A Perceptual Confirmation Bias from Approximate Online Inference

Richard D. Lange[1,2,⋆], Ankani Chattoraj[1], Matthew Hochberg[1], Jeffrey M. Beck[3], Jacob L. Yates[1], Ralf M. Haefner[1]

1 Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA

2 Computer Science, University of Rochester, Rochester, NY 14627, USA

3 Department of Neurobiology, Duke University, Durham, NC 27708, USA

⋆ rlange@ur.rochester.edu

Abstract

The mechanisms underlying evidence accumulation in perceptual decision-making tasks have been the subject of much research. However, existing studies differ in their conclusions on whether the brain weighs evidence optimally over time, or whether it exhibits biases towards evidence presented early (primacy) or later (recency) in the trial. We resolve this discrepancy in the literature by proposing that previous tasks differ in how task-relevant information in the stimulus is partitioned into "sensory information" and "category information." We demonstrate that similar stimulus-dependent biases arise naturally in two common models of approximate inference: neural sampling-based inference, and parametric inference (Variational Bayes). Finally, we test our model by designing a psychophysics task that systematically trades off these two sources of uncertainty in the stimulus against each other while keeping all other aspects of the task the same. We find that subjects' evidence-weighting strategies change in the predicted direction and in a highly robust fashion, individually significant for every one of our 10 subjects.

Keywords: reverse correlation, approximate inference, 2AFC

Introduction

Quickly categorizing and acting on perceptual information is a crucial function of the brain. To elucidate the mechanisms underlying this ability, psychophysics studies often present subjects with noisy or ambiguous stimuli that must be integrated before a subject can make an informed categorical decision. A classic example is the motion-dots task, in which a subject views an aperture of drifting and flickering dots and has to report whether the dots are moving to the right or left. Crucially, a single "frame" of motion only gives weak evidence about the correct choice, but by integrating over many frames subjects are able to better discriminate left from right motion (Gold & Shadlen, 2007). Note that our use of "frame" is not restricted to visual stimuli, but refers to the rapid, sequential presentation of independent stimuli in any modality.

In this and other commonly used tasks, each frame is on average equally predictive of the correct choice, hence *optimal* performance requires weighting each frame equally to arrive at a decision. The *actual* weights used by a given subject in a given task can be measured using reverse correlation. Existing studies differ greatly in their findings, with temporal weight profiles ranging from decreasing (early evidence is weighed most strongly – a primacy effect) (Nienborg & Cumming, 2009; Kiani, Hanks, & Shadlen, 2008) to constant (as would be optimal) (Wyart, Gardelle, Scholl, & Summerfield, 2012; Brunton, Botvinick, & Brody, 2013), to increasing (most recent evidence is weighed most strongly – a recency effect) (Drugowitsch, Wyart, Devauchelle, & Koechlin, 2016). An ideal observer, by definition, cannot explain these systematic biases. Furthermore, existing accounts for any one of these biases ("integration to bound" to explain early weighting, or "forgetting" to explain late weighting) only provide explanations for subsets of the experiments, unable to account for their difference, or under what circumstances which effect should dominate.

Here, we make three novel contributions. First, we observe that previous tasks differ in the nature of the uncertainty introduced into the stimulus in order to make the task difficult for the subject, and we show that those differences are predictive of the kind of bias found in prior studies. Second, we present the results of an experiment in which the stimulus is systematically varied only along the axis that we hypothesize to determine the subject's bias. While previous studies differed widely in sensory modality, subject species, and task-relevant variable, here we confirm the effect of stimulus design on measured evidence-weighting strategy on an individual subject level. Third, We demonstrate how these biases can emerge naturally from the very same *approximate inference* algorithm to perform the task. In particular, we show that two approximate inference models, based on the MCMC Sampling and Variational Bayes algorithms, both exhibit a primacy effect for the same kind of stimulus that produces such a bias empirically. Both model families have been previously proposed as plausible inference mechanisms by neural circuits (Fiser, Berkes, Orbán, & Lengyel, 2010; Pouget, Beck, Ma, & Latham, 2013; Haefner, Berkes, & Fiser, 2016; Raju & Pitkow, 2016). We further show that either explicitly trying to compensate for that bias, or allowing for some noise in the brain's decision-making circuits, will lead to a recency effect specifically for stimuli used in those experiments that find constant weights or recency effects. Either of our approximate inference models is able to explain the full range of primacy to recency effects seen in previous studies.

