely (Wei & Stocker, 2016). We have previously demonstrated how efficient coding can provide a powerful constraint for the formulation of Bayesian observer models that can account for various aspects of perceptual bias and variability that traditional Bayesian observer models are not able to explain (Wei & Stocker, 2012, 2015). Furthermore, assuming efficient representations according to the natural, long-term stimulus statistics, led to the discovery of a lawful relation between perceptual bias and discriminability (Wei & Stocker, 2017) and allowed to successfully predict the neural encoding characteristics of stimulus features in visual cortex from perceptual behavior alone (Zhang & Stocker, 2022). Here we propose a similar, hierarchical Bayesian observer model for subjects' perceived average of an stimulus ensemble, yet consider that efficient coding can also operate on shorter timescales. We fit the data from several experiments presented in two previous studies reporting robust perceptual averaging (de Gardelle & Summerfield, 2011; Li et al., 2017). Our key assumption is that sensory representations are rapidly adapted to efficiently reflect the statistical distributions of the stimulus ensembles within each experiment. We demonstrate that such a model not only accurately accounts for the observed robust averaging behavior but also predicts in detail how the nonuniform weighting profiles depend on the specific stimulus distributions in the experiments. Our results provide a normative explanation for robust averaging in ensemble coding of low-level visual features, representing the outcome of an optimal integration process constrained by limited representational bandwidth.

2. Methods

2.1. Data

We reanalyzed experimental data from two previous discrimination studies, both reporting robust perceptual averaging in ensemble perception. In the following, we briefly describe the different experiments but refer the reader to the original articles for more details.

de Gardelle and Summerfield (2011) tested perceptual averaging for stimulus ensembles varying in shape and color. Ensembles consisted of 8 items equally spaced on a virtual circle (Fig. 1A). Each of the two stimulus features was varied independently within a certain feature range, mapped to a value between 0 and 1. On each trial, feature values for each item were independently sampled from Gaussian distributions with means and variances randomly selected from a predefined set: means varied in two levels on either side of an implicit category boundary, and variances were either low, medium, or high. There were four experiments in total, each with a different number of total subjects ($N_1 = 31, N_2 = 14, N_3 = 16, N_4 = 24$). For each experiment, subjects were told which feature dimension was task-relevant. Subjects were asked to judge whether the average shape (or color) of the 8 elements was more circular/square (or more red/blue) relative to a category boundary set at 0.5. Each subject performed approximately 1000 trials for each experiment. Experiments 1–3 differed in the absolute values of means and variances of the Gaussian distributions. Experiments 4a and 4b used category boundaries set at 0.75 (red/purple) and 0.25 (purple/blue), respectively. Also note that in Experiments 3 and 4, feature values were resampled when necessary to ensure that the mean and standard deviation of the stimulus ensemble on each trial matched the generic values.

Li et al. (2017) tested averaging behavior for stimulus ensembles varying in orientation. Stimulus ensembles consisted of 8 orientated gratings (Fig. 1B). Subjects ($N = 24$) had to judge whether the average orientation was clockwise/anti-clockwise relative to a reference grating displayed in the center. The reference gating was either fixed throughout a block of trials (fixed condition) or varied across trials (variable condition). Orientations of individual items were independently sampled from a Gaussian distribution with varying mean and variance across trials. Means were assigned one of two values on either side of the reference. Each mean was randomly paired with either a high or low variance. Orientation of the reference grating was randomly sampled from a uniform distribution across blocks in the fixed condition and across individual trials in the variable condition. Each subject ran 8 blocks, each consisting of 128 trials, for each of the two conditions.

2.2. Regression analysis

In order to recover the strength with which each item in the stimulus ensemble contributes to the perceived feature average (i.e., weighting profile), we use a logistic regression analysis as in the original studies (de Gardelle & Summerfield, 2011; Li et al., 2017). Specifically, we assume subjects' decision probability $p(\hat{C})$ (e.g., ensemble average orientation clockwise/counter-clockwise of reference orientation) is described by

$$p(\hat{C}) = \frac{1}{1 + e^{-(w_0 + WS)}} , \tag{1}$$

where $WS = \sum_i w_i s_i$ with w_i the weight and s_i the feature value (e.g., orientation) of each item i in the ensemble, and w_0 an offset parameter. The position i of each item in the ensemble is either given by the rank of the item's feature value in each ensemble in ascending order (de Gardelle & Summerfield, 2011), or by the item's feature value according to 8 equally spaced bins over the range of [−45, 45] deg relative to the reference orientation (Li et al., 2017). Subjects' binary responses across trials are used as the dependent variable. Robust averaging refers to a nonuniform weighting profile that shows higher weights for items in the ensemble with feature values close to the decision boundary (inlying ranks or bins) and lower weights to items with features values further away from (outlying ranks or bins). This is illustrated in Fig. 1C.

Note that we utilize the same regression analysis to test the robust averaging behavior of the model. That is, we run our model on the exact same ensemble stimuli as used in the original studies, compute the model's decision probability $p(\hat{C})$ for each trial (see below), and then perform the regression analysis (Eq. (1)) on the overall set of decision probabilities.

2.3. Bayesian observer model constrained by efficient coding

We formulate a hierarchical Bayesian observer model for the specific decision task in the ensemble perception experiments (de Gardelle & Summerfield, 2011; Li et al., 2017). Crucially, we assume that the observer has limited sensory resources/bandwidth and thus efficiently employs those resources according to the stimulus distribution. In the following, we will refer to our model simply as "efficient Bayesian observer model".

We start by describing the efficient encoding of sensory information. We assume that sensory encoding is aimed at maximizing the mutual information $I[s, x]$ between the stimulus feature s and its noisy sensory representation x given an overall resource limit (Wei & Stocker, 2015). This assumption imposes a constraint on encoding precision (measured as Fisher information $J(s)$) in terms of the stimulus distribution $p(s)$ such that

