American Psychological Association

Psychological Review

ISSN: 0033-295X

http://dx.doi.org/10.1037/rev0000193

THEORETICAL NOTE

Benefits of Commitment in Hierarchical Inference

Cheng Qiu
University of Pennsylvania

Long Luu
Columbia University

Alan A. Stocker
University of Pennsylvania

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 biases in cognition and perception. However, committing to a single high-level interpretation seems short-sighted and irrational, and thus it is unclear why humans are motivated to use 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 2 common hierarchical inference tasks, category-based perception and causal cue combination. Using model simulations, we demonstrate that although estimation accuracy is generally impaired when conditioned on only a single high-level interpretation, the reduction is not uniform across the entire feature range. Compared with a full inference strategy that considers all high-level interpretations, accuracy is only worse for feature values relatively close to the decision boundaries but is better everywhere else. That is, for feature values for which an observer has a reasonably high chance of being correct about the high-level interpretation of the feature, a full commitment to that particular interpretation is advantageous. We also show that conditioning on an preceding high-level interpretation provides an effective mechanism for partially protecting the evidence from corruption with late noise in the inference process (e.g., during retention in and recall from working memory). Our results suggest that a top-down inference strategy that solely relies on the most likely high-level interpretation can be favorable with regard to late noise and more holistic performance metrics.

Keywords: hierarchical models, top-down inference, model selection, self-consistency, holistic loss function

Cognitive tasks typically require the brain to perform some form of statistical inference based on uncertain evidence and a learned statistical (generative) model of the task (Helmholtz, 1867; Jaynes, 2003; Lee, 2015; Lee & Mumford, 2003). Previous work has shown that the formalism of Bayesian statistics often provides an accurate description of human behavior in a broad range of tasks associated with perception (Knill & Richards, 1996), cognitive reasoning (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010), economic decision-making (Summerfield & 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 (Jardri & Deneve, 2013; Lieder et al., 2019). Except for simple estimation and decision tasks (Ernst & Banks, 2002; Körding & Wolpert, 2004; Stocker & Simoncelli, 2006), the generative models of these inference processes are hierarchical. Object recognition is one 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, Mamassian, & Yuille, 2004). Various studies have shown that in such tasks humans seem to fully integrate all information in the hierarchical generative model from bottom to top. These studies include models of human judgments of sameness (Van den Berg, Vogel, Josic, & Ma, 2012), of stimulus transparency (Hedges, Stocker, & Simoncelli, 2011), or causal stimulus structure (Körding et al., 2007).

Cheng Qiu, Department of Psychology, University of Pennsylvania; Long Luu, Zuckerman Institute, Columbia University; Alan A. Stocker, Department of Psychology, University of Pennsylvania.

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 preprint posted on BioRxiv (Luu et al., 2017) and were presented at the 2019 Vision Science Society conference (Qiu, Luu, & Stocker, 2019). We also thank Pascal Mamassian for constructive feedback. This work has been supported in part by the National Science Foundation of the United States of America (Awards BCS-1350786 and IIS-1912232), and in part by the University of Pennsylvania.

Correspondence concerning this article should be addressed to Alan A. Stocker, Department of Psychology, University of Pennsylvania, Goddard Laboratories Room 421, 3710 Hamilton Walk, Philadelphia, PA 19106. E-mail: astocker@psych.upenn.edu

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 Figure 1a). Full inference dictates that in order to infer the value of the feature, an observer should consider all possible generative hypotheses (e.g., categorical assignments) and weigh them according to how probable they are given the observed evidence (Figure 1b). This strategy is known as optimal model evaluation (Draper, 1995), or Bayesian model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999), and has been considered a rational account for human behavior in various perceptual and cognitive reasoning tasks (e.g., Anderson, 1991; Duffy, Huttenlocher, & Crawford, 2006; Griffiths et al., 2010; Knill, 2003, 2007; Körding et al., 2007).

However, results of several studies suggest that model averaging is not a general inference strategy. For example, it has been long known that human subjects tend to consider only a single category (model selection) rather than all possible categories when performing category-based induction or prediction of a feature value (Chen, Ross, & Murphy, 2016; Hayes & Newell, 2009; Lagnado & Shanks, 2003; Murphy & Ross, 1994, 2005; Newell, Paton, Hayes, & Griffiths, 2010). More recent results suggest that model selection is also prevalent in low-level perceptual tasks. For example, by making a category assignment a subject's subsequent perceptual estimate of a low-level stimulus feature (e.g., motion direction; Jazayeri & Movshon, 2007; Zamboni, Ledgeway, McGraw, & Schluppeck, 2016 or visual orientation; Fritsche & de Lange, 2019; Luu & Stocker, 2018) is biased toward the assigned category on a per trial basis. This also matches recent results showing that postdecision confidence reports overemphasize information supporting a decision (Peters et al., 2017). These choice-induced 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, Urai, Tsetsos, Usher, & Donner, 2018). Importantly, these biases are not dependent on subjects making an explicit, overt choice about a high-level interpretation; similar biases are observed in tasks that did not require an explicit categorical choice (Ding, Cueva, Tsodyks, & Qian, 2017; Wu, Lu, & Yuille, 2009; Zamboni et al., 2016), indicating that committing to a high-level interpretation may be a quite common inference strategy.

