ARCHIVE

*Computational
Perception and
Cognition
Laboratory*

Running title:	Benefits of conditioning
Authors:	Cheng Qiu and Long Luu and Alan A. Stocker
Affiliation:	Department of Psychology and Department of Electrical and Systems Engineering University of Pennsylvania
Correspondence:	Dr. Alan A. Stocker Computational Perception and Cognition Laboratory Goddard laboratories, Rm 421 3710 Hamilton walk Philadelphia, PA 19106 U.S.A. astocker@psych.upenn.edu Phone: +1 215 573 9341
Repository:	BioRxiv
Classification:	Biological Sciences (Psychology/Neuroscience)
Statistics:	26 pages, 5(+3) figures

Benefits of commitment in hierarchical inference

Cheng Qiu¹, Long Luu², and Alan A. Stocker¹

¹Department of Psychology, University of Pennsylvania

²Zuckerman Institute, Columbia University

Abstract

Humans have the tendency to commit to a single interpretation of what has caused some observed evidence rather than considering all possible alternatives. This tendency can explain various forms of confirmation and reference biases. However, committing to a single high-level interpretation seems short-sighted and irrational, and thus it is unclear why humans seem motivated to pursue such strategy.

In a first step toward answering this question, we systematically quantified how this strategy affects estimation accuracy at the feature level in the context of two universal hierarchical inference tasks, categorical perception and causal cue combination. Using model simulations, we demonstrate that although estimation is generally impaired when conditioned on only a single high-level interpretation, the impairment is not uniform across the entire feature range. On the contrary, compared to a full inference strategy that considers all high-level interpretations, accuracy is actually better for feature values for which the probability of an incorrect categorical/structural commitment is relatively low. That is to say, if an observer is reasonably certain about the high-level interpretation of the feature, it is advantageous to condition subsequent feature inference only on that particular interpretation. We also show that this benefit of commitment is substantially amplified if late noise corrupts information processing (*e.g.*, during retention in working memory). Our results suggest that top-down inference strategies that solely rely on the most likely high-level interpretation can be favorable and at least locally outperform a full inference strategy.

Keywords: hierarchical models, top-down inference, self-consistency, model selection, optimal performance

Introduction

Cognitive tasks typically require some form of statistical inference where the brain has to infer an unknown quantity based on given yet uncertain evidence and a learned statistical (generative) model of the task (Helmholtz, 1867; Jaynes, 2003; Lee and Mumford, 2003). Previous work has shown that this notion in combination with the formalism of Bayesian statistics provides an accurate description of human behavior in a broad range of tasks associated with perception (Knill and Richards, 1996), cognitive reasoning (Griffiths et al., 2010), economic decision-making (Summerfield and Tsetsos, 2012) and motor control (Wolpert, 2007). Furthermore, mental disorders such as autism and schizophrenia have been directly linked to specific computational deficiencies of this inference process (Lieder et al., 2019; Jardri and Deneve, 2013). Except for simple estimation and decision-making tasks (Ernst and Banks, 2002; Körding and Wolpert, 2004; Stocker and Simoncelli, 2006), the generative models of these tasks are hierarchical. Object recognition is a simple example of a hierarchical inference task where, at the top of the hierarchy, object categories are defined as specific distributions over some lower-level feature representation potentially across multiple levels of feature integration. Noisy observations at the lowest feature level then allow to infer the corresponding object category by inverting the hierarchical generative model (Kersten et al., 2004). Various studies have shown that human behavior in such tasks is accurately described by Bayesian statistics that fully integrate all information in the hierarchical generative model (*i.e.*, fully marginalizes) from bottom to top. These studies include models of human judgments of sameness (Van den Berg et al., 2012), of stimulus transparency (Hedges et al., 2011), or causal stimulus structure (Körding et al., 2007).

More interesting and controversial behavior has been observed in tasks that require inference at the feature rather than the top level. In these cases, the hierarchical generative model represents a hypothesis of what caused the feature, and thus ultimately the observed evidence (see Fig. 1a). Full inference dictates that in order to infer the value of the feature, an observer should consider all different possible hypotheses (*e.g.*, categorical assignments) and weigh them according to how probable they are given the observed evidence (Fig. 1b). This strategy is known as optimal model evaluation (Draper, 1995), or Bayesian model averaging (Hoeting et al., 1999), and has been previously suggested to account for human behavior in various perceptual and cognitive reasoning tasks (Duffy et al., 2006; Körding et al., 2007; Griffiths et al., 2010; Knill, 2003, 2007).

However, recent studies have reported conditions in which human subjects seem to abandon the averaging strategy and rather commit to a specific interpretation of what caused the evidence

(model selection). For example, by making an category assignment a subject's subsequent perceptual estimate of a low-level stimulus feature (e.g., motion direction (Jazayeri and Movshon, 2007; Zamboni et al., 2016) or visual orientation (Luu and Stocker, 2018; Fritsche and de Lange, 2019)) is biased towards the assigned category on a per trial basis. These biases can be thought of as a form of consistency (Brehm, 1956) or confirmation bias (Nickerson, 1998) where the perceptual estimate aligns with and confirms the chosen category (Bronfman et al., 2015; Talluri et al., 2018). Importantly, these biases are not dependent on subjects making an explicit, overt choice of a high-level interpretation; similar biases were observed in tasks that did not require an explicit categorical choice (Wu et al., 2009; Zamboni et al., 2016; Ding et al., 2017), indicating that committing to a high-level interpretation could be an inherent inference strategy. As proposed (Stocker and Simoncelli, 2007), and refined and validated more recently (Luu and Stocker, 2018), the behavioral biases in above examples are remarkably well described by a *conditioned observer model*. The model assumes that feature inference is a sequential process: first, an observer selects the most probable hypothesis given the evidence, and then infers the feature value conditioned on the chosen hypothesis (Fig. 1c). In contrast to related models that include meta-cognitive processes with access to additional information (Fleming and Daw, 2017), the conditioned observer model assumes that the observer uses less of the available information as it discards posterior probability information by committing to only one hypothesis. Thus, the expectation is that the conditioned observer performs worse at the feature level (i.e., in terms of estimation accuracy) than an observer that integrates over all possible (high-level) hypotheses.