$$p(s) \propto \sqrt{J(s)} . \tag{2}$$

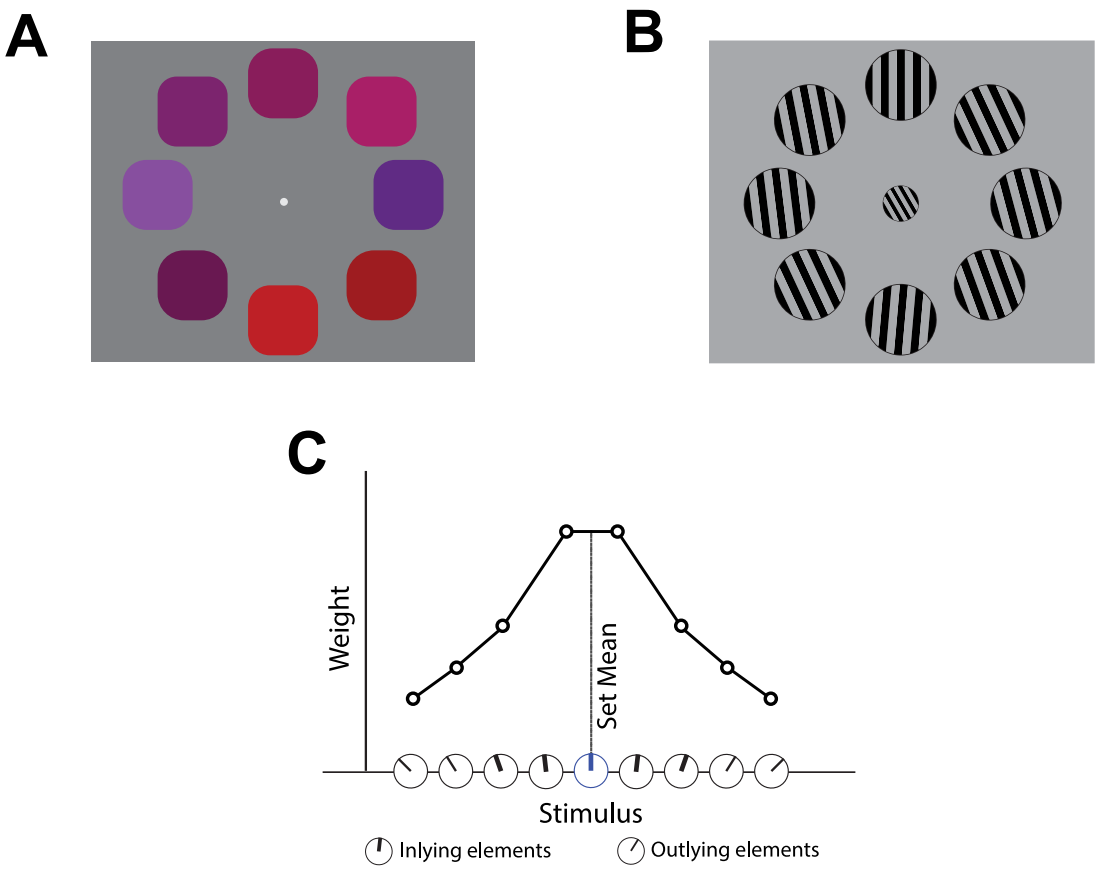

Fig. 1. Stimulus display used in (A) de Gardelle and Summerfield (2011) and (B) Li et al. (2017). Stimulus ensembles comprised a set of 8 items, presented equally spaced on a virtual circle around fixation. In de Gardelle and Summerfield (2011), each element varied along two feature dimensions (color and shape), and observers judged whether the average color (or shape) was more red/blue (or circular/square). In Li et al. (2017), observers were asked to determine whether the average orientation was more clockwise/anti-clockwise relative to the reference orientation shown in the center. In both studies, feature values of each element were randomly sampled from Gaussian distributions with varying means and variances. (C) Illustration of robust averaging in ensemble perception of orientation. Stimuli with feature values close to the set mean (inlying elements) contribute more to the estimated mean than those with feature values away from the set mean (outlying elements). The weights are computed using a logistic regression analysis (see *Methods*). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Different from applications in previous studies, however, we no longer assume that the prior $p(s)$ reflects the long-term, natural stimulus distribution. Rather, we consider that the prior is dominated by the short-term statistics of the relative ensemble stimulus distributions in the experiments. The encoding constraint Eq. (2) states that more frequently occurring feature values (i.e., values with higher prior probability) are represented with higher precision (i.e., higher Fisher information). As a result, the likelihood functions of the efficient Bayesian observer model are inhomogeneous for nonuniform stimulus statistics, contrasting the homogeneous likelihood functions often assumed in standard Bayesian model formulations (Fig. 2A).

We determine the likelihood function from an efficient sensory representation as previously proposed (Wei & Stocker, 2015). We first define a "sensory space" $\tilde{s}$, in which Fisher information is assumed to be uniform. Stimulus values in "stimulus space" s are then mapped to the sensory space via the cumulative of the prior distribution in the stimulus space $F(s)$, hence

$$\tilde{s} = F(s) = \int_{-\infty}^{s} p(s)ds \ . \tag{3}$$

We further assume that the measurement noise and thus the likelihood function is homogeneous and symmetric in sensory space. More specifically, we assume that the noisy measurement x follows a von Mises distribution (or a Gaussian for noncircular variables) with mean $\tilde{s}$ and a concentration parameter κ (or equivalently a variance parameter σ^2 for Gaussian noise), thus

$$p(x|\tilde{s}) = \frac{e^{\kappa \cos(x-\tilde{s})}}{2\pi I_0(\kappa)} \ , \tag{4}$$

where I_0 is the modified Bessel function of order zero. After defining the symmetric likelihood function in sensory space, we can map it back to the stimulus space via the inverse mapping function $F^{-1}(\tilde{s})$. The resulting likelihood function $p(x|s)$ in stimulus space typically shows an asymmetric shape, with long tails away from any peak of the prior distribution (see Fig. 2A).

With the encoding process defined, we model the ensemble decision task in the experiments as a hierarchical Bayesian inference process over the generative model shown in Fig. 2B. The generative model reflects the statistical characteristics of the decision experiments (de Gardelle & Summerfield, 2011; Li et al., 2017). On every trial, the ensemble category $C = \{0, 1\}$ is randomly chosen from two categories with equal probability $p(C) = \frac{1}{2}$. The chosen category determines the distribution from which the feature values in the ensemble are sampled from. Specifically, given each ensemble category, values are independently sampled from a Gaussian with mean μ_j and standard deviation σ_k, each randomly selected from a predefined set $\mu = \{\mu_1, \ldots, \mu_m\}$ and $\sigma = \{\sigma_1, \ldots, \sigma_n\}$. This defines the probability of each feature value in the ensemble for a given trial as $p(S|\mu_j, \sigma_k)$ with $S = (s_1, \ldots, s_8)$ representing the vector of the 8 independent feature values in the ensemble. Finally, the generative model considers that the observer only has access to noisy sensory representations of the ensemble stimuli $p(X|S)$, where we assume that the feature values of individual items are independently but efficiently encoded according to $p(x_i|s_i)$ as described above.