As first proposed (Stocker & Simoncelli, 2007), and refined and validated more recently (Luu & Stocker, 2018), the behavioral biases in above examples are remarkably well described by a *conditioned Bayesian 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 only on the chosen hypothesis (Figure 1c). The conditioned Bayesian 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 general notion is that the conditioned observer performs worse at the feature level than a rational observer that integrates over all possible (high-level) hypotheses. However, a detailed quantitative analysis of how the conditioned inference strategy affects inference accuracy at the feature level has been missing. Such analysis is crucial for understanding the advantages and disadvantages of different inference strategies and, ultimately, to uncover the motivation for why humans tend to commit to a single high-level interpretation.

We set out to fill this gap by systematically assessing how performance is affected by applying a conditioned inference strategy. We quantitatively compared its performance with the perfor-

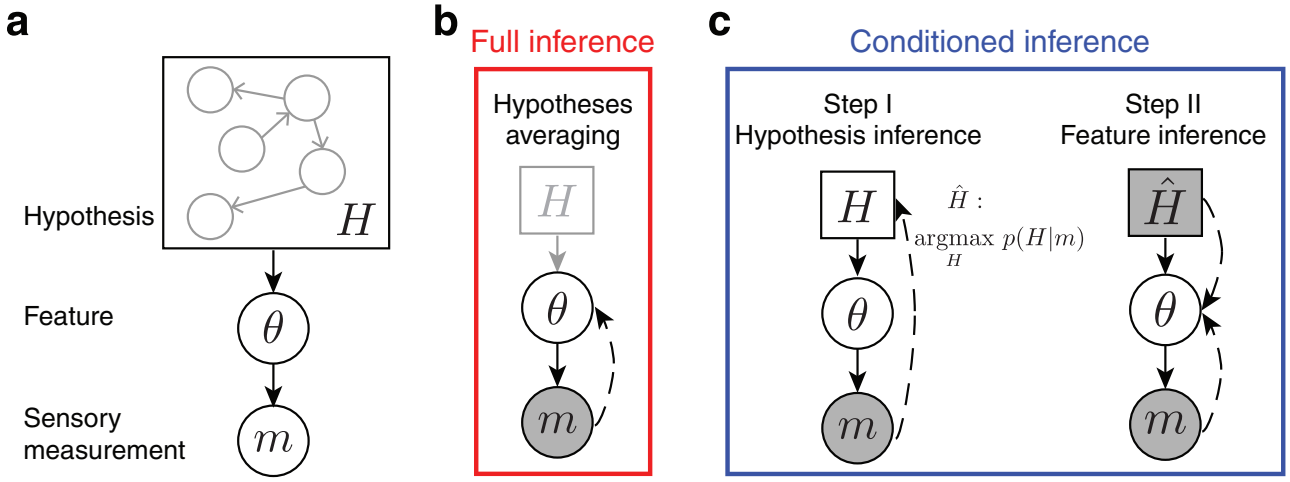

Figure 1. Feature inference in hierarchical generative models. (a) Graphical model that represents the generic class of hierarchical models we consider. Sensory measurement m is assumed to be a noisy sample of the feature value θ, which itself is drawn from a high-level generative hypothesis H. Based 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 only on the committed hypothesis $\hat{H}$. See the online article for the color version of this figure.

mance of a full inference strategy for two simple but common hierarchical inference tasks, category-based perception and causal cue combination. Using simulations we characterized the relative estimation accuracy over a range of different generative model parameters. We show that although a full inference strategy is always globally optimal given the same amount of information, there are distinct local regimes where conditioned inference is better. Our results provide the necessary quantitative analysis for the discussion of different inference strategies and why humans may 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 that can be expressed with a directed graph as shown in Figure 1a. The hierarchical component of the model "above" θ represents a high-level generative hypothesis H about how potential values of θ are generated. 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 (Battaglia, Hamrick, & Tenenbaum, 2013; Kemp & Tenenbaum, 2008; Körding et al., 2007). 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 *full inference* strategy marginalizes over all possible high-level hypotheses when inferring the feature (Figure 1b). With this strategy the posterior over θ becomes a weighted sum

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

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

In contrast, the *conditioned inference* strategy represents a sequential inference process (Luu & Stocker, 2018; Stocker & Simoncelli, 2007): First, a hypothesis $\hat{H}$ is selected according to the posterior probability $p(H \mid m)$ (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 $\hat{H}$ (Figure 1c). The chosen hypothesis imposes a conditioned prior $p(\theta \mid \hat{H}(m))$ that, unlike in the full inference strategy, depends on the sensory evidence m and thus is potentially different in each trial. Accordingly, the posterior probability for this second strategy can be written as

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

In the following we compare the performance of these two strategies in estimating θ. For reasons of simplicity we limit our analysis to scalar features θ.