We set out to systematically assess how strongly performance is hampered with a conditioned inference strategy. We quantitatively explored its performance compared to the full inference strategy for two simple but universal inference examples - categorical perception and causal cue combination. Using simulations over a range of generative model parameters we characterized estimation accuracy for both strategies. We show that there are conditions where the conditioned observer outperforms the full inference observer. These benefits further increase if late noise is affecting the inference process. Our results provide a useful quantitative analysis to the discussion of different inference strategies and why humans may or may not apply one and not the other.

Inference in hierarchical generative models

We focus on the task of inferring the value of a feature θ based on uncertain sensory evidence m (measurement). We assume that θ is embedded in a probabilistic hierarchical generative model

Figure 1: Feature inference in hierarchical generative models. (a) Graphical model that represents the generic class of hierarchical models we consider. Sensory measurement m is assumed to be a noisy sample of the feature value θ , which itself is drawn from a higher-level generative hypothesis H . Based on an observed value of m we consider two inference strategies: (b) Full inference. This strategy requires the observer to marginalize over all possible high-level hypotheses when inferring θ . (c) Conditioned inference. This strategy consists of two steps. First, the observer infers and commits to the most likely hypothesis $\hat{H}$ based on the sensory measurement m . Subsequently, the observer infers the feature value conditioned on the committed hypothesis $\hat{H}$, rather than marginalizing over all possible hypotheses.

that can be expressed with a directed graph as shown in Fig. 1a. The hierarchical component of the model “above θ ” represents a high-level generative hypothesis H about the potential values of θ . A simple example for such a hypothesis is the association of θ with a particular category. More elaborate high-level hypotheses may include different structural assumptions (*i.e.*, different graphs) that capture different contextual or causal dependencies (Körding et al., 2007; Kemp and Tenenbaum, 2008; Battaglia et al., 2013). However, our analysis is agnostic to the specific form of H as it only assumes that the feature θ is at the “bottom” of the hierarchy, and that sensory evidence m only directly depends on θ and not the rest of the hierarchy.

We consider two different inference strategies. The first strategy is to make use of the full posterior probability over hypotheses when inferring the feature (here referred to as “full inference”), that is, to marginalize over all possible high-level hypotheses (Fig. 1b). Thus, we can express the posterior over θ as a weighted sum

$$p(\theta|m) \propto \sum_i p(\theta|m, H_i) p(H_i|m) \quad (\text{full inference}) , \quad (1)$$

with weights given by the posterior probability of each hypothesis given the evidence, $p(H_i|m)$.

The second, alternative strategy represents a sequential inference process (Stocker and Simoncelli, 2007; Luu and Stocker, 2018): First, a hypothesis $\hat{H}$ is selected according to the posterior probability $p(H|m)$ (here, the hypothesis with maximal posterior probability), and then the posterior over θ is computed conditioned on the observed evidence m and the chosen hypothesis $\hat{H}$ (Fig. 1c). The chosen hypothesis imposes a conditioned prior $p(\theta|\hat{H})$ that, unlike in the full inference strategy, depends on the sensory evidence and thus is potentially different for each m (*i.e.*, on each trial). Accordingly, the posterior probability for this second strategy can be written as

$$p(\theta|m, \hat{H}) \propto p(m|\theta) p(\theta|\hat{H}) \quad (\text{conditioned inference}) . \quad (2)$$

In the following we compare the performance of these two strategies in estimating θ . With posterior probabilities Eqs. (1) and (2) and assuming a quadratic loss function (L_2 -norm), we derive optimal estimators for the feature value $\hat{\theta}(m)$ under each inference strategy, and compare their relative expected estimation error for the chosen loss function (see Appendix for results with other loss functions). We apply this error analysis to two well-known examples of hierarchical inference, categorical perception and causal cue combination systematically investigate the relative performance of the two strategies for different levels of sensory uncertainty and additional (late) processing noise.

Example 1: Categorical perception

With *categorical perception* we refer to the task of estimating the value of a low-level (scalar) feature that is associated with multiple high-level categories¹. A broad range of well-studied perceptual tasks and experiments represent forms of categorical perception. Categorical perception assumes the simplest hierarchical generative model where the hypothesis H is represented by a single node C reflecting the categorical assignment of feature θ (Fig. 2a). The generative process first involves the selection of a category C based on a categorical prior $p(C)$; for simplicity, we consider here only two possible categories $C \in \{C_1, C_2\}$. A feature value θ is then sampled from the categorical feature prior $p(\theta|C)$. Finally, sensory evidence m is sampled from the conditional probability $p(m|\theta)$. Also, we explicitly allow the possibility that late noise may deteriorate sensory evidence m due to e.g. retention in working memory. We refer to the deteriorated sensory evidence as m^* distributed according to $p(m^*|m)$. The specific description of the priors and conditional probabilities is provided in Fig. 2 and its caption.

Estimate distributions

Given the generative model, we can now express an optimal estimate $\hat{\theta}(m)$ of the feature value for both inference strategies. The *full inference* strategy (Fig. 1b) marginalizes over all possible categories, resulting in the posterior distribution

$$p(\theta|m) \propto p(m|\theta) \sum_i p(\theta|C_i)p(C_i) = p(m|\theta)p(\theta) . \quad (3)$$

Minimizing mean squared-error (i.e., minimizing L_2 loss) we find the optimal estimator according to the full inference strategy as

$$\hat{\theta}_f(m) = \int_{\theta} \theta p(\theta|m) . \quad (4)$$

In contrast, the *conditioned inference* strategy consists of two inference steps (Fig. 1c). First, the most probable category is chosen based on the sensory evidence m according to

$$\hat{C}(m) = \operatorname{argmax}_C p(C|m) , \quad (5)$$

where the posterior is defined as $p(C|m) \propto p(C) \int_{\theta} p(m|\theta)p(\theta|C)$. Second, the posterior over θ is computed conditioned on the chosen category $\hat{C}(m)$, thus

$$p(\theta|m, \hat{C}(m)) \propto p(m|\theta)p(\theta|\hat{C}(m)) . \quad (6)$$

¹This in contrast to *categorization*, which is the task of inferring the category based on evidence about the feature value.