In the experiments, the observer's task is to infer the ensemble category C from noisy sensory measurements. We assume that the

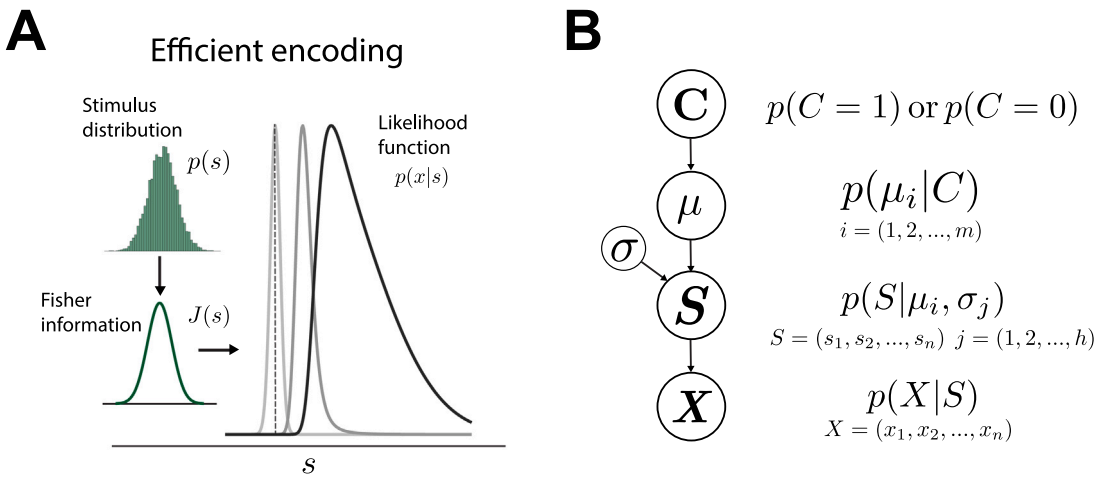

Fig. 2. (A) Efficient sensory encoding predicts inhomogeneous encoding precision (measured in Fisher information $J(s)$) for nonuniform stimulus distributions $p(s)$ (see Eq. (2)). This results in likelihood functions that are narrow and symmetric for measurements x close to the mean (dashed line) of the stimulus distribution and wider and more asymmetric as x is further away (Wei & Stocker, 2015). (B) Hierarchical, generative model of subjects' discrimination task. On each trial, the feature values S of the stimulus ensemble represent samples from a Gaussian distribution with varying mean μ and variance σ^2 depending on the category of the stimulus ensemble C (i.e., which side of the category boundary μ is). Samples were selected such that the feature average across the ensemble in every trial was essentially identical to the mean μ. An observer has to infer the correct category C based on noisy observations $X = (x_1, \dots, x_8)$ of each items feature value.

observer chooses the category with highest posterior probability $p(C|X)$ given the measurement array $X = (x_1, x_2, \dots, x_n)$. Using Bayes' rule, we can define a decision variable d as the posterior ratio

$$d = \frac{p(C=1|X)}{p(C=0|X)} = \frac{p(X|C=1)p(C=1)}{p(X|C=0)p(C=0)} \,. \tag{5}$$

Category likelihoods $p(X|C=1)$ and $p(X|C=0)$ are computed by summing over all possible combinations of μ_i and σ_j within each class. Likewise, the observer needs to marginalize over all possible feature values s given each measurement x_i. Therefore, the decision variable is given by

$$d = \frac{\sum_{k=1}^{h}\sum_{j=1}^{m}\prod_{i=1}^{8}\int p(x_i|s)p(s|\mu_j,\sigma_k)p(\mu_j,\sigma_k|C=1)ds\; p(C=1)}{\sum_{k=1}^{h}\sum_{j=1}^{m}\prod_{i=1}^{8}\int p(x_i|s)p(s|\mu_j,\sigma_k)p(\mu_j,\sigma_k|C=0)ds\; p(C=0)} \,. \tag{6}$$

Finally, the model assumes that the observer reports a decision $\hat{C} = 1$ when $d > 1$, and $\hat{C} = 0$ otherwise. To predict the probability of the observer's decision on each trial, we marginalize over all possible measurements x_i given each item's feature value s_i. Thus, the response probability given the stimulus ensemble S is

$$p(\hat{C}|S) = \int p(\hat{C}|X)p(X|S)dX \,, \tag{7}$$

where S represents the feature values of the stimulus ensemble as define above and X are all possible, independent measurements for each of the items in the ensemble. Because of the high dimensionality, we approximate marginalization by computing the empirical decision probability based on 2000 random samples of the measurement distributions Eq. (4) for every ensemble configuration in the experiment. Note that all model predictions shown in the paper are based on the exact same sets of ensemble stimuli as used in the experiments.

It is worth noting that the generative model (Fig. 2B) can capture more general experimental conditions than those tested in the experiment. For example, it can account for experiments with uneven category priors $p(C)$, as well as asymmetric distributions of the generative means $p(\mu|C)$. We have previously demonstrated that human decision behavior under such asymmetric task conditions can be well accounted for by Bayesian inference over such hierarchical generative models (Luu & Stocker, 2021). For the experiments considered in the current study, however, the higher-level structures are all symmetric and thus do not affect the decision behavior in any meaningful way. As a result, the model's behavior is predominantly determined by the encoding characteristics.

Standard Bayesian observer model We also consider a "standard" Bayesian observer model for comparison. This model uses exactly the same generative model (Fig. 2B). The only difference to the efficient Bayesian observer model is the encoding process $p(x_i|s_i)$: encoding is homogeneous, i.e., the sensory noise and thus the likelihood function is homogeneous in stimulus space.

2.4. Model fits

The efficient Bayesian observer model has two free parameters:

- κ — a noise parameter determining the width of the homogeneous noise distribution in sensory space (σ^2, if noise is assumed to be Gaussian)
- β — a weight parameter that determines the relative contributions of the stimulus distribution in the experiment and a uniform distribution to the stimulus prior $p(s)$. Specifically, we model the stimulus prior as a weighted sum of the stimulus distribution $p(s)$ relative to the decision boundary and a uniform distribution, with β and $1-\beta$ being their respective weights (see Fig. 3).

The standard Bayesian observer model has the same set of parameters (κ and β). For each of the datasets (de Gardelle & Summerfield, 2011; Li et al., 2017), we determined the parameter values that minimized the negative log-likelihood of the models using an adaptive search algorithm (Acerbi & Ma, 2017).

