Efficient coding of numbers explains decision bias and noise

Arthur Prat-Carrabin[1,⋆] and Michael Woodford[1]

[1]Department of Economics, Columbia University, New York, USA

⋆*email: arthur.p@columbia.edu*

Abstract

Human subjects differentially weight different stimuli in averaging tasks. This has been interpreted as reflecting biased stimulus encoding, but an alternative hypothesis is that stimuli are encoded with noise, then optimally decoded. Moreover, with efficient coding, the amount of noise should vary across stimulus space, and depend on the statistics of stimuli. We investigate these predictions through a task in which participants are asked to compare the averages of two series of numbers, each sampled from a prior distribution that differs across blocks of trials. We show that subjects encode numbers with both a bias and a noise that depend on the number. Infrequently occurring numbers are encoded with more noise. A maximum-likelihood decoding model captures subjects' behaviour and indicates efficient coding. Finally, our model predicts a relation between the bias and variability of estimates, thus providing a statistically-founded, parsimonious derivation of Wei and Stocker's "law of human perception".

In many decision problems, someone is presented with an array of variables that must be aggregated in order to identify the optimal decision. How humans combine several sources of information in their decision-making process is a long-standing debate in economics and cognitive science [1, 2, 3, 4, 5]. Recently, a series of experimental studies have focused on averaging tasks, in which subjects are presented with several stimuli (sometimes numbers, but sometimes visual stimuli characterized by their length, orientation, shape, or color) and asked to make a decision about the *average* magnitude of the presented stimuli [6, 7, 8, 9]. Although the contribution of each stimulus to the average should, in theory, be proportional to its true magnitude, the weights attributed to stimuli by human subjects, in their decisions, appear to be nonlinear functions of their magnitudes. Subjects asked to compare the averages of two series of digits, for instance, overweight larger digits when making a decision [6]. What is the origin of this seemingly suboptimal behaviour? Refs. [6, 7] show that if comparison of the average encoded values involves *noise*, then a *nonlinear transformation* of the presented stimuli can partially compensate for the performance loss induced by the noise. The nonlinear distortion of stimuli appears, under this proposal, as a consequence of an optimal encoding strategy, given unavoidable noise at decision time.

Here, we investigate the alternative hypothesis that a presented stimulus is encoded with noise, while the decoding rule used to produce a comparative judgment is deterministic, and indeed represents an optimal inference from the noisy encoded values (rather than a comparison of the simple sums of the encoded values, plus noise). Under the assumption that the estimate of each magnitude is optimally inferred from its encoded representation [10, 11], the noisy estimate that results will generally be biased, to a degree that will vary as a function of the stimulus [12, 13, 14, 15]. Thus the average estimate will be a nonlinear function of the true stimulus magnitude, though for a different reason than in the model of Ref. [6] mentioned above. This nonlinear function will depend on how the encoding noise varies over the stimulus space. Efficient coding theories suggest that the degree of noise with which a stimulus is encoded should be a decreasing function of its probability under the prior distribution of stimuli; in other words, less likely stimuli should be encoded with more noise [16, 17, 18, 19, 20]. This implies that the way that both estimation bias and estimation noise vary over the stimulus space should depend on the prior distribution over that space; we test for such dependence in our data.

In a related model that combines efficient coding with Bayesian decoding, Wei and Stocker derive a relation between estimation noise and estimation bias, a “law of human perception” supported by evidence from numerous sensory domains [21]. We find support for this law in our data on number comparisons as well, and show that it is also predicted by our encoding-decoding model, though for reasons somewhat different than the derivation provided by Wei and Stocker. We also test other implications of both efficient coding and optimal decoding.

Each of these candidate theories highlights the potential role of the prior in human decision-making. Thus we design a task in which subjects are asked to compare the averages of two series of numbers, and manipulate the prior distribution of the presented numbers across different blocks of trials: it is sometimes uniform, but sometimes skewed toward smaller or larger numbers. In each of these three conditions, the weights of numbers in the decisions of subjects appear to vary nonlinearly over the range of presented numbers. We compare the behaviour of our subjects to the predictions of a family of noisy estimation models, and find that the patterns in their responses are best captured by a model in which a nonlinear transformation of a presented number is observed with a degree of noise that itself depends on the number.

Turning to encoding-decoding models, we show how a maximum-likelihood model of inference of the number from noisy representations accounts both for the nonlinear transformation and for the noise, and best captures subjects’ behaviour. The encoding noise, furthermore, varies depending on the prior. Finally, we derive from our model a relation between the nonlinear transformation and the noise. This relation is also predicted by Wei and Stocker's law of human perception, but our derivation relies on different assumptions, and in particular, under our derivation the "law" does not depend on the efficient-coding hypothesis. This suggests that the encoding need not be optimally adapted to the prior for the law to be valid. Nonetheless, we do find evidence that encoding varies with the prior in a way consistent with a theory of efficient coding.

Results

We first present our average-comparison experiment. In each trial of a computer-based task, ten numbers, each within the range $[10.00, 99.99]$, and alternating in color between red and green, are presented to the subject in rapid succession. The subject is then asked to choose whether the five red numbers or the five green numbers had the higher average (Fig. 1a). In different blocks of trials, numbers are sampled from different prior distributions: the **"Uniform"** distribution, under which all numbers in the $[10.00, 99.99]$ range are equally likely, an **"Upward"** distribution (under which higher numbers are more likely), and a **"Downward"** distribution (under which lower numbers are more likely) (Fig. 1b). Within a block of trials, both red and green numbers are sampled from the same distribution. Additional details on the task are provided in Methods.

Although the presented information would allow, in principle, an unambiguous identification of the color of the numbers that have the higher average, subjects make errors in their responses. These errors are not independent of the presented numbers: the proportion of trials in which they choose 'red' (the "choice probability") is an increasing, sigmoid-like function of the difference between the average of the red numbers and that of the green numbers, with about 50% red responses when this difference is close to zero. In other words, subjects make less errors when there is a marked difference between the two averages (Fig. 1c).

In order to isolate the influence of a single number on the decision, we can compute the probability of choosing 'red' conditional on a given (red) number x being presented, $P(\text{'red'}|x)$, and the absolute difference between this probability and 0.5: $\big|P(\text{'red'}|x) - 0.5\big|$. This quantity, which we call the "decision weight" (following Ref. [6]), is a measure of the average impact of a given number on the decision. For an ideal subject who makes no errors, the decision weight should be an increasing, approximately linear function of the absolute difference between the number and the mean of the prior: numbers farther from the prior mean should have a greater impact on the decision, while a number exactly equal to the prior mean has a zero decision weight (Fig. 1d,e, thin lines). In our subjects' data, instead, we find that this relation appears to be nonlinear, with different steepnesses for numbers above and below the prior mean. More precisely, in the three conditions (characterized by the different priors), a number close to the prior mean and below it has a larger decision weight than a number equally far from the mean, but above it. Instead, for numbers farther