Figure 2: Hierarchical generative model of categorical perception, and the predicted estimate distributions for both inference strategies. (a) Generative model. For reasons of simplicity we make the following assumptions: Feature value θ belongs to one of two categories $C \in \{C_1, C_2\}$ with balanced prior probabilities $p(C_{1,2}) = 0.5$; Categorical feature priors $p(\theta | C_i)$ are assumed to be Gaussians with identical standard deviations but different means, $\mathcal{N}(\mu_C = \pm 3, \sigma_C = 3)$; Sensory evidence m is a noisy sample of θ according to additive Gaussian noise with standard deviation σ_s . Late additive Gaussian noise (σ_m) may further deteriorate the sensory measurement m leading to m^* . (b) Estimate distributions according to either full (left) or conditioned inference (right). In each panel, vertical cross-sections represent the estimate distribution $p(\hat{\theta} | \theta)$. Distributions show a characteristic bimodal pattern across the category boundary for the conditioned inference strategy. Top panels show distributions without late noise ($\sigma_s = 2, \sigma_m = 0$); bottom panels show distributions with late noise ($\sigma_s = 2, \sigma_m = 3$). Units of the feature are arbitrary throughout the paper as our analysis is not limited to a specific feature.

Finally, the optimal estimator is given as

$$\hat{\theta}_c(m, \hat{C}(m)) = \int_{\theta} \theta p(\theta|m, \hat{C}(m)) . \quad (7)$$

Note, that for both strategies the optimal estimator represents a monotonic mapping between the evidence and the estimate. Thus, we can obtain the estimate distributions $p(\hat{\theta}_f|\theta)$ and $p(\hat{\theta}_c|\theta)$ with a change of variable for the measurement distribution $p(m|\theta)$, replacing m with the estimate $\hat{\theta}_f(m)$ and $\hat{\theta}_c(m)$ according to Eqs. (4) and (7), respectively. Figure 2b shows the resulting distributions for both strategies given the specific parameter settings of our generative model. Note that the estimate distributions fundamentally differ; the conditioned inference strategy exhibits a characteristic bimodal distribution for θ values close to the category boundary, which matches a range of experimental results (Jazayeri and Movshon, 2007; Zamboni et al., 2016; Luu and Stocker, 2018).

Consider the possibility that late noise may further deteriorate sensory evidence m (e.g. due to retention in working memory (Schneegans and Bays, 2018)) we update the formulations of the optimal estimators accordingly. Assuming that late noise is additive Gaussian (with σ_m), we compute the optimal estimate and the estimate distribution for the full inference strategy as discussed above (Eqs. (3) and (4)) but replacing $p(m|\theta)$ with $p(m^*|\theta) = \int_m p(m^*|m)p(m|\theta)$ where m^* represents the corrupted sensory evidence. For the conditioned inference strategy, we assume that late noise affects feature estimation only but not the selection of the category in step I. Thus, the computation of $\hat{C}$ is as above (Eq. (5)). The optimal feature estimator, however, changes to

$$\hat{\theta}_c(m^*, \hat{C}) = \int_{\theta} \theta p(\theta|m^*, \hat{C}) , \quad (8)$$

where $p(\theta|m^*, \hat{C}) \propto p(m^*|\theta)p(\theta|\hat{C})$. By a change of variable in $p(m^*|\theta, \hat{C})$ substituting m^* with the estimate $\hat{\theta}(m^*, \hat{C})$, we find $p(\hat{\theta}|\theta, \hat{C})$. Then the estimate distributions are obtained as

$$p(\hat{\theta}|\theta) = \sum_{\hat{C}} p(\hat{\theta}|\theta, \hat{C})p(\hat{C}|\theta), \quad (9)$$

where $p(\hat{C}|\theta) = \int_m p(\hat{C}|m)p(m|\theta)$. Figure 2b shows the estimate distributions given the feature value for full and conditioned inference for moderate level of late noise.

Relative accuracy of full vs. conditioned inference strategy

Having defined the optimal estimate distributions for both the full and the conditioned inference strategy, we next perform a systematic quantitative comparison between the estimation accuracy for various model parameters and categorical priors. We defined relative accuracy as the

ratio between the expected estimation errors for both inference strategies. Because the optimal estimators were derived with regard to a quadratic loss function (L_2 -norm), we accordingly defined the expected estimation error as the mean squared-error (MSE). We computed relative accuracy both *locally*, *i.e.*, for both strategies we computed the MSE for each θ separately as $\text{MSE}(\theta) = \int_{\hat{\theta}} p(\hat{\theta}|\theta)(\hat{\theta} - \theta)^2$ and then took the ratio, as well as *globally*, *i.e.*, we marginalized the local error over the total distribution of feature values $\text{MSE} = \int_{\theta} \text{MSE}(\theta)p(\theta)$ and then took the ratio.

We considered two general conditions of categorical feature priors. First, we assumed the priors $p(\theta|C_{1,2})$ to be overlapping Gaussian distributions (Fig. 3a). Local relative accuracy with and without late noise is shown in Fig. 3b. Both curves show a similar, characteristic shape. For θ values close to the category boundary, the full inference strategy provides better estimates. However, for θ values that are farther away than a certain distance from the category boundary, the conditioned inference strategy consistently outperforms the full inference strategy. This came as a surprise. The advantage of the conditioned inference strategy is amplified if we further assume that late noise (σ_m) corrupts the inference process. The distance from the category boundary at which the conditioned inference strategy starts to provide superior estimates corresponds to a certain probability level of making the correct categorical assignment $\hat{C}$, *i.e.*, a certain level of the psychometric function shown in Fig. 3c. It demonstrates that for conditions where the observer is reasonably certain of the correct feature category (*e.g.*, a probability correct of $p = 0.92$ if there is no late noise) the conditioned inference strategy outperforms the full inference strategy. As expected, global relative accuracy for the conditioned inference strategy is always inferior to the full inference strategy if there is no late noise (Fig. 3d). This is because the benefits of the full inference strategy for θ values around the category boundary make up for the deficits for θ values outside of that range. The situation changes, however, if substantial late noise impacts the inference process. In this case the conditioned strategy can also globally outperform the full inference strategy (Fig. 3d).