3. Results

Our efficient Bayesian observer model postulates that subjects' behavior in the ensemble average tasks represents the outcome of an optimal inference process. The model makes two specific predictions. First, it predicts a lower decision accuracy for ensembles with larger variance. With large ensemble variance, more items are, on average, encoded with lower precision because their associated feature values are further away from the peak of the stimulus distribution in the experiment. This results in lower discrimination accuracy for the entire ensemble. Second, it predicts robust averaging (i.e. nonuniform weighting profiles) as revealed by logistic regression (de Gardelle & Summerfield, 2011; Li et al., 2017). Because of efficient coding, inlying and outlying elements are represented with high and low precision, respectively. The Bayesian decision process then implicitly weighs inlying

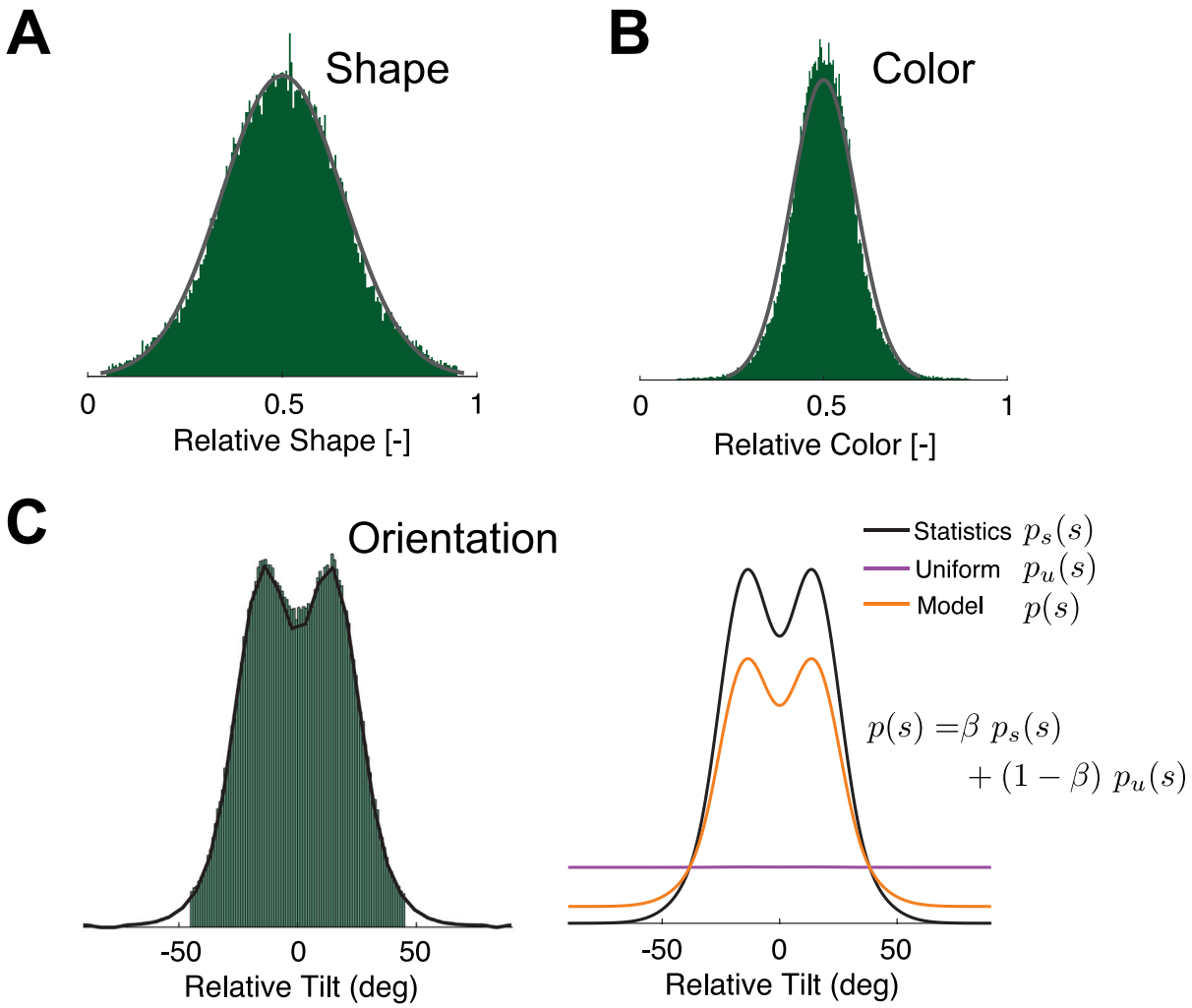

Fig. 3. Stimulus distributions relative to the decision boundary in the psychophysical experiments for shape (A), color (B), and orientation (C). For the model, we assumed the stimulus prior $p(s)$ to be a mixture between a smooth approximation of the relative stimulus distributions (black lines) and a uniform distribution according to a weight parameter β. It is a free model parameter (see Fig. 5B and Supplementary Fig. S2 for fit parameter values of individual subjects). Histogram are according to the psychophysical experiments (Experiments 3a and 3b in de Gardelle & Summerfield, 2011; Li et al., 2017).

elements heavier because they provide more reliable cues (i.e., narrower likelihood functions) for the inference process. These predictions are in stark contrast to those of a standard Bayesian model with homogeneous encoding, which predicts no effect of the ensemble variance on response accuracy and no robust averaging.

To validate our model and test these predictions, we fit the efficient Bayesian observer model (and, for comparison, the standard model) to behavioral data from a series of psychophysical experiments (de Gardelle & Summerfield, 2011; Li et al., 2017). Importantly, we constrained the stimulus prior $p(s)$ to reflect the experimental stimulus distribution of the relevant features in each experiment relative to the decision boundary. Fig. 3 shows these distributions across all ensembles for the three features used in the experiments. In order to allow for some flexibility in how much subjects adapted to the experimental stimulus distributions, we assumed $p(s)$ to reflect an average between these distributions (smoothly approximated) and a uniform distribution weighted by a free model parameter β (see Fig. 3C).

3.1. Dependence between ensemble variance and decision accuracy

Both de Gardelle and Summerfield (2011) and Li et al. (2017) reported that subjects' decision accuracy depends on the ensemble variance for a fixed ensemble mean, with higher variance leading to lower accuracy. As shown in Figs. 4 and 5, the efficient Bayesian observer model well captures this effect for all three experiments using different stimulus features (shape, color, and orientation). Likewise, the model well accounts for subjects' decision accuracy as a function of task difficulty, i.e., depending on how far the ensemble mean μ is away from the decision boundary. The quantitative accurate model account is strong evidence in support of the model as both effects are independent. In comparison, the standard Bayesian observer model does not predict any dependence between ensemble variance and decision accuracy (see Supplementary Fig. S1).