Results

'Sensory Information' vs 'Category Information'

Normative models of two-alternative choice tasks usually begin with the *ideal observer*, who uses Bayes' rule to infer the

best choice on each trial given the stimulus. On a given trial, if the evidence in frame f is e_{f} and the (correct) categorical identity of the stimulus is a binary variable $C \in \{-1, +1\}$, then evidence in favor of $C = +1$ after F independent frames is $\mathrm{p}(C = +1|e_1, \ldots, e_F) \propto \mathrm{p}(C = +1) \prod_{\mathrm{f}=1}^{F} \mathrm{p}(e_{\mathrm{f}}|C = +1)$.

The ideal observer's performance is limited only by (1) the information about C available on a single frame, proportional to $p(e_{\mathrm{f}}|C)$, and (2) the number of frames per trial. In the case of the brain, a decision-making area computing a belief about the correct choice only has access to a sensory representation of the stimulus, which we call x, not to the outside stimulus e directly. As a direct consequence, we can partition the information between outside stimulus and choice into the information between outside stimulus and sensory representation, and the information between sensory representation and choice (Figure 1b). We call these **sensory information** and **category information** respectively. These two kinds of information span a two-dimensional space with a task being defined by a single point (Figure 1c).

Two well-known tasks (the dot motion task and the Poisson clicks task) occupy opposite locations in the space spanned by sensory and category information. In the classic dot motion task, the sensory information is low since the evidence about the direction of motion at any time is very small. The category information, on the other hand, is high, since knowing the "true" motion on a single frame would be highly predictive of the correct choice. In the case of the Poisson clicks task, the sensory information is high since each click is well above threshold, while the category information is low since knowing the side on which a single click was presented provides only little information about the correct choice. To drive home this distinction, consider an orientation discrimination task. Showing multiple frames of high-contrast gratings of the same orientation would constitute a task with high sensory information (the orientation of each frame is clear) *and* high category information (knowing the orientation on any one frame would determined the choice since all frames have the same orientation). Reducing sensory information would mean making each frame noisier (e.g. by reducing contrast or adding pixel noise) and hence increasing the subject's uncertainty about the correct orientation for each individual frame – moving the task left in our space. Reducing category information would mean varying the orientation between frames and hence making each frame less predictive of the correct choice – moving the task downwards in our space (Figure 1c). Subjects will be at threshold performance when the stimulus reaches some level of trade-off between sensory and category information, defining a line illustrated in Figure 1c.

A qualitative placement of prior studies in this space suggests that studies that find early weighting are located in the upper left quadrant and studies with equal or late weighting in the lower right quadrant. A quantitatively precise placement of each study is difficult since the quantitative partition depends on the nature of x. Since prior studies differ in many aspects like subject species (e.g. humans, monkeys, rats), sensory modality (e.g. visual, auditory), sensory variable (e.g. orientation, motion direction, depth), and stimulus parameters like number of frames and frame duration, it is hard to draw a definitive conclusion on whether there is a single variable that determines the shape of the temporal weighting profile.

Visual Discrimination Task

To test our hypothesis that sensory information and category information determine subjects' evidence weighting strategies, we designed a visual discrimination task that allows us to independently manipulate both sources of information while keeping all other aspects of the task constant (Figure 1i-j).