Error Analysis at the Feature Level

With posterior probabilities Equations 1 and 2 and assuming a quadratic loss function (L_2-norm), we derived optimal estimators for the feature value $\hat{\theta}(m)$ under each inference strategy, and computed their relative expected estimation error. We applied this error analysis to two well-known examples of hierarchical inference, category-based perception and causal cue combination, and systematically investigated the relative performance of the two strategies for different levels of sensory uncertainty and additional (late) processing noise.

Example 1: Category-Based Perception

With *category-based perception* we refer to the task of estimating the value of a low-level feature that is associated with multiple high-level categories. The category association typically biases the perceptual estimate at the feature level (Feldman, Griffiths, & Morgan, 2009; Huttenlocher, Hedges, & Vevea, 2000). Category-based perception assumes the simplest hierarchical generative model possible where the hypothesis H is represented by a single node C reflecting the categorical assignment of feature θ (Figure 2a). The generative process first involves the selection of a category C based on a categorical prior $p(C)$; for simplicity, we consider two possible categories $C \in \{C_1, C_2\}$ (see the Appendix for the case of three categories). A feature value θ is then sampled from the categorical feature prior $p(\theta \mid C)$. Finally, sensory evidence m is sampled from the conditional probability $p(m \mid \theta)$. We explicitly allow the possibility that late noise may deteriorate sensory evidence m, for example, due to retention in working memory. We refer to the deteriorated sensory evidence as m^* distributed according to $p(m^* \mid m)$. The specific description of the priors and conditional probabilities is provided in Figure 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 (Figure 1b) marginalizes over all possible categories, resulting in the posterior distribution

$$p(\theta \mid m) \propto p(m \mid \theta) \sum_i p(\theta \mid C_i) p(C_i) = p(m \mid \theta) p(\theta). \tag{3}$$

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

$$\hat{\theta}_f(m) = \int_\theta \theta \; p(\theta \mid m). \tag{4}$$

The *conditioned inference* strategy (Figure 1c) first chooses the most probable category based on the sensory evidence m according to

$$\hat{C}(m) = \underset{C}{\operatorname{argmax}} \; p(C \mid m), \tag{5}$$

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

$$p(\theta \mid m, \hat{C}(m)) \propto p(m \mid \theta) p(\theta \mid \hat{C}(m)). \tag{6}$$

Finally, the optimal estimator under this strategy is

$$\hat{\theta}_c(m, \hat{C}(m)) = \int_\theta \theta \; p(\theta \mid m, \hat{C}(m)). \tag{7}$$