We obtain similar results for non-overlapping categorical feature priors (Fig. 3e). This corresponds to many of the estimation tasks reported in previous studies (Jazayeri and Movshon, 2007; Zamboni et al., 2016; Luu and Stocker, 2018). Local relative accuracy shows very similar curves (Fig. 3f), indicating the supremacy of the conditioned inference strategy for θ values that correspond to similar certainty levels in the categorical assignment (Fig. 3g). Also the global accuracy comparison shows similar behavior (Fig. 3h). While we limited our accuracy analysis in the main paper for the case of minimizing squared-error (*i.e.*, the L_2 norm), we explored other commonly

Figure 3: Relative estimation accuracy for conditioned and full inference in categorical perception. (a) Overlapping categorical feature priors ($\mu_C = \pm 3, \sigma_C = 3$). The categorical prior $p(C)$ is symmetric. (b) Relative accuracy (MSE(θ) ratio) for the two strategies as a function of θ with sensory noise $\sigma_c = 2$ and late noise either absent (light color) or present ($\sigma_m = 3$, dark color). The distance from the category boundary at which the conditioned inference strategy starts to provide superior estimates corresponds to a probability of $p = 0.92$ for the model observer making a categorical assignment $\hat{C}$ ($p = 0.85$ if late noise is present; dashed lines - see also (c)). (c) Psychometric curve of the model observer indicating the probability of assigning θ to category C_1 . (d) Global MSE ratio as a function of σ_s and σ_m . Blue shades represent conditions for which conditioned inference outperforms full inference, red shades indicate the opposite. The black curve labels the conditions for which both strategies lead to equal performance. The two dots correspond to the two conditions shown in (b). (e-h) Prior distributions, psychometric curve, local and global relative accuracy when the two categorical feature priors are box-shaped and non-overlapping.

used error metrics as well. We found that the above results of our analysis qualitatively generalize (see Appendix).

Example 2: Causal cue combination

Our second example is often referred to as causal cue combination (Körding et al., 2007). When tasked to estimate the unknown value of a feature, human observers correctly combine different sensory cues if the cues are in sufficient agreement with the interpretation that they originate from the same feature value (Jacobs, 1999; Ernst and Banks, 2002; Alais and Burr, 2004; Fetsch et al., 2009; Butler et al., 2010). However, human observers do not integrate cues that are inconsistent with such interpretation (cue conflict) and thus signal different underlying feature values (Roach et al., 2006; Wallace et al., 2004).

This integration versus segregation distinction can be modeled within the generative hierarchical framework we discuss here (Fig. 1a). However, different to the categorical perception example, the hypotheses now represent two different causal structures (Fig. 4a): Either the cues $m_{1,2}$ represent evidence of a single feature θ and thus should be combined (common cause hypothesis), or they represent two different features $\theta_{1,2}$, in which case they should be treated independently (independent causes hypothesis). The full inference strategy for estimating the feature values considers both structural hypotheses $S \in \{S_1, S_2\}$ according to their posterior probabilities when inferring the posterior density over the features $p(\theta_1, \theta_2 | m_1, m_2)$. In contrast, the conditioned inference strategy first commits to an interpretation $\hat{S}$ of the most probable structural hypothesis based on the observed sensory evidence, and then computes the posterior density over the feature values conditioned only on the chosen structure, thus $p(\theta_1, \theta_2 | m_1, m_2, \hat{S})$.

Feature estimation

According to the hierarchical generative model shown in Fig. 4a we can define the optimal estimator under each inference strategy. For $S = S_1$ (common cause), the corresponding posterior distribution at the feature level is

$$p(\theta_{1,2} = \theta | m_1, m_2, S_1) \propto p(m_1, m_2 | \theta) p(\theta | S_1) . \quad (10)$$

For $S = S_2$ (independent causes), the posterior distribution changes to

$$p(\theta_1, \theta_2 | m_1, m_2, S_2) \propto p(m_1 | \theta_1) p(m_2 | \theta_2) p(\theta_1, \theta_2 | S_2) . \quad (11)$$

Figure 4: Causal cue combination. (a) Graphical model reflecting two possible generative structures (hypotheses): the observed sensory measurements $m_{1,2}$ may be generated by a common source (S_1) or two independent sources (S_2). For reasons of simplicity we assume the structure prior $p(S)$ to be equal, and the prior over feature values to be normally distributed: $\mathcal{N}(0, \sigma_p = 4)$ for S_1 , and $\mathcal{N}(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 16 & 0 \\ 0 & 16 \end{bmatrix})$ for S_2 . Furthermore, we also assume the observed sensory measurements to be independent samples drawn from Gaussian distributions $\mathcal{N}(\theta_{1,2}, \sigma_s)$. (b) Estimate distributions for the full and the conditioned inference strategy, either without (top row) or with late noise (bottom row; Gaussian $\mathcal{N}(m, \sigma_m)$). Each vertical cross-section represents the estimate density $p(\hat{\theta}_1 | \theta_1, \theta_2 = 0)$ with $\sigma_s = 2$ and late noise σ_m either 0 or 3. (c) Difference of the estimate distributions in (b). For θ_1 values substantially different from $\theta_2 = 0$, conditioned inference produces more veridical estimates than full inference.