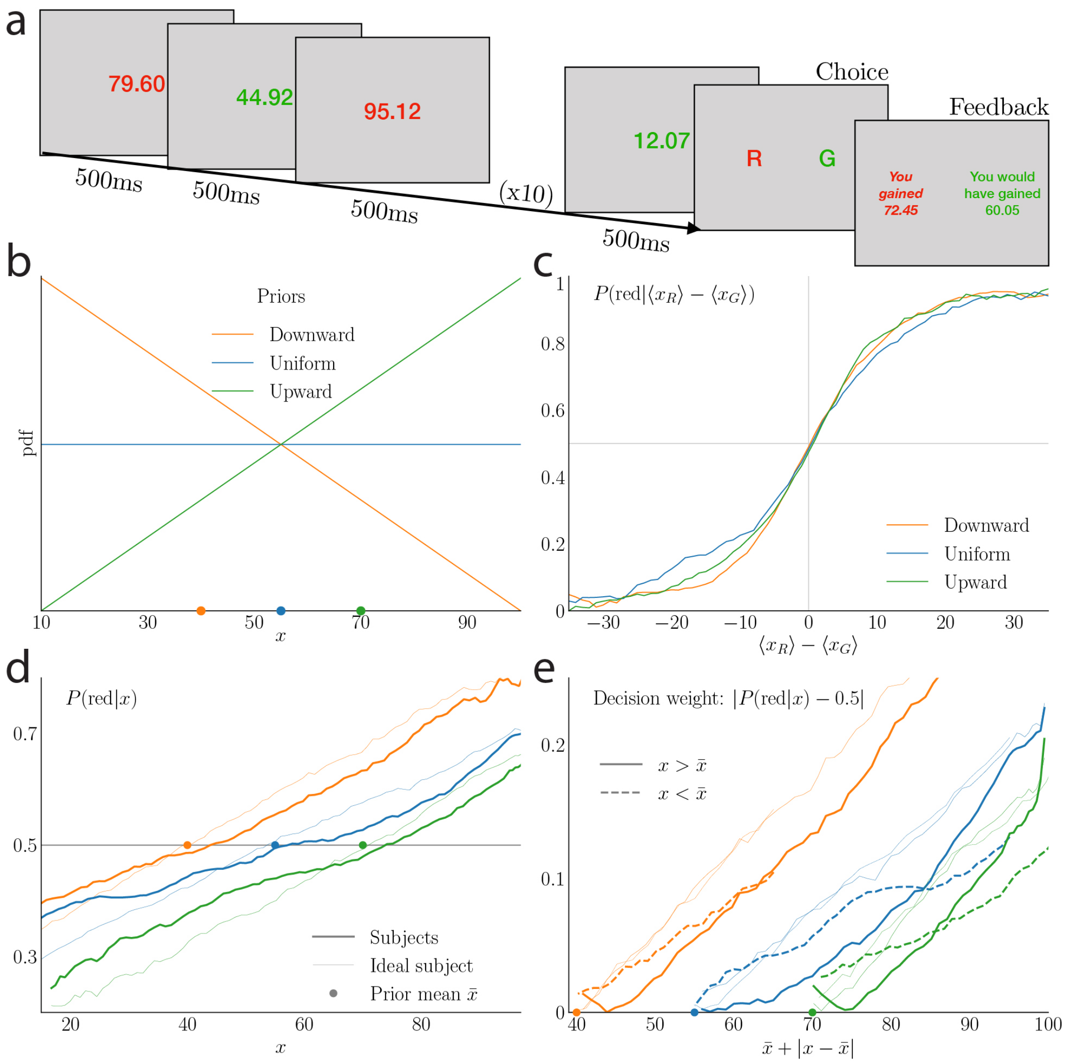

Figure 1: Illustration of the task and subjects choice behaviour. a. In each trial, ten numbers are sequentially presented, alternating in color between red and green. The subject can then choose a color, without time constraint, by pressing the corresponding key. Feedback is given, which reports the two averages and emphasizes the chosen one. **b.** In each trial, both red and green numbers are drawn from the same prior distribution over the range $[10.00, 99.99]$. In different blocks of consecutive trials, the prior is either Downward (triangular distribution with the peak at 10.00, orange line), Uniform (blue line), or Upward (triangular distribution with the peak at 99.99, green line). **c.** Choice probability: in each of the three prior conditions, the fraction of trials in which subjects choose the 'red' average, as a function of the true difference in the averages of the two series of numbers. While an ideal subject would be described by a step function, the choices of subjects result in sigmoid curves. **d.** Probability $P(\text{red}|x)$ of choosing 'red' conditional on a red number x being presented, as a function of x, for the subjects (thick lines) and the ideal subject (thin lines). **e.** Decision weight, defined as $|P(\text{red}|x) - 0.5|$, for the subjects (thick lines) and the ideal subject (thin lines), as a function of the sum of the prior mean, $\bar{x}$, and the absolute difference between the number and the prior mean, $|x - \bar{x}|$. Subjects' decision weights for numbers below the mean (dashed line) and above the mean (solid line) are appreciably different, whereas for the ideal subject there is only a small difference due to sampling error. In **c**, **d**, and **e**, each point of the curves is obtained by taking the average of the quantity of interest (ordinate) over a sliding window of length 10 of the quantity on the abscissa, incremented by steps of length 1.

from the mean, the opposite occurs: a number below the mean has a lower weight than an equally-distant number above the mean (Fig. 1d,e, thick lines).

We wish to explore models of noisy comparison that can account for these regularities. We begin by fitting models to our data that do not separate the processes of encoding and decoding, and instead simply posit that a given number x results in a noisy estimate $\hat{x}$ drawn from a conditional distribution $p(\hat{x}|x)$. We assume that the ten numbers presented are estimated with this procedure (with independent draws), and that the color 'red' is chosen if the average of the estimates of the red numbers is greater than that of the green numbers, i.e., if $\frac{1}{5}\sum_{i=1}^{5}\hat{x}_i^R > \frac{1}{5}\sum_{i=1}^{5}\hat{x}_i^G$, where $\hat{x}_i^R$ and $\hat{x}_i^G$ are the estimates for the red and green numbers. Note that this kind of characterization of the data is equally consistent with the theory of Ref. [6], in which $\hat{x}$ is equal to a deterministic transformation of x plus a random noise term added at the time of the comparison process, and a model in which x is encoded with noise and $\hat{x}$ is then an estimate of the value of x based on the noisy encoded value.

We consider several possible assumptions about the form of the conditional distribution of estimates, $p(\hat{x}|x)$. The simplest kind of model consistent with a sigmoid curve of the kind shown in Figure 1c would be one in which the estimate is normally distributed around the true value of the number, with a constant standard deviation, s, i.e.: $\hat{x}|x \sim N(x, s^2)$. In other words, numbers are perceived with constant Gaussian noise, but the estimates are unbiased. The probability of choosing the red color, conditioned on the ten presented numbers, would then be

$$P\left(\sum_{i=1}^{5}\hat{x}_i^R > \sum_{i=1}^{5}\hat{x}_i^G \middle| x_{1:5}^R, x_{1:5}^G\right) = \Phi\left(\frac{\sum_{i=1}^{5} x_i^R - \sum_{i=1}^{5} x_i^G}{\sqrt{10s^2}}\right), \tag{1}$$

where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The choice probability in this model is thus a sigmoid function of the difference between the averages of red and green numbers, similarly to the choice probability of subjects (Fig. 2a). However, the decision weight of a number, in this model, is an approximately linear function of the absolute difference between the number and the prior mean, and two numbers equally distant from the mean have the same weight (Fig. 2c, top left panel). Thus this simple model does not reproduce the subjects' unequal weighting of numbers in decisions documented in Figure 1e.