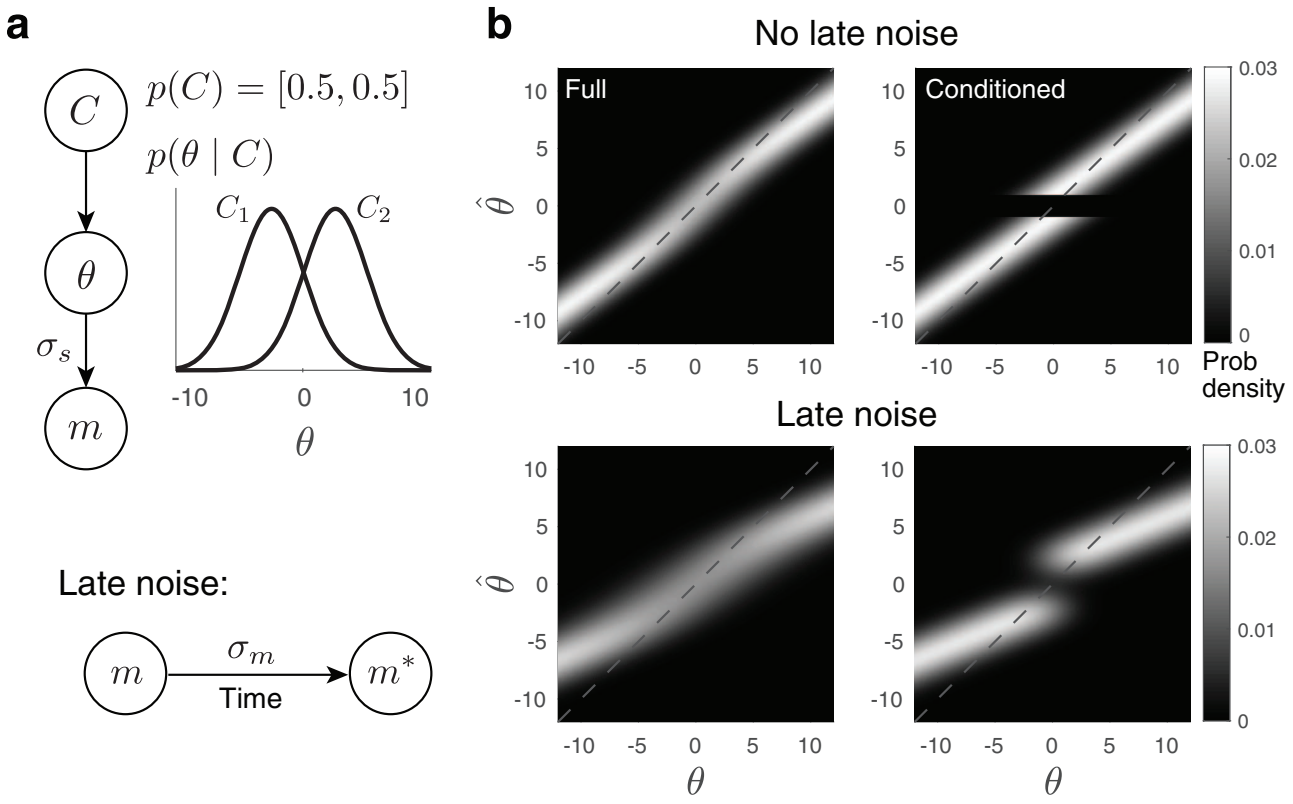

Figure 2. Hierarchical generative model of category-based perception, and the predicted estimate distributions for both inference strategies. (a) Generative model. For reasons of simplicity we make the following assumptions: Feature value θ belongs to one of two categories $C \in \{C_1, C_2\}$ with equal prior probabilities $p(C)$; Categorical feature priors $p(\theta \mid C_i)$ 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 $p(\hat{\theta} \mid \theta)$. 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.

For both strategies the optimal estimator represents a monotonic mapping between the evidence and the estimate. Thus, we can obtain the estimate distributions $p(\hat{\theta}_f \mid \theta)$ and $p(\hat{\theta}_c \mid \theta)$ with a change of variable and the corresponding density transformation for the measurement distribution $p(m \mid \theta)$, replacing m with the estimate $\hat{\theta}_f(m)$ and $\hat{\theta}_c(m)$ according to Equations 4 and 7, respectively. This is computed numerically. 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 & Movshon, 2007; Luu & Stocker, 2018; Zamboni et al., 2016).

Considering the possibility that late noise may further deteriorate sensory evidence m (e.g., due to retention in working memory; Schneegans & Bays, 2018), we update the formulations of the optimal estimators accordingly. We compute the optimal estimate and the estimate distribution for the full inference strategy as discussed above (Equations 3 and 4) but replace $p(m \mid \theta)$ with

$$p(m^* \mid \theta) = \int_m p(m^* \mid m) p(m \mid \theta) \tag{8}$$

where m^* represents the corrupted sensory evidence. Note, we assume that late noise affects feature estimation but not the selection of the category $\hat{C}$ which remains as above (Equation 5). The optimal feature estimator, however, changes to

$$\hat{\theta}_c(m^*, \hat{C}) = \int_\theta \theta \, p(\theta \mid m^*, \hat{C}), \tag{9}$$

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

$$p(\hat{\theta} \mid \theta) = \sum_{\hat{C}} p(\hat{\theta} \mid \theta, \hat{C}) p(\hat{C} \mid \theta), \tag{10}$$

where $p(\hat{C} \mid \theta) = \int_m p(\hat{C} \mid m) p(m \mid \theta)$. Figure 2b shows the estimate distributions given the feature value for full and conditioned inference for no and moderate levels of late noise (additive Gaussian, σ_m).

Relative Accuracy of Full Versus Conditioned Inference

Having defined the optimal estimate distributions for both the full and the conditioned inference strategy, we next performed a systematic quantitative comparison between the estimation accuracy for various model parameters. 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 $\mathrm{MSE}(\theta) = \int_{\hat{\theta}} p(\hat{\theta} \mid \theta)(\hat{\theta} - \theta)^2$ and then took the ratio), as well as globally (i.e., we marginalized the local error over the total distribution of feature values $\mathrm{MSE} = \int_{\theta} \mathrm{MSE}(\theta)p(\theta)$ and then took the ratio).