3.2. Robust averaging

To recover the regression weights predicted by the models, we first generated the response probability of each trial using the best-fitting parameter values for each model. We then performed logistic regression with either stimulus ranks (as in de Gardelle & Summerfield, 2011) or bins (as in Li et al., 2017) as the predictors and response probabilities as the dependent variable. Consistent with human data, the efficient Bayesian observer model behavior exhibits nonuniform regression weights for both shape and color conditions (Fig. 4). Specifically, the weights for inlying elements (i.e., ranks of 3–6) are higher than those for outlying elements (i.e., ranks of 1,2, and 7, 8). In addition, the predicted weight difference between inlying and outlying

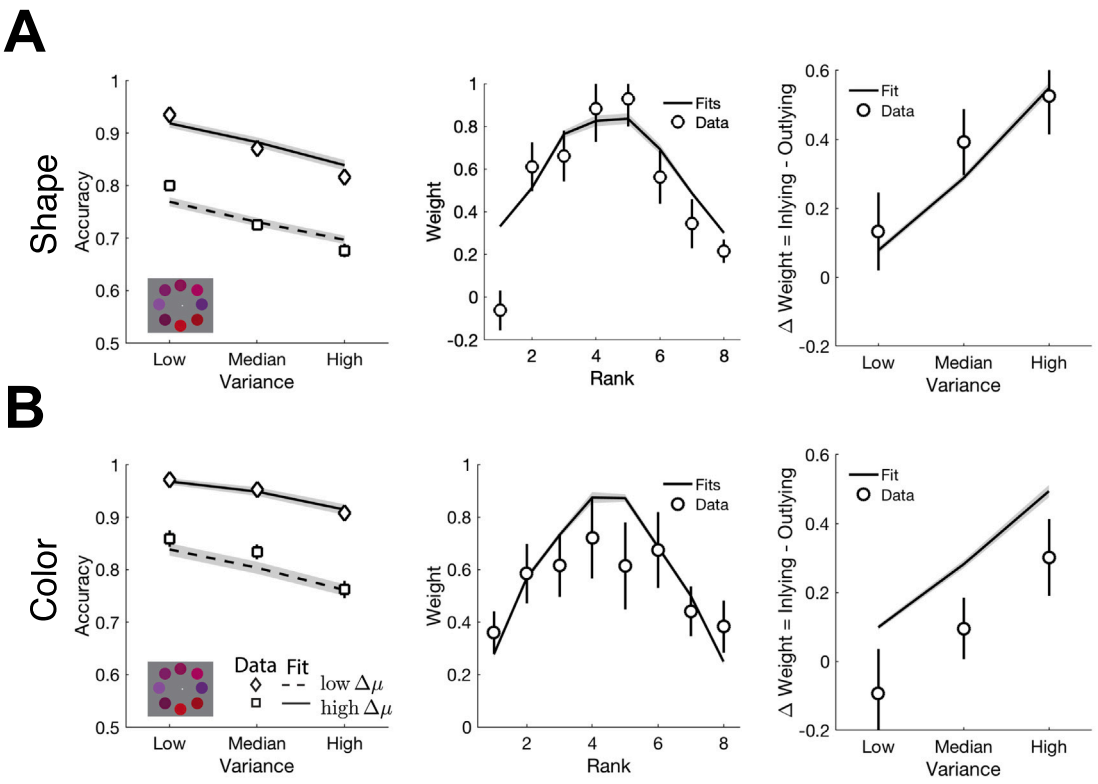

Fig. 4. Data and fit model predictions for the ensemble average decision task (color and shape — Data reanalyzed from Experiments 3a and 3b in de Gardelle and Summerfield (2011)). Decision accuracy (left), recovered regression weights for item rank (middle), and regression weight differences between inlying and outlying items (right) for shape (A) and color (B) ensembles. Open marks and solid lines indicate the average response accuracy and regression weights, and the model predictions, respectively, both based on analyses/fits to individual subjects data. Error bars and shaded areas represent the standard error of the subject population mean (SEM). Fit model parameter values for individual subjects are shown in Supplementary Fig. S2.

elements increases with the ensemble variance, which is also in line with the data.

The same is true for the orientation experiment (Fig. 5B). The model's predicted regression weights exhibit the same signature of robust averaging as the human data: higher weights for inlying elements (i.e., bins close to the ensemble mean) and lower weights for outlying elements (i.e., bins away from the ensemble mean). Moreover, the predicted weights were overall higher in the fixed compared to the variable condition. Examining the fit model parameters reveals that the model attributes the difference in decision accuracy and regression weights between fixed and variable conditions primarily to an increase in sensory noise (Fig. 5C). Across all subjects, the fit noise parameter σ is consistently higher for the variable condition. This makes sense given that in the variable condition, the orientation of the reference stimulus was randomly changed in each trial, and thus subjects had to establish the decision boundary on every trial anew via a perceptual process, whereas in the fixed condition they could form a stable representation of that boundary over an entire block of trials (see Fig. 5A). For simplicity, we kept the generative model simple and did not include potential noise in the percept of the decision boundary (Fig. 2B). Thus the model simply accounts for this extra noise via an increase in sensory noise of the ensemble stimulus. In contrast, yet not surprisingly, the fit standard Bayesian observer model predicts uniform regression weights (see Supplementary Fig. S1).

3.3. Effect of the stimulus prior on decision accuracy and regression weights

A defining characteristic of our model is that the precision with which individual elements in the ensemble stimuli are encoded is determined by the stimulus prior $p(s)$ according to the efficient coding assumption Eq. (2). The prior parameter β thereby determines how strongly an observer adapted the stimulus prior $p(s)$, and thus the sensory encoding characteristics, to the ensemble statistics in the experiment. More specifically, it specifies the relative weight of the empirical stimulus distribution in the experiment in a linear mixture with a uniform distribution in determining $p(s)$. Fit β values vary substantially for individual subjects throughout the different experiments (Fig. 5C and Supplementary Fig. S2). We ran model simulations to illustrate how different β values correspond to individual differences in decision behavior. Fig. 6 shows the predicted decision accuracies and the retrieved regression weights for the color discrimination experiment (Fig. 4B) for three different β values.