219 With a full inference strategy, we compute the total posterior as the average posterior under both
220 hypotheses (Eqs. (10) and (11)) weighted by the posterior probability of each hypothesis, thus

$$p(\theta_1, \theta_2 | m_1, m_2) = \sum_i p(\theta_1, \theta_2 | m_1, m_2, S_i) p(S_i | m_1, m_2) . \quad (12)$$

221 The posterior $p(S | m_1, m_2)$ is different for the two hypothesis, namely

$$p(S = S_1 | m_1, m_2) \propto \int_{\theta} p(m_1, m_2 | \theta) p(\theta | S_1) p(S_1) \quad (13)$$

222 and

$$p(S = S_2 | m_1, m_2) \propto \int_{\theta_1} \int_{\theta_2} p(m_1 | \theta_1) p(m_2 | \theta_2) p(\theta_1, \theta_2 | S_2) p(S_2) . \quad (14)$$

223 As in our first example, we define the optimal estimate of the feature vector $[\theta_1, \theta_2]$ with regard
224 to the L_2 -norm which corresponds to the expectation over the total posterior Eq. (12) (mean of
225 posterior), thus

$$(\hat{\theta}_1, \hat{\theta}_2) = E[\theta_1, \theta_2 | m_1, m_2] . \quad (15)$$

226 With the conditioned inference strategy we first infer the most probable causal structure,

$$\hat{S} = \operatorname{argmax}_S p(S | m_1, m_2), \quad (16)$$

227 with $p(S | m_1, m_2)$ as defined above (Eqs. (13) and (14)). The optimal estimate of the feature vector
228 is again the expectation over the posterior distributions; however, now conditioned on $\hat{S}$ as well,
229 thus

$$(\hat{\theta}_1, \hat{\theta}_2) = E[\theta_1, \theta_2 | m_1, m_2, \hat{S}] , \quad (17)$$

230 where the posterior is given as in Eqs. (10) and (11), according to $\hat{S}$.

231 For both strategies, the estimate distributions, $p(\hat{\theta}_1, \hat{\theta}_2 | \theta_1, \theta_2)$, are obtained by marginalizing over
232 the measurement distributions $p(m_1, m_2 | \theta_1, \theta_2)$ (change of variable). Figure 4b shows the estimate
233 distributions for $\hat{\theta}_1$ for both strategies with setting $\theta_2 = 0$, equal structural priors $p(S = S_{1,2}) = 0.5$,
234 and Gaussian distributed feature values $\theta_{1,2}$ and measurement distributions (see figure/caption for
235 details). Distributions with and without late noise are shown where late noise was again assumed
236 to be additive, independent Gaussian noise.

237 Unlike in the categorical perception example, the resulting estimate distributions for both inference
238 strategies are very similar given the chosen model parameters. Plotting the difference in distribu-
239 tion density (Fig. 4c), however, shows that the estimates are more veridical for the full inference
240 strategy for values of θ_1 close to θ_2 , whereas the conditioned strategy leads to more veridical
241 estimates for conditions where there is a substantial cue conflict, *i.e.*, $\theta_1 \neq \theta_2$.

Comparing estimation accuracy of full vs. conditioned inference strategy

Similar to the previous example, we analyzed the relative estimation accuracy of the two inference strategies by computing both the local $\text{MSE}(\theta_1, \theta_2)$ as well as the global MSE (integral over all $\theta_{1,2}$). Results are shown in Fig. 5. For feature value pairs (θ_1, θ_2) that support both the common and the independent source hypothesis, full inference outperforms conditioned inference (Fig. 5c). For feature values, however, that clearly favor one structural interpretation over the other, the conditioned inference strategy is beneficial. The relative accuracy comparison is similar to the categorical perception example as shown in Fig. 5d. Again, in situations where observers are reasonably certain about the high-level interpretation, conditioned inference proves to be the better strategy. This advantage is further amplified if late noise affects feature inference (Fig. 5e).

Discussion

Since first proposed (Stocker and Simoncelli, 2007), data from a growing number of psychophysical studies suggest that humans perform conditioned inference in certain situations (Jazayeri and Movshon, 2007; Zamboni et al., 2016; Wu et al., 2009; Ding et al., 2017; Fritsche and de Lange, 2019; Luu and Stocker, 2018); *i.e.*, they commit to a single high-level interpretation rather than taking into account all possible high-level interpretations when performing low-level feature inference. This also matches recent results showing that post-decision confidence reports overemphasize information supporting a decision (Peters et al., 2017). In this paper, we presented an extensive quantitative analysis of how a conditioned inference strategy affects the accuracy with which an observer is able to estimate low-level features. Using model simulations, we show that considering all possible high-level interpretations does not necessarily provide better accuracy across the entire feature range. Specifically, under conditions where the observer is reasonably certain about the high-level interpretation, committing to a single interpretation is better. Although our conclusions must necessarily remain limited to the specific examples we have analyzed, categorical perception and causal cue combination, the results for these two well-known yet quite different inference problems are qualitatively very similar and robust to variations in model parameters (such as *e.g.* different prior distributions - see Appendix). This suggests that these *benefits of commitment* may generally apply to the type of hierarchical problems in perception and cognition we consider here (Fig. 1a).

We also demonstrated that a conditioned inference strategy becomes increasingly favorable when

Figure 5: Relative performance between conditioned and full inference for causal cue combination. (a) Probability of inferring a structure with independent sources (S_2) as a function of feature values. When θ_1 and θ_2 are far apart, they most likely represent independent sources, whereas when they are close, a common source is more plausible. (b) Cross-section through the probability density surface in (a) along the off-diagonal (dashed line). (c) Relative performance in estimating θ_1 and θ_2 (ratio of net MSE) for sensory noise $\sigma_s = 2$ and no ($\sigma_m = 0$) or moderate late noise ($\sigma_m = 3$). Blue indicates feature values for which conditioned inference outperforms full inference. These values correspond to a regime where decisions about $\hat{S}$ are relatively certain as shown in (a). Red indicates feature values for which full inference is better. With late noise, the benefits of conditioned inference are amplified. (d) Local relative MSE along the off-diagonal $\theta_1 = -\theta_2$ (dashed lines in (c)). Conditioned inference outperforms full inference when the observer is relatively certain that θ_1 and θ_2 come from independent sources (dashed lines in (b)). (e) Global MSE ratio as a function of sensory (σ_s) and late noise (σ_m).

late noise is affecting the inference process, to a point where it can globally outperform a full inference strategy. This has implications for models and theories of working memory formation and recall, and in general, for inference tasks that evolve over time (Gold and Stocker, 2017). Specifically, it suggests that the commitment to a high-level interpretation can be seen as creating a robust summary representation of sensory evidence, which then, at the time of recall, is used in order to improve inference accuracy. In fact, conditioned inference can predict some of the memory biases that have been reported in the recall of color (Bae et al., 2015) or other low-level visual features (Ding et al., 2017; Luu et al., 2017). It provides a normative motivation for memory biases and potentially other similar forms of confirmation biases (Talluri et al., 2018; Lange et al., 2019).