First, we assumed overlapping categorical feature priors $p(\theta \mid C_{1,2})$ (Figure 3a). Local relative accuracy with and without late noise is shown in Figure 3b. Both curves show a similar, characteristic shape. For θ values close to the category decision boundary, the full inference strategy provides better estimates. However, for θ values that are farther away from the category decision boundary, the conditioned inference strategy consistently outperforms the full inference strategy. The advantage of the conditioned inference strategy is amplified if we further assume that late noise 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 $\hat{C}$ (Figure 3c). It demonstrates that for feature values for which the observer has a reasonably high chance of correctly inferring the category (e.g., in this case a probability correct of $p = .81$ 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 (Figure 3d). This is because the benefits of the full inference strategy for θ values around the decision boundary make up for the deficits at θ values outside of that range. The situation changes, however, if substantial late noise (σ_m) impacts the inference process. In this case the conditioned strategy can also globally outperform the full inference strategy (Figure 3d).

We also compared estimation accuracy for nonoverlapping prior distributions (Figure 3e–h) that correspond to the estimation tasks used in some previous studies (Jazayeri & Movshon, 2007; Luu & Stocker, 2018; Zamboni et al., 2016), as well as priors that lead to multiple categorical decision boundaries (Figure 3i–k) such as when the two category distributions share the same mean but differ in their spread (Berg et al., 2012; Qamar et al., 2013). The general pattern throughout all these examples is that the advantage of the full inference strategy in estimation accuracy is limited to feature values close to the optimal decision boundaries; that is for feature values for which the decision-maker has the highest probability of making a categorical assignment error. Outside of these ranges the conditioned inference strategy performs better. This is further confirmed by cases where the optimal decision boundaries move, for example, due to changes in the category prior probability $p(C)$ (see Appendix Figure A2).

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 used symmetric error metrics as well. We found that the above results of our analysis qualitatively generalize (see Appendix Figure A3).

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 (Alais & Burr, 2004; Butler, Smith, Campos, & Bülthoff, 2010; Ernst & Banks, 2002; Fetsch, Turner, DeAngelis, & Angelaki, 2009; Jacobs, 1999). However, human observers do not integrate cues that are inconsistent with such interpretation (cue conflict) and thus signal different underlying feature values (Roach, Heron, & McGraw, 2006; Wallace et al., 2004).

This integration versus segregation distinction can be modeled within the generative hierarchical framework discussed here (Figure 1a). However, different from the category-based perception example, the hypotheses now represent two different causal structures (Figure 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 $p(\theta_1, \theta_2 \mid m_1, m_2)$. 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 $p(\theta_1, \theta_2 \mid m_1, m_2, \hat{S})$.

Feature Estimation

According to the hierarchical generative model shown in Figure 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

$$p(\theta_{1,2} = \theta \mid m_1, m_2, S_1) \propto p(m_1, m_2 \mid \theta)p(\theta \mid S_1). \tag{11}$$

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

$$p(\theta_1, \theta_2 \mid m_1, m_2, S_2) \propto p(m_1 \mid \theta_1)p(m_2 \mid \theta_2)p(\theta_1, \theta_2 \mid S_2). \tag{12}$$

With a full inference strategy, we compute the total posterior as the average posterior under both hypotheses (Equations 11 and 12) weighted by the posterior probability of each hypothesis, thus

$$p(\theta_1, \theta_2 \mid m_1, m_2) = \sum_i p(\theta_1, \theta_2 \mid m_1, m_2, S_i)p(S_i \mid m_1, m_2). \tag{13}$$

The posterior $p(S \mid m_1, m_2)$ for the two hypothesis is

$$p(S = S_1 \mid m_1, m_2) \propto \int_{\theta} p(m_1, m_2 \mid \theta)p(\theta \mid S_1)p(S_1) \tag{14}$$

and

$$p(S = S_2 \mid m_1, m_2) \propto \int_{\theta_1} \int_{\theta_2} p(m_1 \mid \theta_1)p(m_2 \mid \theta_2)p(\theta_1, \theta_2 \mid S_2)p(S_2), \tag{15}$$

respectively. As in our first example, we define the optimal estimate of the feature vector $[\theta_1, \theta_2]$ with regard to the L_2-norm which corresponds to the expectation over the total posterior Equation 13 (mean of posterior), thus

$$(\hat{\theta}_1, \hat{\theta}_2) = \mathrm{E}[\theta_1, \theta_2 \,|\, m_1, m_2]. \tag{16}$$

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

$$\hat{S} = \underset{S}{\operatorname{argmax}}\, p(S \,|\, m_1, m_2), \tag{17}$$

with $p(S \,|\, m_1, m_2)$ as defined by (Equations 14 and 15). The optimal estimate of the feature vector is again the expectation over