In the model just presented, estimates are unbiased ($\mathbb{E}\hat{x} = x$), and all numbers are estimated with the same amount of noise ($\text{Var}\hat{x} = s^2$). But we also consider models that are more flexible in either or both of these respects. For example, we can allow for bias by assuming that the estimate, $\hat{x}$, of a presented number, x, is normally distributed around a nonlinear transformation, $m(x)$, of the number: $\hat{x}|x \sim N(m(x), s^2)$. Such a model is equivalent to the one assumed in Refs. [6, 7], in which noise is added to the sum of nonlinearly transformed values of the stimuli. Alternatively, we might assume that the variability of estimates is not constant over the stimulus space, as is found to be true in many stimulus domains [14]. The simplest case of this kind would be one in which $\hat{x}|x \sim N(x, s^2(x))$, where now the estimation noise, $s(x)$, varies with the number. Our most general model combines the features of these last two, positing that a transformation of the number is observed with varying noise: $\hat{x}|x \sim N(m(x), s^2(x))$. In this model, the probability of choosing the red color, conditioned on the ten presented numbers, will be

$$P\left(\sum_{i=1}^{5} \hat{x}_i^R > \sum_{i=1}^{5} \hat{x}_i^G \middle| x_{1:5}^R, x_{1:5}^G \right) = \Phi\left(\frac{\sum_{i=1}^5 m(x_i^R) - \sum_{i=1}^5 m(x_i^G)}{\sqrt{\sum_{i=1}^5 s^2(x_i^R) + \sum_{i=1}^5 s^2(x_i^G)}} \right). \tag{2}$$

The first three models are special cases of the fourth one, with either an absence of bias (i.e., an identity transformation $m(x) = x$), a constant noise ($s(x) = s$), or both (Eq. (1)).

We fit these models to the behavioural data by maximizing their likelihoods. We flexibly specify the functions $m(x)$ and $s(x)$, by allowing them to be arbitrary low-order polynomials, defined by their values at a small number of points (between 2 and 8), with the order chosen to minimize the Bayesian Information Criterion (BIC). Running simulations of the fitted models, using the numbers presented to the subjects, we find that each of them predicts that choice frequency should be a similar sigmoid function of the difference between the two averages. The models differ, however, in the implied graphs for the decision weights. The unbiased model with varying noise ($N(x, s^2(x))$) still results in approximately linear decision weights, with equal weights for numbers at equal distances from the mean (Fig. 2c, top right panel). Instead the two models that include a transformation $m(x)$ of the presented numbers yield nonlinear decision weights resembling those of the subjects: numbers below the mean and close to it have higher weights that numbers above the mean and equally distant from it, while the opposite occurs for numbers further from the mean (Fig. 2c, bottom panels. More details on how the transformation $m(x)$ affects the decision weights can be found in Methods.) In summary, the two models which feature a transformation of the number, whether with a constant or variable noise, better capture the patterns observed in behavioural data.

In order to quantitatively compare the degree of explanatory success of the models, we compute the BIC for the best-fitting model in each class. We can either assume that the parameters of the models are identical under the three priors ("homogeneous parameters"), or, conversely, that the parameters depend on the prior ("prior-specific parameters"). The latter results in a larger number of parameters, which is penalized by the BIC. In spite of this penalty, in all four models the BIC is lower with prior-specific parameters than with homogeneous parameters, suggesting that subjects' decision-making process is adapted to the prior distribution (Fig. 2b). The two best-fitting one-stage models include a transformation $m(x)$ of the presented number, in accordance with our observations about the decision weights, and the best-fitting

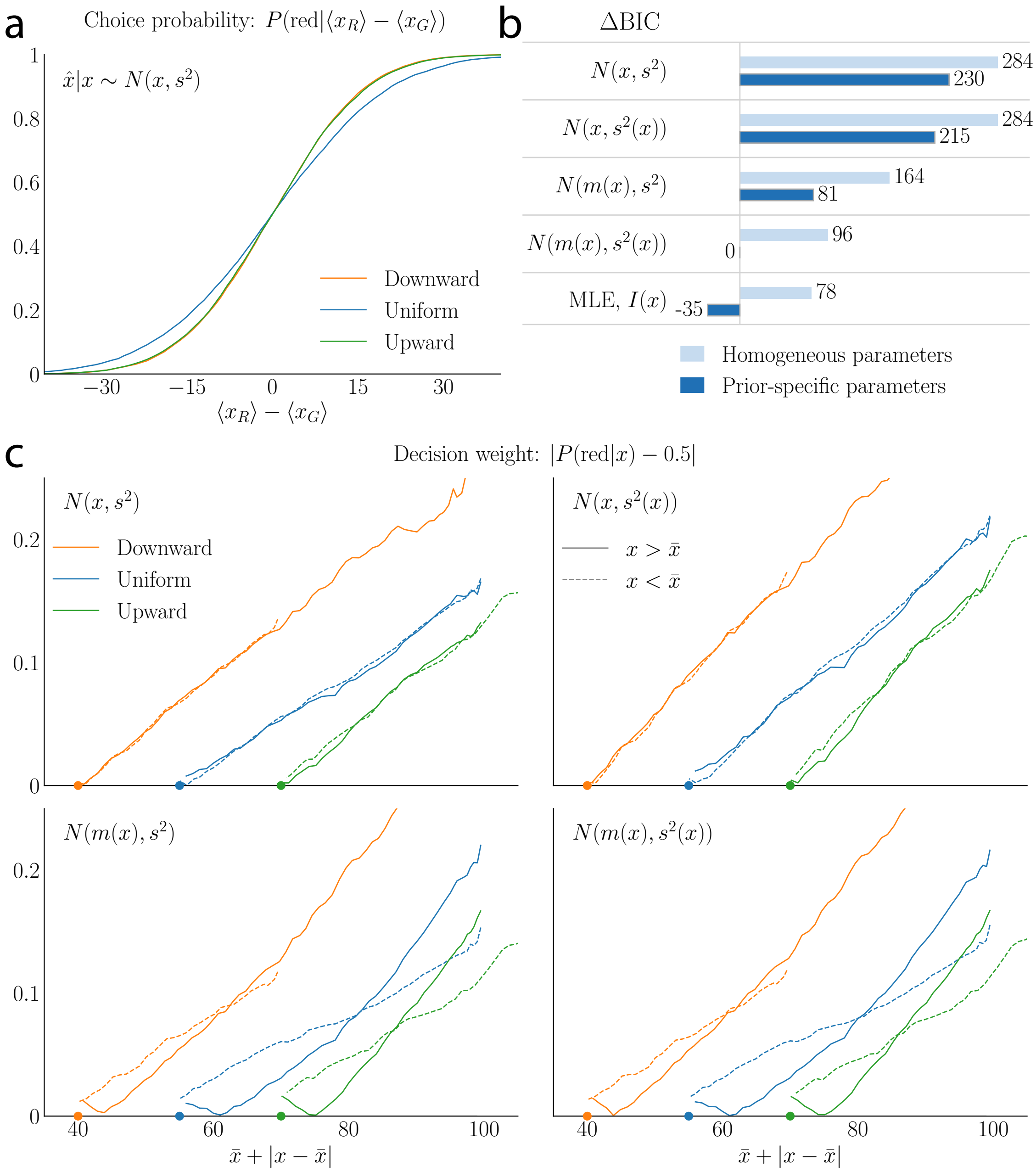

Figure 2: Models with both nonlinear transformation and varying noise best capture subjects' behaviour. a. Choice probability under the unbiased, constant-noise model ($N(x, s^2)$) as a function of the difference in the averages of the presented numbers, for the three prior conditions. **b.** Model comparison statistics for the four one-stage models of noisy estimation and for the MLE decoding model, both with homogeneous parameters and with prior-specific parameters. For each model class, the difference in BIC from the least restrictive model class (general transformation, variable noise, prior-specific) is reported. The lower BIC for the MLE model (a special case of the general one-stage model) indicates that it captures subjects' behaviour more parsimoniously. **c.** Decision weights $|P(\text{red}|x) - 0.5|$ for the unbiased, constant-noise model (top left), the unbiased, varying-noise model (top right), the model with transformation of the number and constant noise (bottom left), and the model with transformation of the number and varying noise (bottom right). Only the two models with transformation of the number reproduce the behaviour of subjects shown in Fig. 1e.