Our results also have theoretical implications. While a full inference strategy provides globally better estimation accuracy, a conditioned inference strategy is more accurate in situations where the observer can be reasonably certain that the high-level interpretation is correct. While this certainty level does not correspond to a fixed probability value and depends on the specifics of the generative model, we found that it lies within a relative limited range for a given generative model (see dashed lines in Fig. 3). This suggests that a mixed inference strategy that uses either conditioned or full inference depending on the certainty in the high-level interpretation could globally outperform full inference under any condition. It would require a meta-cognitive mechanism that monitors and evaluates the observer's certainty in all possible high-level hypotheses. We are not aware of any previous proposal of such a “supra-optimal” mixed inference strategy. However, it would provide a quantitative argument for why meta-cognitive processes that monitor decision confidence may be useful. Future work will show if human inference behavior shows any evidence for such a mixed strategy. Our results also provide as a useful quantitative analysis for the ongoing general debate about optimality/suboptimality in human perceptual and cognitive behavior (Rahnev and Denison, 2018; Stocker, 2019).

An obvious but willful limitation of our study is that we only analyzed accuracy at the feature level (*i.e.*, a loss function only including θ). However, cognition involves simultaneous and co-joint inference processes at all hierarchical levels that together form a holistic representation. Thus, it is likely that human inference strategies have evolved with regard to error metrics that are defined at the level of such holistic representations. For example, in the case of categorical perception (see Fig. 2a), the error metric would not only include errors at the feature but simultaneously also at the category level. Furthermore, one can easily make the case that such metric includes certain measures of consistency across the hierarchy; errors at the feature level would be irrelevant if at

the same time the category assignment of the feature is already wrong (“why bother about error at the low level if inference is wrong at the top level?”). It is easy to show that conditioned inference globally outperforms full inference if the analysis is limited to those trials where the observer is correct at the category level: As both strategies have equal performance in getting the category right, conditioned inference is always beneficial because it uses the more specific feature prior $p(\theta|\hat{C})$ where $\hat{C}$ is correct. Also, we have recently shown how conditioned inference avoids inconsistencies of the representations across the hierarchy under conditions of late noise (Luu and Stocker, 2018), which may be a mechanism to avoid states of cognitive dissonance (Brehm, 1956; Festinger, 1957).

Our comparison of the inference strategies did not consider potential differences in computational and resource costs. There is growing evidence that cognitive and perceptual inference processes are optimized also with regard to such costs, leading to inference strategies that are commonly referred to as bounded rationality (Simon, 1984; Gershman et al., 2015). Intuitively, conditioned inference is a simpler and computationally less costly strategy because it performs feature inference only for one high-level hypothesis. Future work will be needed to investigate in more detail the benefits (or drawbacks) of a conditioned inference strategy with regard to hierarchical error metrics and other costs such as computational complexity.

Finally, given that many inference problems are of the general hierarchical type considered here, conditioned inference strategies may be more ubiquitous than we recognize at the moment. The problem is that conditioned inference not always leads to clearly identifiable behavioral signatures such as the bimodal estimate distributions in categorical perception (Fig. 2b). In some cases behavioral effects may be small and difficult to extract from experimental data (Fig. 4b,c). Potential candidates for conditioned inference strategies are context-dependent perceptual tasks where different contextual interpretations represent the different high-level hypotheses. Examples include tilt estimation of textured surfaces with competitive priors (Knill, 2003), orientation estimation affected by center-surround integration or segmentation (Schwartz et al., 2009; Coen-Cagli et al., 2015; Qiu et al., 2013), lightness perception affected by perceived surface curvature (Knill and Kersten, 1991), or the perceived brightness of a gray patch depending on its spatial context (Adelson, 1993). A detailed quantitative modeling approach will allow us to determine the extent to which conditioned inference is a ubiquitous and general inference mechanism of the human brain.

Acknowledgments

We thank current and past members of the Computational Perception and Cognition laboratory for their various forms of feedback and inspirations. Some of the ideas have been described in a paper posted on BioRxiv (Luu et al., 2017). This work has been supported in part by the National Science Foundation of the United States of America (CAREER award 1350786 to AAS), and in part by the University of Pennsylvania.

References

- E. Adelson. Perceptual organization and the judgement of brightness. *Science*, 262:2042–2044, December 1993.
- D. Alais and D. Burr. The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14:257–262, February 2004.
- G.-Y. Bae, M. Olkkonen, S. R. Allred, and J. Flombaum. Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, 144(4):744–763, 2015.
- P. Battaglia, J. Hamrick, and J. Tenenbaum. Simulation as an engine of physical scene understanding. *Proceeding of the National Academy of Sciences of the U.S.A.*, 2013. doi: 10.1073/pnas.1306572110.
- J. Brehm. Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology*, 52(3):384ff., 1956.
- Z. Z. Bronfman, N. Brezis, R. Moran, K. Tsetsos, T. Donner, and M. Usher. Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society of London B: Biological Sciences*, 282(1810), 2015. doi: 10.1098/rspb.2015.0228.
- J. S. Butler, S. T. Smith, J. L. Campos, and H. H. Bülthoff. Bayesian integration of visual and vestibular signals for heading. *Journal of Vision*, 10(11):23–23, 09 2010. doi: 10.1167/10.11.23.
- R. Coen-Cagli, A. Kohn, and O. Schwartz. Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, 18:1648 EP, 10 2015.