The stimulus in our task consisted of ten visual frames. Each frame consisted of band-pass-filtered noise with excess orientation power either in the -45deg or the $+45$deg orientation (Nienborg & Cumming, 2014). Here, the excess orientation power, parameterized by $0 \leq \kappa < \infty$, determines the uncertainty over orientation for each frame (sensory information). The probability, $0.5 \leq p \leq 1$, that the orientation of any one frame matched the rewarded choice corresponds to the category information. The stimulus was presented as an annulus around the fixation marker in order to minimize the effect of small fixational eye movements. Using this stimulus, we ran 10 human subjects (7 naive, 3 authors) comparing two conditions. Starting with both high sensory and high category information, we either ran a staircase lowering the sensory information (κ) until subjects reached threshold performance while keeping category information constant ("noise" condition), or we ran a staircase lowering category information while keeping sensory information constant ("ratio" condition). For each condition, we used regularized logistic regression to infer subjects' temporal evidence weighting profiles (Figure 1k-l).

In agreement with our hypothesis, we find predominantly flat or increasing weighting profiles in the "ratio" condition, and predominantly decreasing weighting profiles in our "noise" condition. A within-subject comparison revealed that the change in average slope between the two conditions was individually significant for every single subject.

Approximate Online Inference Explains Data

While these significant changes in evidence weighting for different stimulus statistics could reflect a fundamental change in subjects' strategies, we show here that they arise naturally in common models of approximate inference proposed for the brain. In particular, we show that both a neural sampling-based approximation and a parametric (mean field) approximation to sequential decision-making induce a bias towards overweighting early evidence when sensory information is low and category information is high, as seen in the data (Figure 1e-h).

The central assumption in both models is that the brain computes a posterior over both C and x given the external evidence, i.e. $\mathrm{p}(x, C|e)$, not just over the variable C which happens to be task-relevant in our particular context. As a result, the sensory representation encoding the brain's belief about

x will depend both on the external evidence, e, via the likelihood, but also on the brain's current belief about C, via the prior. When the sensory representation is then used to update the brain's belief about C, care must be taken to not "double-count" the current belief about C which appears both in the belief update about C in the prior over x. We call this double-counting a **perceptual confirmation bias** since it results in a positive feedback loop between beliefs about a stimulus and the percept.

As in the Sequential Probability Ratio Test (Gold & Shadlen, 2007), we assume the brain approximately computes beliefs about the correct choice as

$$\begin{aligned}\log\frac{\mathrm{p_f}(C=+1)}{\mathrm{p_f}(C=-1)} &\equiv \log\frac{\mathrm{p}(C=+1|e_1,\ldots,e_\mathrm{f})}{\mathrm{p}(C=-1|e_1,\ldots,e_\mathrm{f})}\\ &= \log\frac{\mathrm{p_{f-1}}(C=+1)}{\mathrm{p_{f-1}}(C=-1)} + \log\frac{\mathrm{p}(e_\mathrm{f}|C=+1)}{\mathrm{p}(e_\mathrm{f}|C=-1)}\\ &= \log\frac{\mathrm{p_{f-1}}(C=+1)}{\mathrm{p_{f-1}}(C=-1)} + \underbrace{\log\frac{\int_x \mathrm{p}(e_\mathrm{f}|x)\mathrm{p}(x|C=+1)\mathrm{d}x}{\int_x \mathrm{p}(e_\mathrm{f}|x)\mathrm{p}(x|C=-1)\mathrm{d}x}}_{\text{update per frame}}\end{aligned} \quad (1)$$

where the last line makes explicit the need to marginalize over beliefs about sensory variables x on each frame.

Neural sampling-based approximation Our first model makes three crucial assumptions. First, as described above, we assume that sensory areas of the brain represent a posterior over x using the current belief about C as a prior. Second, we assume that this posterior is represented by samples over time. Note that these two assumptions alone do not yet preclude exact inference; computing the integrals in equation (1) via importance sampling gives an asymptotically exact update rule that effectively accounts for the prior fed back from C to x. Our third assumption in this model is that, due to a fundamental limit on sampling speed, the brain must rely on a small number of samples per frame. These three assumptions are sufficient to reproduce the transition from primacy to flat evidence weighting as stimuli move from the low to high sensory information regime along the threshold performance line (Figure 1e-f). Including an active bias correction (subtracting $\gamma\log\frac{\mathrm{p_f}(C=+1)}{\mathrm{p_f}(C=-1)}$ on each update, Figure 1g-h) or noise in evidence accumulation (not shown) qualitatively reproduces the full range of primacy to recency effects seen in the data.