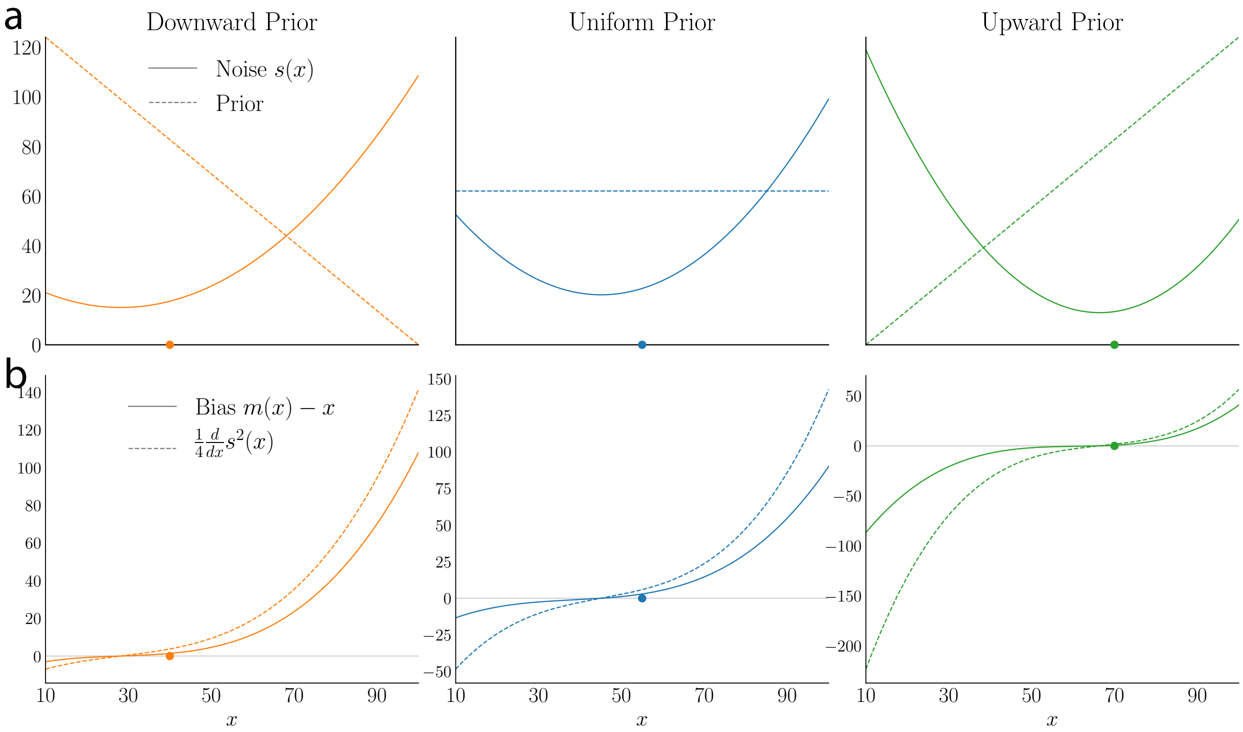

Figure 3: Best-fitting noise and bias: subjects encode less frequent numbers with greater noise. **a.** The best-fitting noise function $s(x)$ in the $N(m(x), s^2(x))$ model (solid line), and prior distribution (dashed line), in the Downward (left), Uniform (middle) and Upward (right) conditions. The scale refers to the noise (scale of the prior pdf not shown.) **b.** The bias (solid line) and the derivative of the noise variance (divided by 4, dashed line) for the best-fitting $N(m(x), s^2(x))$ model, as a function of x, in the three prior conditions. The MLE model requires these two functions to be the same (Eq. (6)).

model also features a variable noise $s(x)$. We note that the BIC in this case is significantly lower than with a transformation but maintaining constant noise ($\Delta\text{BIC} = 81$), in spite of the additional parameters of the noise function $s(x)$. In the Supporting Information we discuss the finer features of subjects' behaviour that are better captured by allowing for variable noise, showing that a nonlinear transformation alone (as in Ref. [6, 7]) does not suffice to account for our data.

The shape of the best-fitting noise function $s(x)$ for each prior is shown in Figure 3a. In each case, we find that it is U-shaped: minimal at a value close to the mean of the prior (but slightly below it), and increasing as a function of the distance from this minimum. In the Downward-prior condition, larger numbers (which have a low probability) are perceived with a much larger amount of noise than small numbers (which have a high probability). The noise function in the Upward condition approximately mirrors that in the Downward condition, with small numbers perceived with more noise than large numbers. In other words, the least likely numbers are perceived with the greatest randomness, suggesting that the process by which subjects estimate numbers is adapted to the prior distribution from which they are drawn.

The finding that the variance of $\hat{x}$ differs for different numbers x is a natural one if we suppose that, rather than summing transformed values of the individual numbers and adding noise only at the comparison stage, each individual number is encoded with noise, with the estimate $\hat{x}$ then representing an *inference* about the likely value of x based on its noisy internal representation r. In general, an optimal rule of inference will make the estimate $\hat{x}(r)$ a nonlinear function of r, so that the variance of $\hat{x}$ should depend on x even if we suppose that the variance of r does not. At the same time, an encoding-decoding model of this kind, in which decoding is assumed to be optimal given the nature of the noisy encoding, will imply that the function $m(x)$, indicating the bias in the average decoded value, will not be independent of $s(x)$. Thus this class of models represents a special case of the general specification $N(m(x), s^2(x))$, though a different restricted class than any of those considered above.

More precisely, suppose that a presented number, x, elicits in the brain a series of n signals, $r = (r_1, \ldots, r_n)$, each drawn independently from a distribution conditioned on the presented number, $p(r_i|x)$. And suppose further that the estimate $\hat{x}$ is the *maximum-likelihood estimator* (MLE) of x, given r: $\hat{x}(r) = \text{argmax}_x\, p(r|x)$. This implies a distribution of values for $\hat{x}$ in the case of any number x.