- 361 S. Ding, C. Cueva, M. Tsodyks, and N. Qian. Visual perception as retrospective bayesian decoding
362 from high- to low-level features. *Proc. National Academies of Sciences U.S.A.*, Early edition,
363 October 2017.
- 364 D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistical*
365 *Society B*, 57:45–97, 1995.
- 366 S. Duffy, J. Huttenlocher, and L. Crawford. Children use categories to maximize accuracy in
367 estimation. *Developmental Science*, 9(6):597–603, 2006.
- 368 M. Ernst and M. Banks. Humans integrate visual and haptic information in a statistically optimal
369 fashion. *Nature*, 415:429ff, January 2002.
- 370 L. Festinger. *Theory of Cognitive Dissonance*. Stanford University Press, Stanford, CA, 1957.
- 371 C. R. Fetsch, A. H. Turner, G. C. DeAngelis, and D. E. Angelaki. Dynamic reweighting of visual and
372 vestibular cues during self-motion perception. *Journal of Neuroscience*, 29(49):15601–15612,
373 2009. doi: 10.1523/JNEUROSCI.2574-09.2009.
- 374 S. Fleming and N. Daw. Self-evaluation of decision-making: A general Bayesian framework for
375 metacognitive computation. *Psychological Review*, 124(1):91–114, 2017.
- 376 M. Fritsche and F. P. de Lange. Reference repulsion is not a perceptual illusion. *Cognition*, 184:
377 107 – 118, 2019. doi: <https://doi.org/10.1016/j.cognition.2018.12.010>.
- 378 S. Gershman, E. Horvitz, and J. Tenenbaum. Computational rationality: A converging paradigm for
379 intelligence in brains, minds, and machines. *Science*, 349(6245):273ff, 2015.
- 380 J. I. Gold and A. A. Stocker. Visual decision-making in an uncertain and dynamic world. *Annual*
381 *Review of Vision Science*, 3(1):227–250, 2017. doi: 10.1146/annurev-vision-111815-114511.
- 382 T. Griffiths, N. Chater, C. Kemp, A. Perfors, and J. Tenenbaum. Probabilistic models of cognition:
383 exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364,
384 2010.
- 385 J. Hedges, A. Stocker, and E. Simoncelli. Optimal inference explains the perceptual coherence of
386 visual motion stimuli. *Journal of Vision*, 11(6):1–16, May 2011.
- 387 H. v. Helmholtz. *Handbuch der Physiologischen Optik*. Allg. Enzyklopädie der Physik 9. Bd. Voss,
388 Leipzig, Germany, 1867.

- 389 J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.
- 391 R. Jacobs. Optimal integration of texture and motion cues to depth. *Vision Research*, 39:3621–3629, 1999.
- 393 R. Jardri and S. Deneve. Circular inferences in schizophrenia. *Brain*, 136(11):3227–3241, 2013. doi: doi: 10.1093/brain/awt257.
- 395 E. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.
- 396 M. Jazayeri and J. Movshon. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature*, 446:912ff, April 2007.
- 398 C. Kemp and J. Tenenbaum. The discovery of structural form. *Proc. National Academies of Sciences U.S.A.*, 105(31):10687–10692, 2008.
- 400 D. Kersten, P. Mamassian, and A. Yuille. Object perception as Bayesian inference. *Annual Review of Psychology*, 55:271–304, 2004.
- 402 D. Knill. Mixture models and the probabilistic structure of depth cues. *Vision Research*, 43:831–854, 2003.
- 404 D. Knill. Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision*, 7(7):1–24, May 2007.
- 406 D. Knill and W. Richards, editors. *Perception as Bayesian Inference*. Cambridge University Press, 1996.
- 408 D. C. Knill and D. Kersten. Apparent surface curvature affects lightness perception. *Nature*, 351(6323):228–230, 1991. doi: 10.1038/351228a0.
- 410 K. Körding and D. Wolpert. Bayesian integration in sensorimotor learning. *Nature*, 427(15):244–247, January 2004.
- 412 K. P. Körding, U. Beierholm, W. J. Ma, S. Quartz, J. B. Tenenbaum, and L. Shams. Causal inference in multisensory perception. *PLoS One*, 2(9), 2007.
- 414 R. D. Lange, A. Chattoraj, J. M. Beck, J. L. Yates, and R. M. Haefner. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *bioRxiv*, 2019. doi: 10.1101/440321.

- 417 T. Lee and D. Mumford. Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A*,
418 20(7):1434–1448, 2003.
- 419 I. Lieder, V. Adam, O. Frenkel, S. Jaffe-Dax, M. Sahani, and M. Ahissar. Perceptual bias reveals
420 slow-updating in autism and fast-forgetting in dyslexia. *Nature Neuroscience*, 22(2):256–264,
421 2019. doi: 10.1038/s41593-018-0308-9.
- 422 L. Luu and A. Stocker. Post-decision biases reveal a self-consistency principle in perceptual infer-
423 ence. *eLife*, 7:e333334, May 2018. doi: 10.7554/eLife.33334.
- 424 L. Luu, C. Qiu, and A. A. Stocker. High- to low-level decoding does not generally improve percep-
425 tual performance. *bioRxiv*, (240390), December 2017.
- 426 R. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General*
427 *Psychology*, 2(2):175–220, 1998.
- 428 M. Peters, T. Thesen, Y. Ko, B. Maniscalco, C. Carlsón, M. Davidson, W. Doyle, R. Kuzniecky,
429 O. Devinsky, E. Halgren, and H. Lau. Perceptual confidence neglects decision-incongruent
430 evidence in the brain. *Nature Human Behaviour*, 1:1–8, 2017.
- 431 C. Qiu, D. Kersten, and C. A. Olman. Segmentation decreases the magnitude of the tilt illusion.
432 *Journal of Vision*, 13(13):1–19, 11 2013. doi: 10.1167/13.13.19.
- 433 D. Rahnev and R. N. Denison. Suboptimality in perceptual decision making. *Behavioral and Brain*
434 *Sciences*, pages 1–107, 2018. doi: DOI: 10.1017/S0140525X18000936.
- 435 N. W. Roach, J. Heron, and P. V. McGraw. Resolving multisensory conflict: a strategy for balancing
436 the costs and benefits of audio-visual integration. *Proceedings of the Royal Society of London*
437 *B: Biological Sciences*, 273(1598):2159–2168, 2006. doi: 10.1098/rspb.2006.3578.
- 438 S. Schneegans and P. M. Bays. Drift in neural population activity causes working mem-
439 ory to deteriorate over time. *Journal of Neuroscience*, 38(21):4859–4869, 2018. doi:
440 10.1523/JNEUROSCI.3440-17.2018.
- 441 O. Schwartz, T. J. Sejnowski, and P. Dayan. Perceptual organization in the tilt illusion. *Journal of*
442 *Vision*, 9(4):19–19, 04 2009. doi: 10.1167/9.4.19.
- 443 H. Simon. *Models of Bounded Rationality*. MIT Press, Cambridge, MA, 1984.
- 444 A. Stocker and E. Simoncelli. Noise characteristics and prior expectations in human visual speed
445 perception. *Nature Neuroscience*, pages 578–585, April 2006.