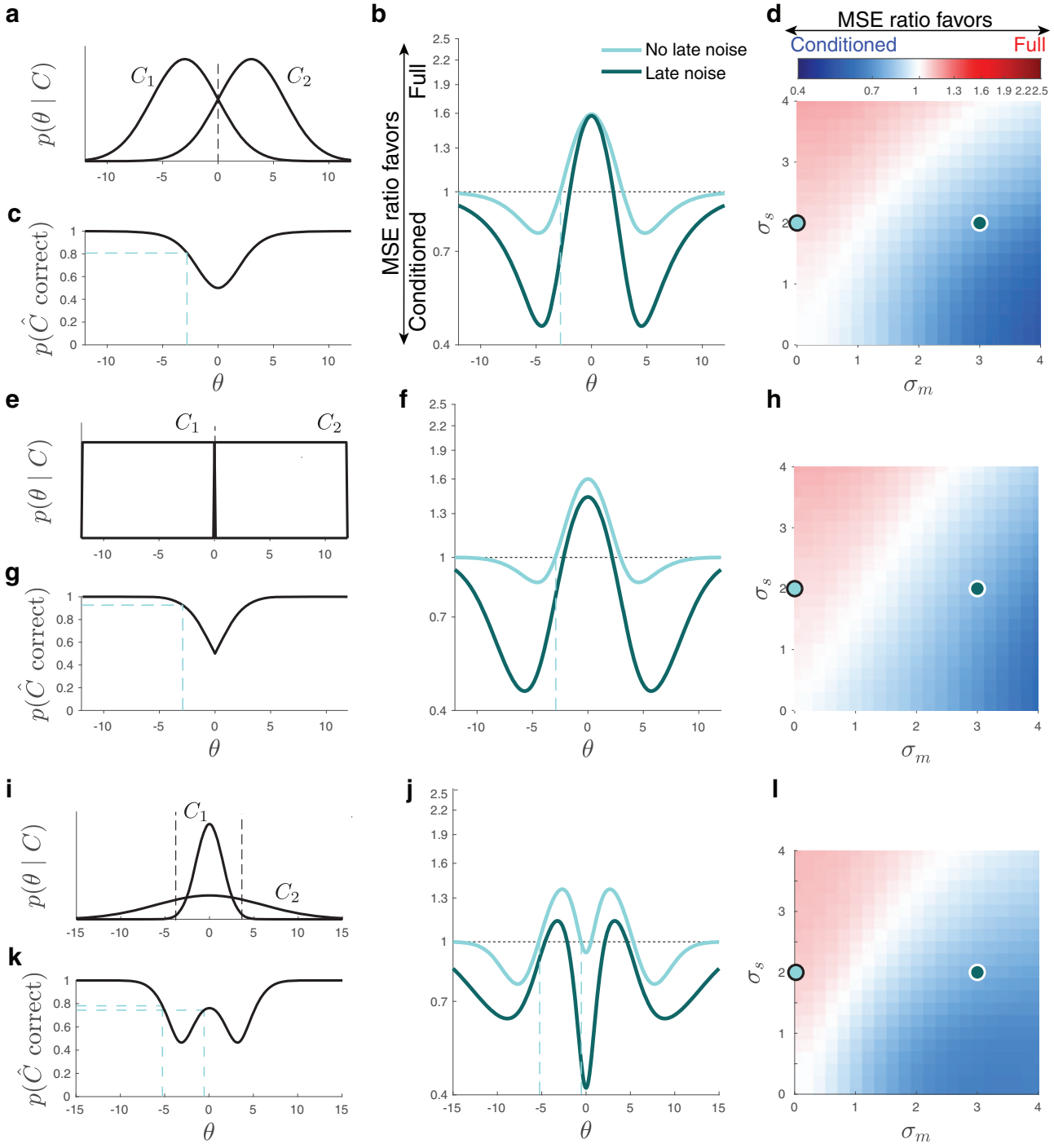

Figure 3 (opposite)

the posterior distributions; however, now conditioned on $\hat{S}$ as well, thus

$$(\hat{\theta}_1, \hat{\theta}_2) = \mathrm{E}[\theta_1, \theta_2 \mid m_1, m_2, \hat{S}], \tag{18}$$

where the posterior is given as in Equations 11 and 12, according to $\hat{S}$.

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

Unlike in the category-based perception example, the resulting estimate distributions for both inference strategies are very similar given the chosen model parameters. Plotting the difference in distribution density (Figure 4c), however, shows that the estimates are more veridical for the full inference strategy for values of θ_1 close to θ_2, whereas the conditioned strategy leads to more veridical estimates for conditions where there is a substantial cue conflict, that is, $\theta_1 \neq \theta_2$.