The MLE is known to be unbiased and efficient up to order $1/\sqrt{n}$, i.e.,

$$\sqrt{n}(\hat{x} - x) \xrightarrow{d} N\Big(0, \frac{1}{I_1(x)}\Big), \tag{3}$$

where $I_1(x)$ is the Fisher information of the likelihood $p(r_i|x)$ for an individual signal. To this order of approximation, the estimate is normally distributed around the presented number, with a variance equal to the inverse of the total Fisher information: $\hat{x}|x \sim N(x, 1/I(x))$, where $I(x) = nI_1(x)$. But to order $1/n$, the MLE is biased (see Ref. [22, 23, 24] and Methods). If the likelihood $p(r_i|x)$ is Gaussian, this bias is given (to order $1/n$) by

$$\text{Bias} \;\equiv\; \mathbb{E}(\hat{x} - x) \;=\; \frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}x}\Big(\frac{1}{I(x)}\Big). \tag{4}$$

Thus we can approximate the distribution of estimates as

$$\hat{x}|x \sim N\left(x + \frac{1}{4}\frac{\mathrm{d}}{\mathrm{d}x}\Big(\frac{1}{I(x)}\Big), \frac{1}{I(x)}\right). \tag{5}$$

As in the one-stage models considered above, the estimate is normally distributed around a transformation of the number. This is a special case of the general model considered above ($N(m(x), s^2(x))$), with the added constraint that both the nonlinear transformation $m(x)$ and the noise $s(x)$ are here determined by a single function, the Fisher information $I(x)$. In particular, the MLE model predicts a relation between the bias and the noise:

$$m(x) - x = \frac{1}{4}\frac{d}{dx}s^2(x), \tag{6}$$

so that the bias is predicted to be proportional to the derivative of the noise variance, $s^2(x)$.

We can test whether our data are consistent with the additional restriction given by Eq. (6). In Figure 3b, we plot the functions corresponding to the two sides of this equation, implied by our best-fitting estimates of $m(x)$ and $s(x)$ for each of the three conditions, when the theoretical restriction implied by the MLE model is not imposed in the estimation. Here it is important to note that the transformation $m(x)$ can be identified from choice data only up to an arbitrary affine transformation; this gives us two free parameters to choose in plotting the left-hand side of the equation. As explained in Methods, we choose these parameters to make the implied function more similar to the right-hand side of the equation. When we do so, we obtain the functions plotted in Figure 3b.

As noted above, the noise $s(x)$ is U-shaped; thus the derivative of the noise variance, $\frac{d}{dx}s^2(x)$, increases as a function of x; it vanishes at a value close to the prior mean, is negative below this value, and is positive above it (Fig. 3b, dashed line). With an appropriate choice of normalization for $m(x)$, the bias vanishes at the same value; in addition, we note that for all three priors the bias is also negative below this value and positive above it, and increases as a function of x (Fig. 3b, solid line). In other words, numbers below the prior mean are underestimated, while numbers above the prior mean are overestimated. Thus we find that the two functions, when fitted to subjects' data, are qualitatively consistent with the relation predicted by MLE decoding (Eq. (6)).

We can also estimate the MLE model, as defined by Eq. (5), finding the Fisher information function $I(x)$ that minimizes the BIC. As with our one-stage models, prior-specific parameters allow us to obtain a lower BIC than with homogeneous parameters. Furthermore, the BIC of the MLE model is lower than that of the model in which the functions $m(x)$ and $s(x)$ are unrestricted (Fig. 2b, $\Delta\text{BIC} = 35$). We conclude that the MLE decoding model captures more parsimoniously the behaviour of subjects.

In this model, a single function, the Fisher information, completely determines the statistics of responses. The best-fitting function differs, however, under the three priors. In the Downward condition, the Fisher information is a decreasing function of the number over most of the range of presented numbers (Fig. 4, left panel). In the Upward condition, conversely, the fitted Fisher information increases as a function of the number (except at large numbers; Fig. 4, right panel). Hence, in both the Downward and Upward conditions, the Fisher information is lower for numbers that are less likely under the prior. In the Uniform condition, the Fisher information peaks around 30, but varies less than in the other two conditions (Fig. 4, middle panel).

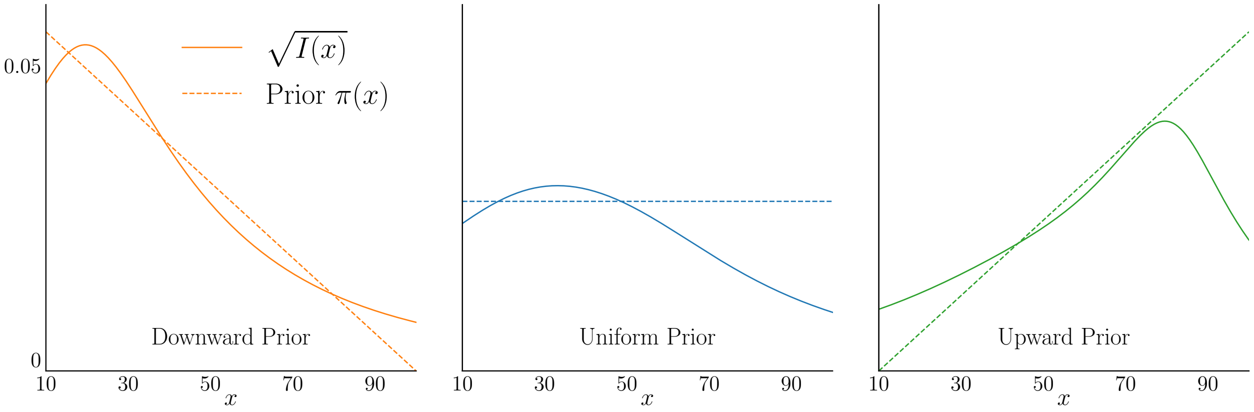

Figure 4: MLE model: the Fisher information, fitted to subjects' data, is adapted to the prior. a. Prior distribution (dashed lines) and square-root of the Fisher information fitted to subjects' data (solid lines), in the Downward (left), Uniform (middle) and Upward (right) conditions. The ordinate scale refers to the Fisher information (scale for the prior pdf not shown).

The behavioral implications of these fitted Fisher information functions are shown in Figure 5. Estimation noise is predicted to be U-shaped, and to reach a minimum where the Fisher information peaks (Fig. 5a). The bias vanishes at this same point; for smaller x, it is negative, while for higher x it is positive (Fig. 5b). Consequently the MLE model reproduces the unequal weighting of numbers in subjects' decisions (Fig. 5c,d) and the sigmoid shape of the choice probability curve (Fig. 5e).

Our estimates best fit subjects' data (even penalizing the additional free parameters) when the Fisher information is allowed to differ depending on the prior distribution from which numbers are sampled. Context-dependent encoding of this kind is predicted by theories of efficient coding; this leads us to ask whether the differing encoding rules that we observe represent efficient adaptations to the different priors, in the sense of maximizing average reward in our task. We investigate this hypothesis by examining the performance, over numbers sampled from a given prior (say, Downward), of a model subject equipped with the Fisher information fitted to subjects' data in the context of another prior (say, Upward). In other words, we look at how successful the "Upward encoding" (and associated MLE decoding) would be on "Downward data" (we use these shorthands below). We measure performance as follows. In each trial of our task, the score is augmented by the chosen average, so that the subject is guaranteed to receive at least the minimum of the two averages. The additional value to be captured in a trial i is thus the absolute difference between the two, i.e., $\Delta_i = |\langle x^R \rangle_i - \langle x^G \rangle_i|$. We define the "Performance ratio" as the fraction of this value captured by a model subject, i.e., $\sum \delta_i \Delta_i / \sum \Delta_i$, where δ_i is 1 if the model subject makes the correct choice in trial i and 0 otherwise.