For $\beta = 1$ the model is fully and thus most efficiently adapted to the stimulus distribution in the experiment, showing highest overall decision accuracy (Fig. 6A) as well as strongest robust averaging (Fig. 6B). Decision accuracy depends on the ensemble variance because with larger variance more items in the ensemble are at the tail end of the prior distribution and thus encoded with lower precision. Overall decision accuracy and the differential weighting between inlying and outlying elements gradually decrease with decreasing values of β. For $\beta = 0$, the stimulus prior $p(s)$ is uniform and thus encoding precision is homogeneous and thus least adapted to the experimental stimulus statistics. As a result, overall decision accuracy is lowest and not depending on ensemble variance, and the regression weights are approximately uniform. Note that for the experiments considered in our study, the model behavior for $\beta = 0$ is functionally identical to the behavior predicted by the standard Bayesian observer model (see Methods).

3.4. Results summary

The proposed efficient Bayesian observer model provides an accurate account of the data obtained from various different ensemble average discrimination experiments. The model assumes that subjects'

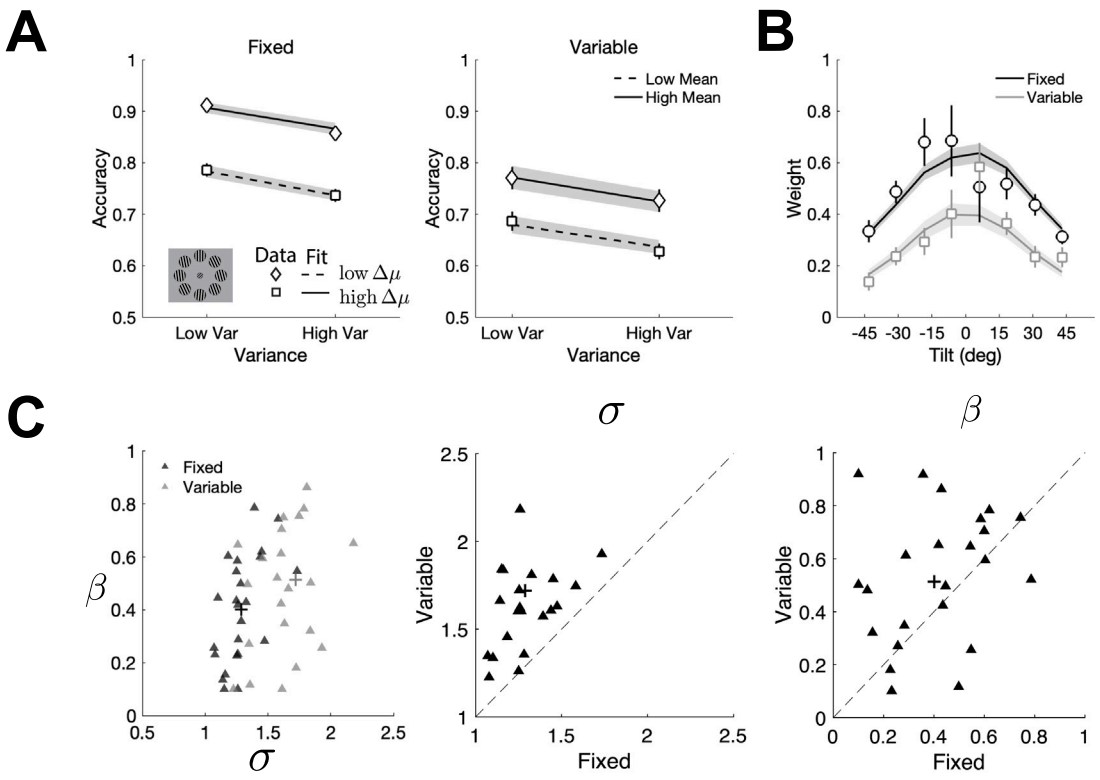

Fig. 5. Data, fit model predictions, and fit parameter values for ensemble average decision task (orientation — Data reanalyzed from Li et al. (2017)). (A) Decision accuracy in the fixed (left) and variable (right) reference conditions. (B) Recovered regression weights for each of the 8 equally-spaced bins in both conditions (right). Error bars and shaded areas represent SEMs of the data and model fits, respectively. (C) Fit σ (internal noise) and β (relative weight of the experimental stimulus distribution (Fig. 3C) to the mixture prior) values in the fixed (black triangle) and variable (gray triangle) conditions. σ values were converted from the best fitting concentration parameter κ. Triangles represent fit parameter values for individual subjects. The cross represents average fit parameter value across all subjects. Model regression weights for individual subjects are shown in Supplementary Fig. S4.

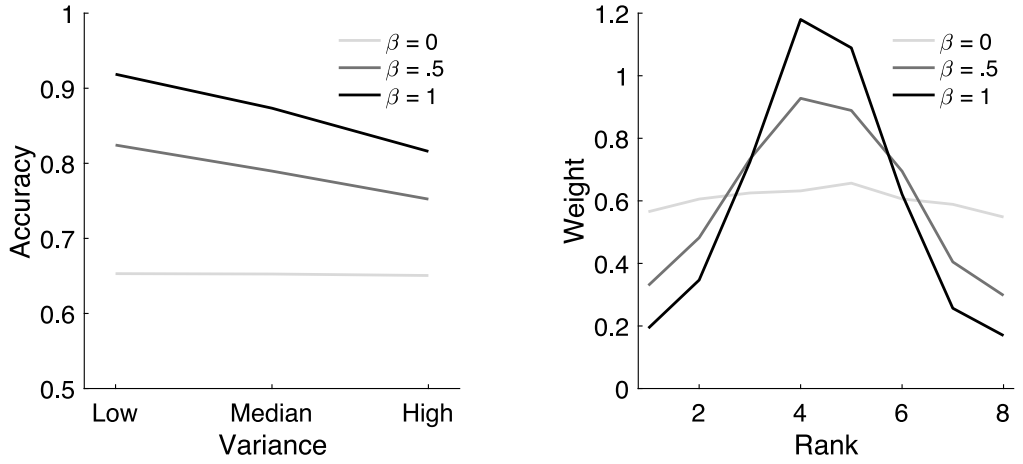

Fig. 6. The impact of the prior parameter β on the expected decision accuracy and regression weights. Shown are the predicted decision accuracies (A) and regression weights (B) of the model for the color discrimination experiment (Fig. 4A) for $\beta = [0, 0.5, 1]$. Note that the sensory noise level σ was fixed, and thus the overall amount of available sensory encoding resources $E = \int \sqrt{J(s)}$ is identical for each condition (Wei & Stocker, 2016). The increase in overall decision accuracy for increasing β is thus a direct reflection of the performance gain obtained with efficient coding.

encoding precision and prior beliefs follow the stimulus distribution over the course of the experiment. This assumption is further supported by the fact that the model also provides a good account of behavior in experiments where the stimulus distribution was shifted relative to the decision boundary (Experiments 4a and 4b in de Gardelle and Summerfield (2011) — see Supplementary Fig. S3).