- 446 A. Stocker and E. Simoncelli. A Bayesian model of conditioned perception. In J. Platt, D. Koller,
447 Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems NIPS*
448 *20*, pages 1409–1416, Cambridge, MA, December 2007. MIT Press.
- 449 A. A. Stocker. Credo for optimality. *Behavioral and Brain Sciences*, 41:38–39, 2019. Invited
450 commentary.
- 451 C. Summerfield and K. Tsetsos. Building bridges between perceptual and economic decision-
452 making: Neural and computational mechanisms. *Frontiers in Neuroscience*, 6:70, 2012. doi:
453 10.3389/fnins.2012.00070.
- 454 B. C. Talluri, A. E. Urai, K. Tsetsos, M. Usher, and T. H. Donner. Confirmation bias through
455 selective overweighting of choice-consistent evidence. *Current Biology*, 28:1–8, October 2018.
456 doi: 10.1016/j.cub.2018.07.052.
- 457 R. Van den Berg, M. Vogel, K. Josic, and W. Ma. Optimal inference of sameness. *Proc. National*
458 *Academies of Sciences U.S.A.*, 109(8):3178–3183, 2012.
- 459 M. T. Wallace, G. E. Roberson, W. D. Hairston, B. E. Stein, J. W. Vaughan, and J. A. Schirillo.
460 Unifying multisensory signals across time and space. *Exp Brain Res*, 158(2):252–258, Sep
461 2004.
- 462 D. Wolpert. Probabilistic models in human sensorimotor control. *Human movement science*, 26
463 (4):511–524, 2007. doi: 10.1016/j.humov.2007.05.005.
- 464 S. Wu, H. Lu, and A. Yuille. Model selection and velocity estimation using novel priors for motion
465 patterns. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural*
466 *Information Processing Systems 21*, pages 1793–1800. 2009.
- 467 E. Zamboni, T. Ledgeway, P. McGraw, and D. Schluppeck. Do perceptual biases emerge early
468 or late in visual processing? Decision-biases in motion perception. *Proc. of Royal Society of*
469 *London B*, 283(1833), 2016. doi: 10.1098/rspb.2016.0263.

A Appendix

A.1 Categorical perception

Our analysis in the main text was limited to comparing optimal estimators under a L_2 -norm loss function (MSE estimates) and symmetric top-level priors. In order to get a sense of the generality of our results, we also compared the accuracy of optimal estimators formulated under both inference strategies for a L_1 -norm loss function (measured with regard to absolute error, correspondingly) as well as for asymmetric category priors $p(C)$ (for L_2 -norm).

A.1.1 Minimizing L_1 -norm loss

Using the same generative model and model parameters as in the main text (Fig. 2a) we formulated optimal estimators for each inference strategy as the median of the posterior distributions Eqs. (3) and (6), respectively, and performed the same accuracy comparison as in the main text with regard to absolute error. As shown in Fig. A.1, the results are qualitatively very similar to the L_2 -norm case (Fig. 3).

Figure A.1: Comparison of the full and conditioned inference strategy when the observer is optimized with regard to a L_1 -norm loss function. Accuracy comparison is correspondingly assessed in terms of L_1 -norm loss (absolute error). Panels correspond to panels in Fig. 3. All other parameters were identical to the example in the main text.

A.1.2 Asymmetric category priors

We also explored the scenario that one category was more likely than the other, *i.e.*, the case $p(C_{1,2}) = [0.2, 0.8]$. Figure A.2 shows the results for both overlapping and non-overlapping categories (panels are similarly organized as in Fig. 3). With unequal category priors, the psychometric curves $p(C|\theta)$ are shifted away from the categorical boundary (*e.g.*, $p(C_1|\theta)$ is shifted away from the more prevalent category C_2). Compared to the equal prior case, the pattern of local relative accuracy maintains, although it is also shifted (Fig. A.2b,f). The asymmetric categorical priors exaggerate the differences between posteriors $p(C_1|m)$ and $p(C_2|m)$ making full inference act more like conditioned inference. The differences in terms of the performance between the two strategies are consequently reduced. In the extreme case where $p(C_{1,2}) = [0, 1]$ (only one category is possible), the two strategies are functionally identical.

A.2 Causal cue combination: different feature prior

We analyzed the causal cue combination example for the case of an approximately uniform categorical feature prior ($\sigma_p = 100$). The relative performance pattern is qualitatively similar compared to the example in the main text.

Figure A.2: Estimation accuracy for asymmetric category priors. The panels are similarly organized and labeled as in Fig. 3. (a) Categorical feature priors multiplied with the categorical prior for illustration purposes. (b) Local performance comparison between conditioned and full inference. Larger benefits are seen when there is reasonable evidence for committing to the less prevalent category (as shown in (c)). (c) Psychometric curve indicating the probability of assigning θ to category $C_{1,2}$. (d) Global MSE ratio as a function of sensory (σ_s) and late noise (σ_m). (e-h) Same as (a-d) for non-overlapping categorical feature priors.

Figure A.3: Simulations of causal cue combination with approximately uniform feature prior ($\sigma_p = 100$). (a,b) Estimate distributions and their differences between the two strategies. (c-g) Relative MSE accuracy. Panels correspond to panels in Fig. 5.