With Downward data, the Downward encoding yields the largest performance ratio (87.5%), whereas the Upward encoding results in the lowest ratio (82.0%, Fig. 5f, left bars). Conversely, with Upward data

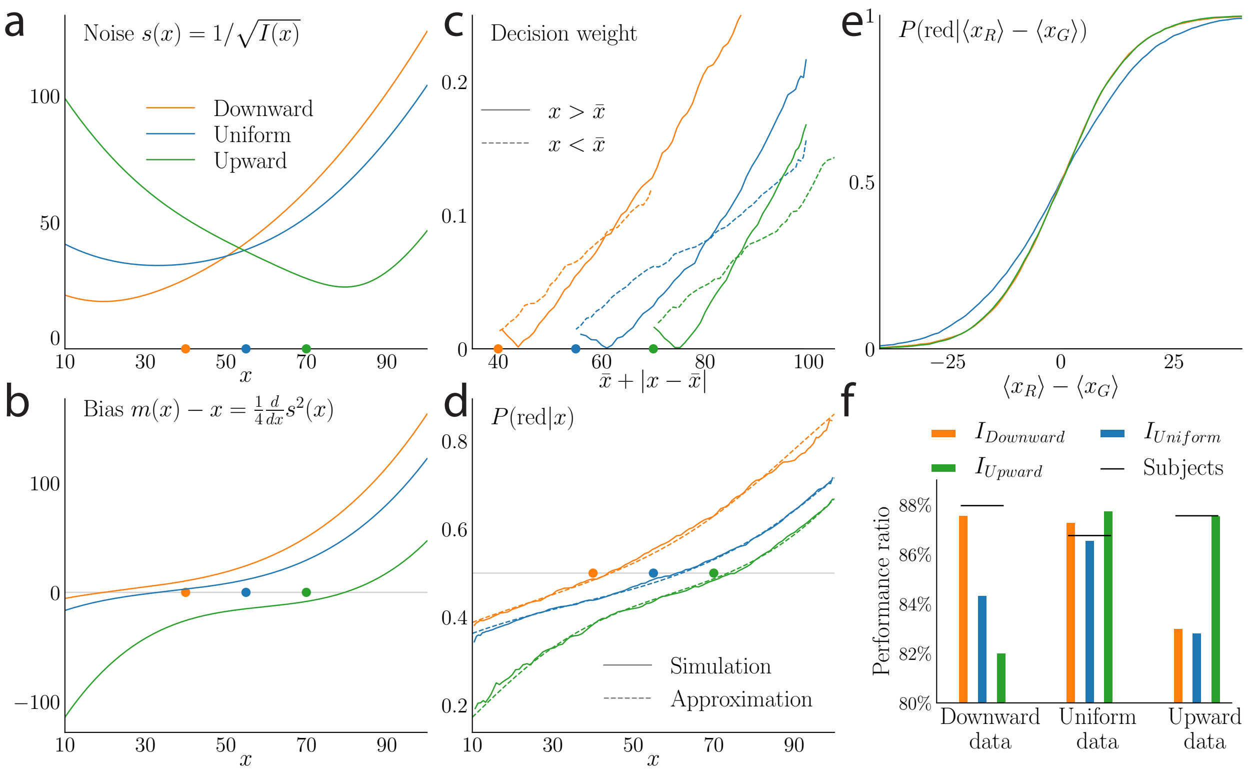

Figure 5: The MLE model reproduces subjects' behaviour, and the Fisher information functions fitted to subjects' data improve the performance ratio, in the Upward and Downward conditions. a. Noise $s(x)$ implied by the Fisher information, in the three prior conditions. **b.** Bias $m(x)-x$ in the three prior conditions, also implied by the Fisher information. **c.** Decision weights $|P(\text{red}|x) - 0.5|$ predicted by the MLE model. **d.** Probability of choosing 'red' conditional on a red number x being presented, in the MLE model (solid line), and our approximation to the model prediction (dashed line; see Methods). **e.** Choice probabilities implied by the MLE model as a function of the difference between the averages of the numbers presented. **f.** Performance ratios, over numbers sampled from the Upward (left), Uniform (middle) and Downward (right) priors, for the subjects (black lines) and implied by the estimated encoding rules (bars) for each of the three prior conditions. With numbers sampled from the Downward prior, the Downward encoding rule, $I_{Downward}$, yields the best performance (left orange bar), whereas with numbers sampled from the Upward prior, the Upward encoding rule, I_{Upward}, results in the best performance (right green bar).

the Upward encoding outperforms the Downward encoding (87.5% vs. 83.0%, Fig. 5f, right bars). In other words, the encoding rule fitted to the behaviour of subjects in the context of a given prior, Upward or Downward, results in higher rewards in the context of numbers sampled from this same prior, suggesting that the choice of encoding rule is efficient. With Uniform data, the performance ratio of the Uniform encoding is not quite as good as those of the Downward and Upward encoding rules (86.5% vs. 87.3% and 87.7%). However, these differences in performance are appreciably smaller than those observed with the Downward and Upward data (about 1 percentage point vs. about 5 percentage points). Moreover, the performance ratio of the subjects in the Uniform condition (86.8%) is lower than that in the Downward (88.0%) and Upward (87.6%) conditions. Consistent with this, the fitted Uniform encoding rule is noisier than the other two; and it is primarily this fact (rather than the way that the precision of encoding varies for different numbers) that makes the fitted Uniform encoding less efficient even for Uniform data.

Discussion

We designed an average-comparison task in which we changed the prior distribution from which numbers were sampled across blocks of trials. In their choices, subjects seem to differentially weight numbers that should be equally relevant to the correct decision. Subjects' behaviour can be characterized by a model of noisy perception of the size of numbers, in which both the average estimation bias and the variability of estimates depends on the magnitude of the number. Furthermore, we introduced an encoding-decoding model, in which the variable precision of encoding is specified by a Fisher information function, and the decoded value of each number is the maximum-likelihood estimate based on the encoded evidence. This is equivalent to a constrained version of the model of noisy perception (Eq. (5)), in which both the bias and the variable noise are determined by a single function, the Fisher information. This implies a relation between the bias and the derivative of the noise variance (Eq. (6)). The MLE model yields the lowest BIC among the models considered, and reproduces the patterns observed in behaviour. Furthermore, the Fisher information fitted to subjects' data varies depending on the prior distribution. With the two skewed priors (Upward and Downward priors,) the Fisher information is lower for numbers less likely to appear under the prior (Fig. 4). This resulted in a higher performance in the task (Fig. 5f), suggesting that within the one-hour timeframe of the experiment, subjects efficiently adapted their decision-making process to the prior distribution of presented numbers.

Reference [6] had previously found that a model featuring a nonlinear transformation of presented numbers (i.e., a bias) reproduced the apparent unequal weighting of numbers in subjects' decisions, and interpreted the bias as a strategy to compensate for decision noise. We provide a different account, that relies on two further observations. First, the noise in the perceived value of a number seems to vary with the number, and to depend on the prior. Second, the way that the bias and the noise function vary from one condition to another is consistent with the theoretical relation between these two functions predicted by our model of MLE decoding. Under this account, estimation noise, estimation bias, and thus the unequal weighting of numbers in average comparisons and the sigmoid-shape choice probability function, all derive from the form of the Fisher information function. In turn, the Fisher information is an increasing function, at least roughly, of the prior density from which numbers are sampled, suggesting that the encoding is efficiently adapted to the prior.

The idea that the nervous system encodes stimuli efficiently was first proposed by H. Barlow [25], and finds a modern formulation in the recent literature on neural coding [17, 18, 19, 20, 26]. One well-known formulation of efficient coding (e.g., [20]) implies conditional probabilities $p(r|x)$ such that the square root of the Fisher information should be proportional to the prior density, i.e.,

$$\sqrt{I(x)} \propto \pi(x), \tag{7}$$

where π is the prior. Here, under the hypotheses of our model (in which no a priori assumptions on the Fisher information were made,) and on the basis of decision data, we empirically estimate the Fisher information implied by subjects' encoding rule in the three prior conditions. We find that its square root is somewhat similar to the prior density (Fig. 4); thus our results are consistent with the idea that more "representational capacity" is allocated to stimuli more likely to appear.