Parametric approximation Our second model uses a parametric representation of both C and x, but makes the mean field assumption that the posterior $\mathrm{p}(x,C|e)$ can be approximated by a factorized distribution, $\mathrm{q}(x)\mathrm{q}(C)$. This assumption is plausible for the brain because a central challenge for parametric approximations is the explosion of the number of parameters necessary to approximate joint posteriors over many variables. As a result, it is commonly assumed that the brain explicitly accounts for dependencies only between subsets of variables, and not between variables represented by different layers of the cortical hierarchy (e.g. sensory and decision areas) (Raju & Pitkow, 2016).

We simulated inference in this model with Variational Bayes. As in the sampling model, we use the running posterior estimate of $\mathrm{p_f}(C)$ as a prior over x. This parametric model displays the same behavior as the sampling model: a transition from primacy to flat kernels as sensory information increases, with recency effects emerging when a bias correction or noise in the decision-making area are added. Whereas the limited number of samples was the key deviation from optimality in the sampling model, here it is the assumption that the brain represents a factorized posterior over x and C.

References

Brunton, B. W., Botvinick, M. M., & Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. Science, 340(6128), 95–8.

Drugowitsch, J., Wyart, V., Devauchelle, A.-D., & Koechlin, E. (2016). Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. Neuron, 92(6), 1398–1411.

Fiser, J. J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: from behavior to neural representations. Trends in cognitive sciences, 14(3), 119–30.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annual review of neuroscience, 30(30), 535–574.

Haefner, R. M., Berkes, P., & Fiser, J. (2016). Perceptual Decision-Making as Probabilistic Inference by Neural Sampling. Neuron, 90(3), 649–660.

Kiani, R., Hanks, T. D., & Shadlen, M. N. (2008). Bounded integration in parietal cortex underlies decisions even when viewing duration is dictated by the environment. The Journal of neuroscience, 28(12), 3017–3029.

Nienborg, H., & Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. Nature, 459(7243), 89–92.

Nienborg, H., & Cumming, B. G. (2014). Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. The Journal of neuroscience, 34(10), 3579–85.

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. Nature Neuroscience, 16(9), 1170–8.

Raju, R. V., & Pitkow, X. (2016). Inference by Reparameterization in Neural Population Codes. NIPS, 30.

Wyart, V., Gardelle, V. D., Scholl, J., & Summerfield, C. (2012). Rhythmic Fluctuations in Evidence Accumulation during Decision Making in the Human Brain. Neuron, 76(4), 847–858.