Comparing Estimation Accuracy of Full Versus Conditioned Inference Strategy

Similar to the previous example, we analyzed the relative estimation accuracy of the two inference strategies by computing both the local *MSE* (as a function of $\theta_{1,2}$) as well as the global *MSE* (integral over all $\theta_{1,2}$). Results are shown in Figure 5. For feature value pairs (θ_1, θ_2) that may support both the common and the independent source hypotheses, full inference outperforms conditioned inference (Figure 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 shown in Figure 5d. It is similar to the category based perception example. In situations where there is a reasonably high probability that the observer is correct in the high-level interpretation, conditioned inference proves to be the better strategy. This advantage is further amplified if late noise affects feature inference (Figure 5e).

Performance as a Function of Sensory Measurement

So far, we have looked at the performance ratio as a function of the feature value θ. Alternatively, however, we can analyze performance from the observer's perspective as a function of the sensory signal m^*. Assuming that m^* is the only trial-specific information available to the observer at the time of estimation, we computed the relative estimation accuracy of the two inference strategies as a function of m^* rather than the feature value θ. Because every m^* corresponds to an estimate $\hat{\theta}(m^*)$ for each strategy according to Equations 4 and 9, respectively, we can calculate the *MSE* ratio between the two estimates averaged across all potential θ values that could have generated m^* according to $p(\theta \mid m^*)$. We show this analysis for the category-based perception example with overlapping categories (see Figure 3a) but the results are general.

Without memory noise the *MSE* ratio is consistently in favor of the full inference strategy for all values of m^* (Figure 6a). This is expected and reflects the fact that the full inference strategy is optimal. However, if memory noise starts to corrupt the sensory measurement conditioned inference performs increasingly better up to the point where it globally outperforms full inference. The trade-off in accuracy between the two strategies is limited to values of m^* close to the decision boundaries. This becomes even clearer when we separately analyze relative accuracy based on whether $\hat{C}(m)$ was correct or not: conditioned inference is always better than full inference for trials with correct $\hat{C}(m)$ (Figure 6b) and is always worse for trials with incorrect $\hat{C}(m)$ regardless of the specific noise conditions (Figure 6c), with the biggest differences occurring for m^* values close to the decision boundary.

Trial by Trial Strategy Selection

In principle, an observer could decide on each trial whether to use conditioned or full inference, therefore potentially combining the best of both worlds. Without memory noise, however, there is no benefit in strategy switching because full inference is always optimal in terms of *MSE*. Even with memory noise, our results rule out any active decision strategy to improve overall accuracy that only has access to the measurement m^*; there is no clear correspondence between being correct in the high-level interpretation (Figure 6d) based on m^* and a potential performance advantage of one over the other strategy (Figure 6a).

Yet, because the estimation accuracy becomes quickly indistinguishable for measurements farther away from the decision boundary, an active strategy could abandon full inference and switch to the computationally simpler conditioned inference strategy for m^* for which there is only a small chance of an incorrect high-level interpretation; essentially not bothering to consider a high-level interpretation that only has a small probability to be correct be-

Figure 3 (*opposite*). Relative estimation accuracy for conditioned and full inference in category-based perception. (a) Overlapping categorical feature priors (Gaussians: $\mu_C = \pm 3$, $\sigma_C = 3$). The categorical prior $p(C)$ is symmetric. The dashed line indicates the optimal decision boundary for making a category assignment. (b) Relative accuracy (MSE(θ) ratio) for the two strategies as a function of θ assuming Gaussian sensory noise ($\sigma_s = 2$) and late noise either absent (light color) or present (Gaussian: $\sigma_m = 3$, dark color). While worse for values close to the decision boundary, the conditioned inference strategy provides superior estimates for feature values further away. (c) The probability for making a correct categorical assignment $\hat{C}$ of the model observer. Dashed lines in (b) and (c) indicate probability correct for feature values beyond which conditioned inference provides more accurate estimates. These probabilities do not correspond to a fixed value but depend on the parameter of the generative model such as the categorical feature prior (see also the following examples). (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 two dots correspond to the two conditions shown in (b). (e–h) Same analysis for nonoverlapping box prior distributions and (i–l) Gaussian prior distributions with identical mean ($\mu_C = 0$) but different standard deviations ($\sigma_{C_1} = 1.5$, $\sigma_{C_2} = 6$). MSE = mean squared-error. See the online article for the color version of this figure.