4. Discussion

In this paper, we propose that "robust averaging" in ensemble perception naturally emerges from an efficient allocation of limited sensory bandwidth in an otherwise optimal perceptual integration process (Wei & Stocker, 2015). We validated this hypothesis by deriving a hierarchical Bayesian observer model constrained by efficient coding.

We fit this model to data from a series of ensemble discrimination experiments, demonstrating that it can quantitatively well account for all the key characteristics of subjects' behavior in this task (de Gardelle & Summerfield, 2011; Li et al., 2017).

A key assumption of our model is that sensory encoding can quickly adapt to the ensemble statistics of the experiment. This extends previous formulations, which assumed that both encoding and decoding are adapted to rather long-term stimulus statistics (Wei & Stocker, 2012, 2015). Results from a few previous studies on sensory adaptation and perceptual learning (Noel, Zhang, Stocker, & Angelaki, 2021; Wei & Stocker, 2017) and also higher-level inference tasks (Polania, Woodford, & Ruff, 2019), already suggested that resource-rational, efficient Bayesian observer models may also accurately describe human perceptual behaviors operating on short-term stimulus statistics. Our results provide further evidence for that. Other studies have shown that human subjects can quickly learn the shape of uniform, Gaussian, and even bimodal distributions over a small set of trials (Chetverikov, Campana, & Kristjansson, 2017; Hansmann-Roth, Chetverikov, & Kristjánsson, 2019), and then use this information as their prior beliefs in a perceptual inference task (Chalk, Seitz, & Seriès, 2010; Jazayeri & Shadlen, 2010; Roach, McGraw, Whitaker, et al., 2017). In light of our results, we suggest to revisit those studies and investigate whether Bayesian inference models that also consider efficient encoding changes according to these priors would provide an even better account of the data.

Our results naturally raise the question about the underlying neural mechanism that could give rise to such quick changes in sensory representation. Attention seems the obvious candidate; studies have shown that ensemble encoding accuracy depends on whether attention is being equally distributed over all items in the ensemble display or focused only on a small subset (Baijal, Nakatani, van Leeuwen, & Srinivasan, 2013; Chong, Joo, Emmmanouil, & Treisman, 2008; Chong & Treisman, 2005). The benefit of distributed attention, however, comes with a cost in that it leads to compressed representations of each individual item compared to their independent representation (Alvarez & Oliva, 2009; Baijal et al., 2013). These findings fit the notion that attention is a limited encoding resource, and that its graded distribution to individual items in the ensemble performs the function of efficient coding as we propose. Results showing that items that receive high attention contribute stronger to the estimated ensemble average also fit this idea (De Fockert & Marchant, 2008; Im et al., 2015). For at least one of the experiments we considered in our study (the 'variable condition' experiment (Li et al., 2017)) such an attentional mechanism would need to operate on the sub-second timescale of a single trial because the decision boundary, and therefore the corresponding efficient stimulus representation, varied with every trial. The role the reference stimulus plays in establishing these fast changes in sensory representation remains an interesting open question, in particular with regard to estimation tasks that do not include a reference (e.g., Rosenbaum et al., 2021). Alternatively, a bottom-up process driven by the entire stimulus ensemble may fully guide the attentional change in encoding accuracy. Examining in detail the relation between efficient coding theories and attention as a short-term allocation mechanism of sensory resources/bandwidth presents an interesting future research direction.

Our model provides an alternative to a previously proposed, heuristic account of robust averaging. In Li et al. (2017), the authors showed that a compressive nonlinear transformation (e.g., an exponential function with an exponent smaller than 1) of individual feature values can mimic the nonuniform contribution of individual stimulus items to subjects' estimates of the ensemble average. They showed that downweighting outlying elements can improve decision accuracy in case of late decision noise, which, they proposed, is the reason why robust averaging occurs in ensemble perception. Our efficient Bayesian observer model (i.e., its efficient coding assumption) provides a normative and quantitative definition of the proposed nonlinear transformation, and predicts that the nonlinearity will depend on the ensemble statistics of the specific experiment. Thus, future experiments with different ensemble statistics will allow us to distinguish whether the robustness to late decision noise is the main motivation for robust averaging, or rather a welcome side-effect of the proposed efficient coding assumption.

Another alternative model suggests that robust averaging is the result of unequal weighting at the decision level (Teng, Li, & Zhang, 2021). Specifically, the model proposes that a cost function that is robust to large errors (e.g., an "inverse" Gaussian) reduces the contribution of outlying items to the integration process. Aside from the conceptual difficulty to incorporate this robust cost function approach into a complete model of the decision tasks we consider here, one can indeed show that such a loss function assumption can qualitatively explain many of the robust averaging results presented here. However, the assumption imposes robust averaging by design, i.e., the inverse Gaussian shape of the loss function determines the expected Gaussian weighting profile from the regression analysis, independent of the stimulus distribution. In contrast, the efficient coding assumption of our model does not lead to regression weights that follow a predefined profile but rather depend on the shape of the stimulus distribution. Thus ensemble discrimination experiments with relative stimulus distributions that substantially deviate from a Gaussian will allow us to discriminate the two models.

Finally while our study focused on robust averaging in ensemble perception of low-level features (i.e., color, shape, orientation), the characteristic downweighting of outliers has also been reported in ensemble perception of high-level visual stimuli such as faces (Haberman & Whitney, 2010). A recent study indicates that the weighting profile obtained from subjects' estimates of the average value of a sequence of numbers is modulated by the probability of the different numbers (Prat-Carrabin & Woodford, 2022). Specifically, numbers associated with higher probabilities contribute stronger to the estimate regardless of whether they are small or large, even though the demonstrated quantitative match between the strength of contribution and these prior probabilities is less accurate than what we show here. However, the general finding well aligns with the predictions of our proposed efficient Bayesian observer model, although these predictions may not fully hold when averaging small integers (Spitzer, Waschke, & Summerfield, 2017). It indicates that our proposed model may generalize beyond ensemble perception of low-level visual features to more cognitive representations.