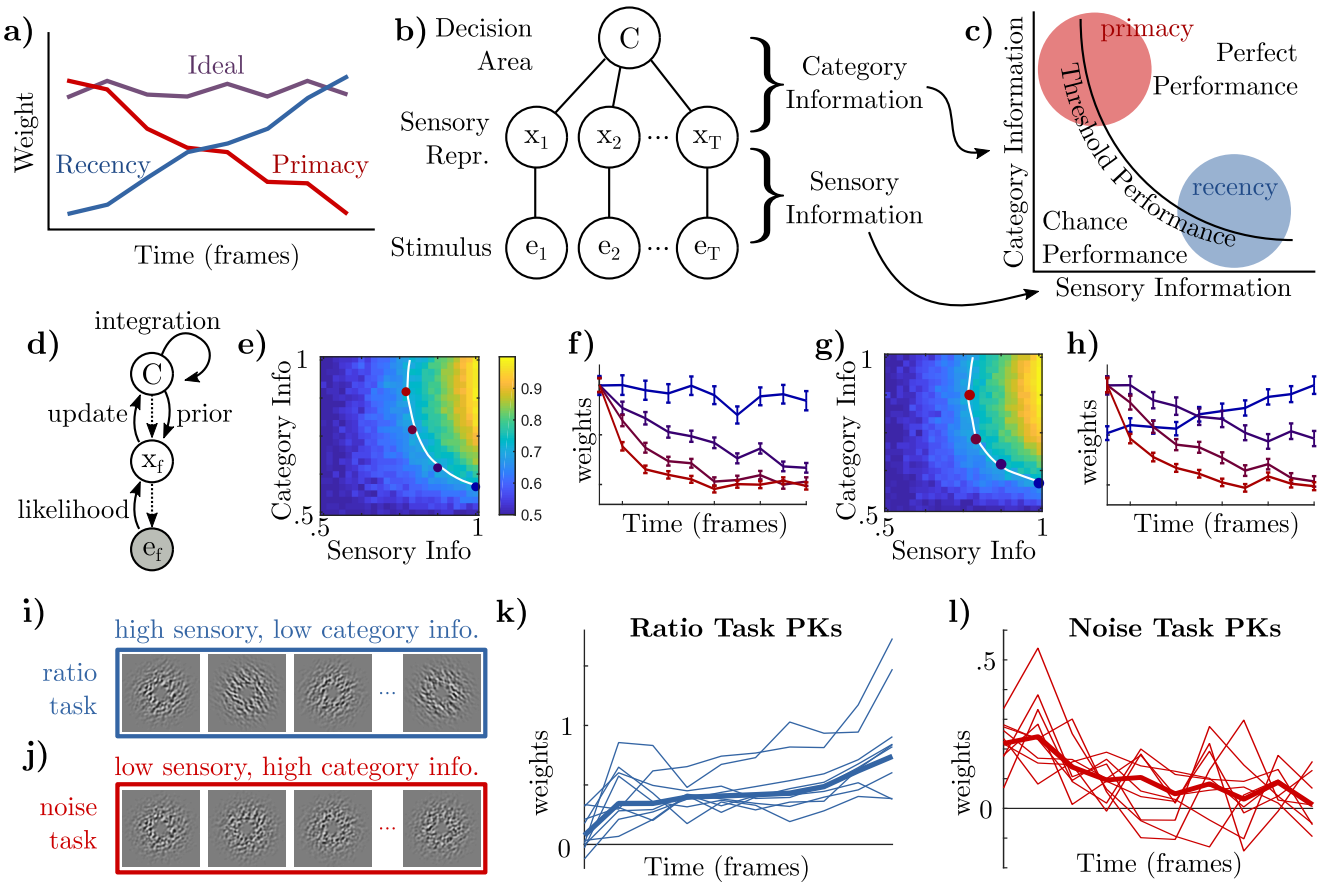

Figure 1: **a)** Possible temporal weight profiles for evidence integration tasks. **b)** The information between time-changing external evidence, e_t, and category C can be split into "sensory information" between e_t and its sensory representation in the brain, x_t, and "category information" between the sensory representation and the correct overall choice. **c)** Any given task can be thought of as a point in category-info versus sensory-info space. Previous studies finding primacy effects primarily reduce sensory information to reach threshold; previous studies finding recency effects primarily reduce category information to reach threshold. **d)** Simplified approximate inference model; dotted lines show generative model, solid lines show information flow with our assumption that a posterior over x_f is represented and evidence integration happens in a decision area representing C. Unless prior information to x_f can be completely "subtracted out" during the update of the belief about C, a positive feedback loop emerges between C and x. **e)** Sampling model performance across the entire space of category and sensory information. White line is threshold performance. Colored circles denote points used for the plots in g. **f)** (Normalized) temporal evidence weighting of the sampling model at different points in the task space, computed using regularized logistic regression. Without integration noise or a bias correction term, weight profiles vary between primacy and flat. **g-h)** Same as (e-f) but with a small "bias correction" term ($\gamma = 0.1$) leading to a recency effect in the same part of space where it is seen in the data. **Not shown:** Variational model shows same patterns as seen in (e-h). **i)** Example band-passed grating stimulus in the "ratio" task. Each frame has a clear orientation, but orientations change from frame to frame. **j)** Example band-passed grating stimulus in the "noise" task. All frames contain noisy rightward information. **k)** Temporal weighting strategy inferred for all subjects in the ratio task (thin lines) and average across subjects (thick line) show consistent recency effects. **l)** Same as (k) but for noise task; subjects consistently show primacy effect.