In the MLE model, the magnitude of the Fisher information for a given stimulus is the inverse of the variance of the subjective estimate of that stimulus. In addition, the estimate should be biased, and the MLE model predicts that a mathematical relation should exist between the bias and the derivative of the estimation variance (Eq. (6)). A quantity related to the estimation variance and often measured in perceptual tasks is the discrimination threshold, $DT(x)$, which quantifies the sensitivity of an observer to small changes in the stimulus. From Eq. (6) we can derive, as shown in Methods, a relation between the bias and the discrimination threshold, as

$$\mathbb{E}(\hat{x} - x) \propto \frac{d}{dx} DT^2(x). \tag{8}$$

Wei and Stocker [21] derive this same relation from different assumptions. They show that it is supported by numerous empirical results obtained in perceptual tasks, and thus they call it a "law of human perception". Our MLE model predicts this law, so that our model is also supported by the numerous results just mentioned, in addition to being the best-fitting model in the context of our task.

A key assumption in Wei and Stocker's derivation of Equation (8), however, is the efficient-coding relation (Eq. (7)). Although our experimental results are broadly consistent with this relation, our derivation of (8) does not rely on it. Hence, the law of perception might derive not from efficient coding, but simply from maximum-likelihood decoding of noisy internal representations. Further investigations are necessary to determine whether one can discriminate between these two competing explanations.

In the context of our task, the Fisher information implied by subjects' choices (interpreted using the MLE model) is roughly consistent with the efficient-coding relation, as mentioned above — but discrepancies remain. First, there is a "boundary effect," whereby the Fisher information decreases (despite the rising prior density) near the lower boundary of the interval of presented numbers, in the Downward condition, and near the upper boundary, in the Upward condition (Fig. 4). Second, in the Uniform condition (in which the prior is constant), the Fisher information is not exactly constant; nor is it symmetric around the center of the interval, despite the symmetry of the problem in this case. Instead, the Fisher information is somewhat lower for large numbers than for small numbers.

These discrepancies might indicate that the standard theory of efficient coding does not perfectly apply to our case, for any of a variety of reasons. First, the standard theory assumes an encoding rule that maximizes mutual information between the stimulus and the representation. In our task, subjects might instead maximize the financial reward they can expect; the optimal Fisher information is presumably somewhat different for a different objective. Second, the efficient-coding relation assumes additive, Gaussian, and vanishingly small noise. When noise is large, the optimal Fisher information differs from the prediction of Eq. (7), in particular at the boundaries [20]. And third, studies of numerosity perception suggest that the internal representation of numbers is consistent with Fechner's law, i.e., that larger numbers are discriminated less accurately than smaller ones [27]. Thus it is possible that the asymmetry of the Fisher information in the Uniform condition reflects a logarithmic encoding of numbers. We leave these questions to future investigations.

Methods

Experiment, subjects, and reward. The experiment was conducted at Columbia University (IRB Protocol Number: IRB-AAAR9375). 37 subjects, 18 female and 19 male, aged 24.5 on average, participated in the experiment. Each subject participated in two blocks of trials, in each of which all numbers were samples from a single distribution (Uniform, Upward or Downward), so that each subject experienced two of the three conditions. Subjects were explicitly told the current distribution, and at the beginning of each new block they were presented with a series of random samples, in order to familiarize them with the distribution. Each block of trials consisted of 200 trials, so that 2x200=400 decisions were collected per subject. Each session lasted about one hour. On screen, the numbers were presented with two decimal digits, and for a duration of 500ms. In each trial, the color of the first presented number was chosen between red and green with equal probability. The score of the subject was augmented at each trial by the chosen average, and thus increased over the course of the experiment. At the end of the experiment, the subject received a financial reward, which is a linear function of the total score, with a $10 "show-up" minimum. The expected reward, not taking into account the $10 minimum, for a hypothetical subject providing random responses (i.e., choosing "red" with probability 0.5 at all trials) was $10, and an accuracy of 80% of correct responses yielded an average of $25. The average reward over the 37 subjects was $28.

The transformation $m(x)$ and decision weights. To shed light on the relation between the transformation $m(x)$ and the decision weights, we derive an approximation to the probability $P(red|x)$ of choosing 'red' conditional on a red number x being presented. Consider first the model with constant noise ($N(m(x), s^2)$). The probability can be computed by marginalization of the probability of choosing 'red' conditional on ten numbers as

$$\begin{aligned} P(red|x) &= \int \dots \int P(red|x, x^R_{2:5}, x^G_{1:5}) \pi(x^R_2) \dots \pi(x^G_5) dx^R_2 \dots dx^G_5 \\ &= \int \Phi\Big(\frac{m(x)+\Delta_m}{s\sqrt{10}}\Big) \pi_\Delta(\Delta_m) d\Delta_m, \end{aligned} \tag{9}$$

where

$$\Delta_m \equiv \sum_{i=2}^{5} m(x^R_i) - \sum_{i=1}^{5} m(x^G_i) \tag{10}$$

and π_Δ is the prior density for this random quantity. We approximate π_Δ by a Gaussian distribution with the same mean and variance, so that

$$\pi_\Delta \approx N(-\bar{m}, 9\text{Var}m), \tag{11}$$

where $\bar{m}$ is mean of $m(x)$ under the prior and Var$m$ is the variance. Substituting this approximation in Eq. (9) results in

$$P(\text{`red'}|x) \approx \Phi\left(\frac{m(x)-\bar{m}}{\sqrt{10s^2 + 9\text{Var}m}}\right). \tag{12}$$

We find this approximation to be fairly close to the conditional probabilities obtained through simulations of the model. In case of variable noise ($N(m(x), s^2(x))$), we replace in Eq. (12) the noise variance s^2 by its average under the prior, $\mathbb{E}s^2(x)$, and despite this coarse approximation, we also obtain a close match to simulated data. Figure 5d provides an example of the quality of these approximations, in the case of the MLE model.

Equation (12) implies that the decision weight, $|P(\text{`red'}|x) - 0.5|$, vanishes at approximately the number whose transformation equals the average transformation. In the absence of bias ($m(x) = x$), this number would be the prior mean; but it is slightly greater in the case of the fitted transformations. Hence the decision weights at the prior mean do not fall to zero, and consequently numbers above and below the prior mean have different weights (Fig. 2c).

Bias of the maximum-likelihood estimator. References [22, 23, 24] show that the MLE derived from n samples drawn from a distribution parameterized by x has the following bias:

$$\mathbb{E}(\hat{x}-x) = -\frac{I'(x)+I_3(x)}{4nI^2(x)}, \tag{13}$$

where I is the Fisher information, I' is the derivative of I and

$$I_3(x) = \int p(r_i|x)\Big(\frac{\mathrm{d}\ln p(r_i|x)}{\mathrm{d}x}\Big)^3 \mathrm{d}r_i. \tag{14}$$

(We provide in Supporting Information the main steps of a demonstration of this result.) Hence, approximately,

$$\hat{x}|x \sim N\Big(x - \frac{I'(x)+I_3(x)}{4nI^2(x)}, \frac{1}{nI(x)}\Big). \tag{15}$$

We make the additional assumption that the likelihood is a Gaussian distribution centered on a transformation $\mu(x)$ of the number and with constant variance ν^2, i.e., $r_i|x \sim N(\mu(x), \nu^2)$. In this case, the quantity $I_3(x)$ vanishes, and we obtain the model described by Eq. (5). We make this Gaussianity assumption for illustrative purposes; in fact we only need to assume that $I_3 = 0$.