4.1. Conclusion

We present a normative explanation for the well-known "robust averaging" phenomenon in ensemble perception of low-level visual features. We demonstrate that a Bayesian observer model that efficiently allocates its limited sensory resources according to the ensemble statistics can accurately account for several, previously reported datasets. Our results further support the notion that "resource-rational" Bayesian observer models can account for seemingly irrational behavior.

Data and code availability

All data in this paper have been previously published and were obtained directly from the authors. Matlab code for simulating and fitting the computational models is available at the public repository Github (https://github.com/cpc-lab-stocker/ensemble-perception-2022).

Acknowledgments

We thank Vincent De Gardelle for sharing his behavioral data. We also thank the members of the Computational Perception and Cognition Laboratory for many helpful discussions of the work, and the reviewers for valuable and constructive feedback. This work was supported by the University of Pennsylvania and in part by grant IIS-1912232 from the National Science Foundation (AAS).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105334. It contains model parameters for Exp. 3 and model fits for Exps. 1, 2, 4a and 4b of de Gardelle and Summerfield (2011), as well as predicted regression weights for individual subjects in Li et al. (2017). It also shows fits of the standard Bayesian model to all data.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems*, *30*, 1837–1847.

Alvarez, G. A., & Oliva, A. (2009). Spatial ensemble statistics are efficient codes that can be represented with reduced attention. *Proceedings of the National Academy of Sciences*, *106*(18), 7345–7350.

Anderson, N. (1968). Averaging of space and number stimuli with simultaneous presentation. *Journal of Experimental Psychology*, *77*(3), 383–392.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Baijal, S., Nakatani, C., van Leeuwen, C., & Srinivasan, N. (2013). Processing statistics: An examination of focused and distributed attention using event related potentials. *Vision Research*, *85*, 20–25.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Chalk, M., Seitz, A. R., & Seriès, P. (2010). Rapidly learned stimulus expectations alter perception of motion. *Journal of vision*, *10*(8), 1–18.

Chetverikov, A., Campana, G., & Kristjansson, A. (2017). Rapid learning of visual ensembles. *Journal of vision*, *17*(2), 21.

Chong, S. C., Joo, S. J., Emmmanouil, T.-A., & Treisman, A. (2008). Statistical processing: Not so implausible after all. *Perception & Psychophysics*, *70*(7), 1327–1334.

Chong, S. C., & Treisman, A. (2005). Statistical processing: Computing the average size in perceptual groups. *Vision Research*, *45*(7), 891–900.

De Fockert, J. W., & Marchant, A. P. (2008). Attention modulates set representation by statistical properties. *Perception & Psychophysics*, *70*(5), 789–794.

Epstein, M. L., Quilty-Dunn, J., Mandelbaum, E., & Emmanouil, T. A. (2020). The outlier paradox: The role of iterative ensemble coding in discounting outliers. *Journal of Experimental Psychology: Human Perception and Performance*, *46*(11), 1267–1279.

de Gardelle, V., & Summerfield, C. (2011). Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, *108*(32), 13341–13346.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825–1838.

Hansmann-Roth, S., Chetverikov, A., & Kristjánsson, Á. (2019). Representing color and orientation ensembles: Can observers learn multiple feature distributions? *Journal of vision*, *19*(9), 2.

Hubert-Wallander, B., & Boynton, G. M. (2015). Not all summary statistics are made equal: Evidence from extracting summaries across time. *Journal of vision*, *15*(4), 5.

Im, H. Y., Park, W. J., & Chong, S. C. (2015). Ensemble statistics as units of selection. *Journal of Cognitive Psychology*, *27*(1), 114–127.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*(8), 1020–1026.

Juni, M. Z., Gureckis, T. M., & Maloney, L. T. (2012). Effective integration of serially presented stochastic cues. *Journal of vision*, *12*(8), 1–16.

Li, V., Castañón, S. H., Solomon, J. A., Vandormael, H., & Summerfield, C. (2017). Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology*, *13*(8), Article e1005723.

Luu, L., & Stocker, A. A. (2021). Categorical judgments do not modify sensory representations in working memory. *PLoS Computational Biology*, *17*(6), 1–28.

Noel, J.-P., Zhang, L.-Q., Stocker, A. A., & Angelaki, D. E. (2021). Individuals with autism spectrum disorder have altered visual encoding capacity. *PLoS Biology*, *19*(5), Article e3001215.

Polania, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, *22*(1), 134–142.

Prat-Carrabin, A., & Woodford, M. (2022). Efficient coding of numbers explains decision bias and noise. *Nature Human Behaviour*, *6*, 1142–1152.

Roach, N., McGraw, P., Whitaker, D., et al. (2017). Generalization of prior information for rapid Bayesian time estimation. *Proceedings of the National Academy of Sciences*, *114*(2), 412–417.

Rosenbaum, D., de Gardelle, V., & Usher, M. (2021). Ensemble perception: Extracting the average of perceptual versus numerical stimuli. *Attention, Perception, and Psychophysics*, *83*, 956–969.

Spencer, J. (1963). A further study of estimating averages. *Ergonomics*, *6*(3), 255–265.

Spitzer, B., Waschke, L., & Summerfield, C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, *1*(8), 1–8.

Teng, T., Li, S., & Zhang, H. (2021). The virtual loss function in the summary perception of motion and its limited adjustability. *Journal of vision*, *21*(5), 2.

Tong, K., Dubé, C., & Sekuler, R. (2019). What makes a prototype a prototype? Averaging visual features in a sequence. *Attention, Perception, & Psychophysics*, 1–17.

Vandormael, H., Castañón, S. H., Balaguer, J., Li, V., & Summerfield, C. (2017). Robust sampling of decision information during perceptual choice. *Proceedings of the National Academy of Sciences*, *114*(10), 2771–2776.

Wei, X.-X., & Stocker, A. A. (2012). Efficient coding provides a direct link between prior and likelihood in perceptual Bayesian inference. In *Advances in neural information processing systems 25* (pp. 1304–1312).

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, *18*(10), 1509.

Wei, X.-X., & Stocker, A. A. (2016). Mutual information, Fisher information, and efficient coding. *Neural Computation*, *28*(2), 305–326.

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, *114*(38), 10244–10249.

Whitney, D., & Yamanashi-Leib, A. (2018). Ensemble perception. *Annu. Rev. Psychol.*, *69*, 105–129.

Zhang, L.-Q., & Stocker, A. A. (2022). Prior expectations in visual speed perception predict encoding characteristics of neurons in area MT. *Journal of Neuroscience*, *42*(14), 2951–2962.