Model fitting and identification. When fitting subjects' data to the general one-stage model $N(m(x), s^2(x))$, it is important to note that the transformation $m(x)$ can be identified from choice data only up to an arbitrary affine transformation: a model with transformation $\alpha m(x) + \beta$ and noise $\alpha s(x)$ makes the same predictions as the model with transformation $m(x)$ and noise $s(x)$, for any non-zero α and any β (see Eq. (2)). In calculating the bias plotted in Figure 3b, we choose the 'scale and location' parameters α and β to satisfy two additional desiderata. First, for any choice of α, we choose β so that the left- and right-hand sides of (6) are exactly equal at the particular value of x where the right-hand side is equal to zero. And second, given this, we choose α so as to minimize

$$\frac{\int(LHS(x) - RHS(x))^2 dx}{\langle|RHS|\rangle^2},$$

a measure of the relative difference between the two functions $LHS(x)$ and $RHS(x)$ defined by the two sides of the equation.

Relations between bias, variance, and discrimination threshold. The discrimination threshold $DT(x)$ is defined as the difference δ in stimulus magnitude for which a subject distinguishes two stimuli x and $x+\delta$ with a given success rate (e.g., 75%.) In a model in which an estimate $\hat{x}$ of presented number x is normally distributed around a transformation $m(x)$ of the number, with varying noise $s(x)$, i.e., $\hat{x} \sim N(m(x), s^2(x))$, the probability of telling $x_1 = x$ and $x_2 = x + \delta$ apart is, assuming δ small,

$$\begin{aligned} P(\hat{x}_2 > \hat{x}_1) &= \Phi\Big(\frac{m(x_2) - m(x_1)}{\sqrt{s^2(x_2) + s^2(x_1)}}\Big) \\ &\approx \Phi\Big(\frac{m'(x)}{s(x)}\delta\Big) \\ &\approx \frac{1}{2} + \frac{1}{\sqrt{2\pi}}\frac{m'(x)}{s(x)}\delta, \end{aligned}$$

This implies

$$DT(x) \propto \frac{s(x)}{m'(x)}, \tag{16}$$

where the proportionality factor depends on the chosen target success rate. In the MLE model, the bias $m(x) - x$ is proportional to the derivative of the inverse of the Fisher information, $I(x) = nI_1(x)$, therefore the bias is of order $\mathcal{O}(1/n)$, and $m'(x) \approx 1$. Equations (16) and (6) then immediately results in Wei and Stocker's law of human perception [21] (Eq. (8)):

$$\begin{aligned} \frac{d}{dx}DT^2(x) &\propto \frac{d}{dx}s^2(x) \\ &\propto m(x) - x = \mathbb{E}(\hat{x} - x). \end{aligned}$$

We note, in addition, that the converse derivation appears in the Supporting Information of Ref. [21]: from the relation involving the discrimination threshold (Eq. (8)), the authors derive a relation involving the noise variance, as in Eq. (6).

References

[1] Amos Tversky. Elimination by aspects: a theory of choice. *Psychological Review*, 79(4), 1972.

[2] John W Payne, R Bettman, and Eric J Johnson. *The adaptive decision maker.* Cambridge University Press, 1993.

[3] Gerd Gigerenzer and Daniel G Goldstein. Reasoning the Fast and Frugal Way : Models of Bounded Rationality. *Psychological Review*, 103(4):650–669, 1996.

[4] Eric J Johnson and Roger Ratcliff. Computational and Process Models of Decision Making in Psychology and Behavioral Economics. In *Neuroeconomics*, chapter 3, pages 35–48. Elsevier Inc., 2014.

[5] Christopher Summerfield and Konstantinos Tsetsos. Do humans make good decisions ? *Trends in Cognitive Sciences*, 19(1):27–34, 2015.

[6] Bernhard Spitzer, Leonhard Waschke, and Christopher Summerfield. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, 1(8):1–8, 2017.

[7] Vickie Li, Santiago Herce Castañon, Joshua A Solomon, Hildward Vandormael, and Christopher Summerfield. Robust averaging protects decisions from noise in neural computations. *PLoS Computational Biology*, 13(8):1–19, 2017.

[8] Vincent De Gardelle and Christopher Summerfield. Robust averaging during perceptual judgment. *Proceedings of the National Academy of Sciences*, 108(32), 2011.

[9] Konstantinos Tsetsos, Rani Moran, James Moreland, Nick Chater, Marius Usher, and Christopher Summerfield. Economic irrationality is optimal during noisy decision making. *Proceedings of the National Academy of Sciences*, 113(11):3102–3107, 2016.

[10] David C Knill and Whitman Richards. *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, 1996.

[11] Marc O Ernst and Martin S Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, jan 2002.

[12] Alan A Stocker and Eero P Simoncelli. Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, 9(4):578–585, apr 2006.

[13] Ahna R Girshick, Michael S Landy, and Eero P Simoncelli. Cardinal rules: visual orientation perception reflects knowledge of environmental statistics. *Nature Neuroscience*, 14(7):926–932, jul 2011.

[14] Frederike H. Petzschner, Stefan Glasauer, and Klaas E. Stephan. A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, 19(5):285–293, 2015.

[15] Xue-Xin Wei and Alan A. Stocker. A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience*, 18(10):1509–1517, 2015.

[16] Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60, 1994.

[17] Nicolas Brunel and J P Nadal. Mutual information, Fisher information, and population coding. *Neural computation*, 10(7):1731–57, 1998.

[18] Deep Ganguli and Eero P Simoncelli. Implicit encoding of prior probabilities in optimal neural populations. *Advances in neural information processing systems*, 2010:658–666, 2010.

[19] Deep Ganguli and Eero P. Simoncelli. Neural and perceptual signatures of efficient sensory coding. *ArXiv e-prints*, pages 1–24, feb 2016.

[20] Xue-Xin Wei and Alan A Stocker. Mutual Information, Fisher Information, and Efficient Coding. *Neural Computation*, 326:305–326, 2016.

[21] Xue-Xin Wei and Alan A Stocker. Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences*, 114(38):10244–10249, 2017.

[22] J. B. S. Haldane and Sheila Maynard Smith. The Sampling Distribution of a Maximum-Likelihood Estimate. *Biometrika*, 43:96–103, 1956.

[23] L Shenton. The Distribution of Moment Estimators. *Biometrika*, 46(3/4):296–305, 1959.

[24] D. R. Cox and E. J. Snell. A General Definition of Residuals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 30(2):248–275, 1968.

[25] H. B. Barlow. Possible Principles Underlying the Transformations of Sensory Messages. In Walter A. Rosenblith, editor, *Sensory Communication*, chapter 13, pages 217–234. The MIT Press, Cambridge, MA, sep 1961.

[26] Mark D. McDonnell and Nigel G. Stocks. Maximally informative stimuli and tuning curves for sigmoidal rate-coding neurons and populations. *Physical Review Letters*, 101(5):1–4, 2008.

[27] Véronique Izard and Stanislas Dehaene. Calibrating the mental number line. *Cognition*, 106(3):1221–1247, 2008.