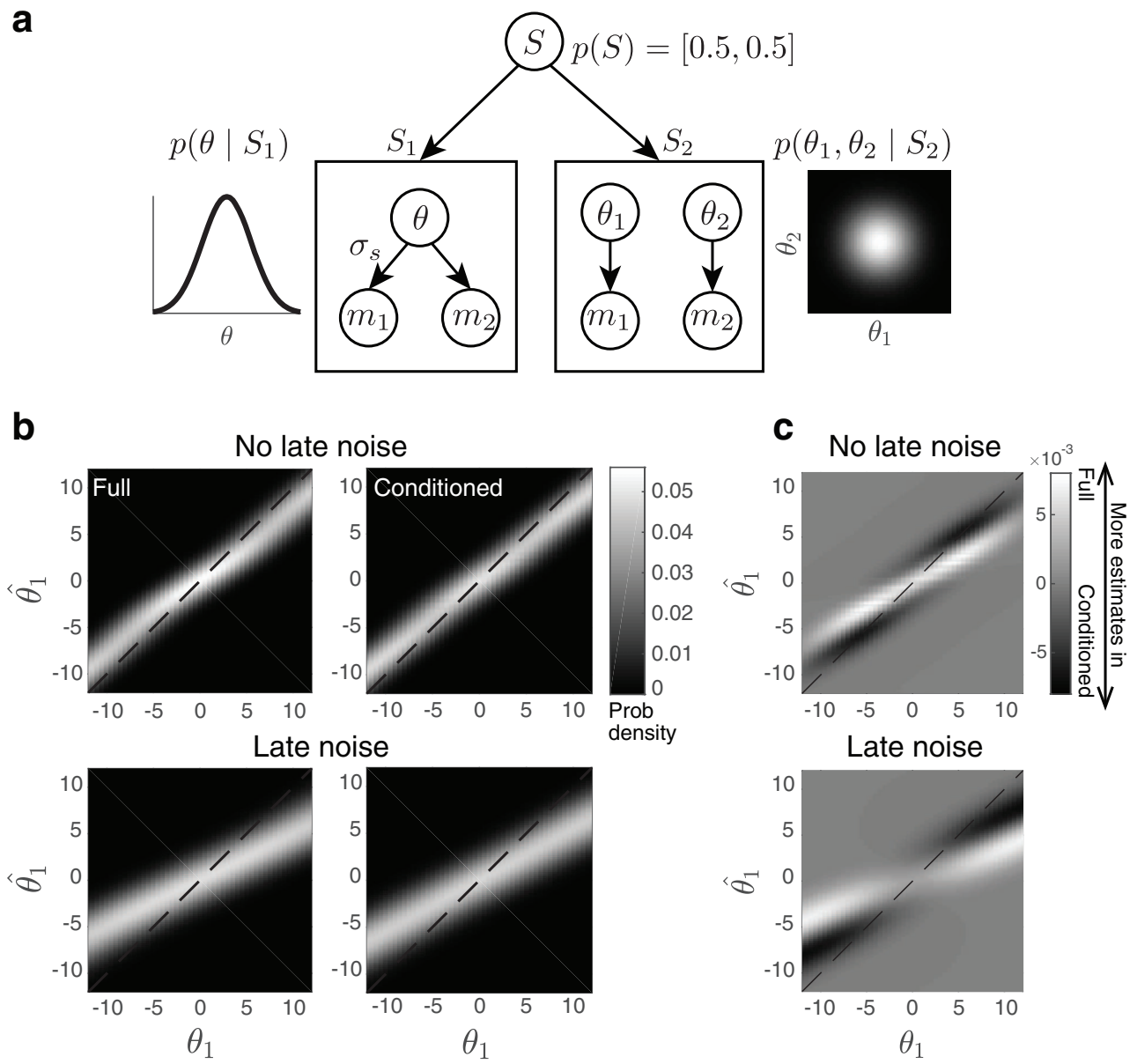

Figure 4. Causal cue combination. (a) Graphical model reflecting two possible generative structures (hypotheses): The observed sensory measurements $m_{1,2}$ may be generated by a common source (S_1) or two independent sources (S_2). For reasons of simplicity we assume the structure prior $p(S)$ to be equal, and the prior over feature values to be normally distributed: $\mathcal{N}(0, \sigma_p = 4)$ for S_1, and $\mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \Sigma = \begin{bmatrix}16 & 0\\0 & 16\end{bmatrix}\right)$ 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 $p(\hat{\theta}_1 \mid \theta_1, \theta_2 = 0)$ 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.

cause the expected gain in accuracy is negligible. Theoretically, there is also the possibility that a metacognitive process that has access to additional information about the correctness of the high-level interpretation (Fleming & Daw, 2017; Mamassian, 2016), could actively guide the selection of the appropriate inference strategy to achieve better estimation accuracy. The important assumption, however, is that the additional information can only be used to compute a confidence signal but not to infer the feature values. The potential impact of metacognition in strategy selection remains a topic for future investigations.

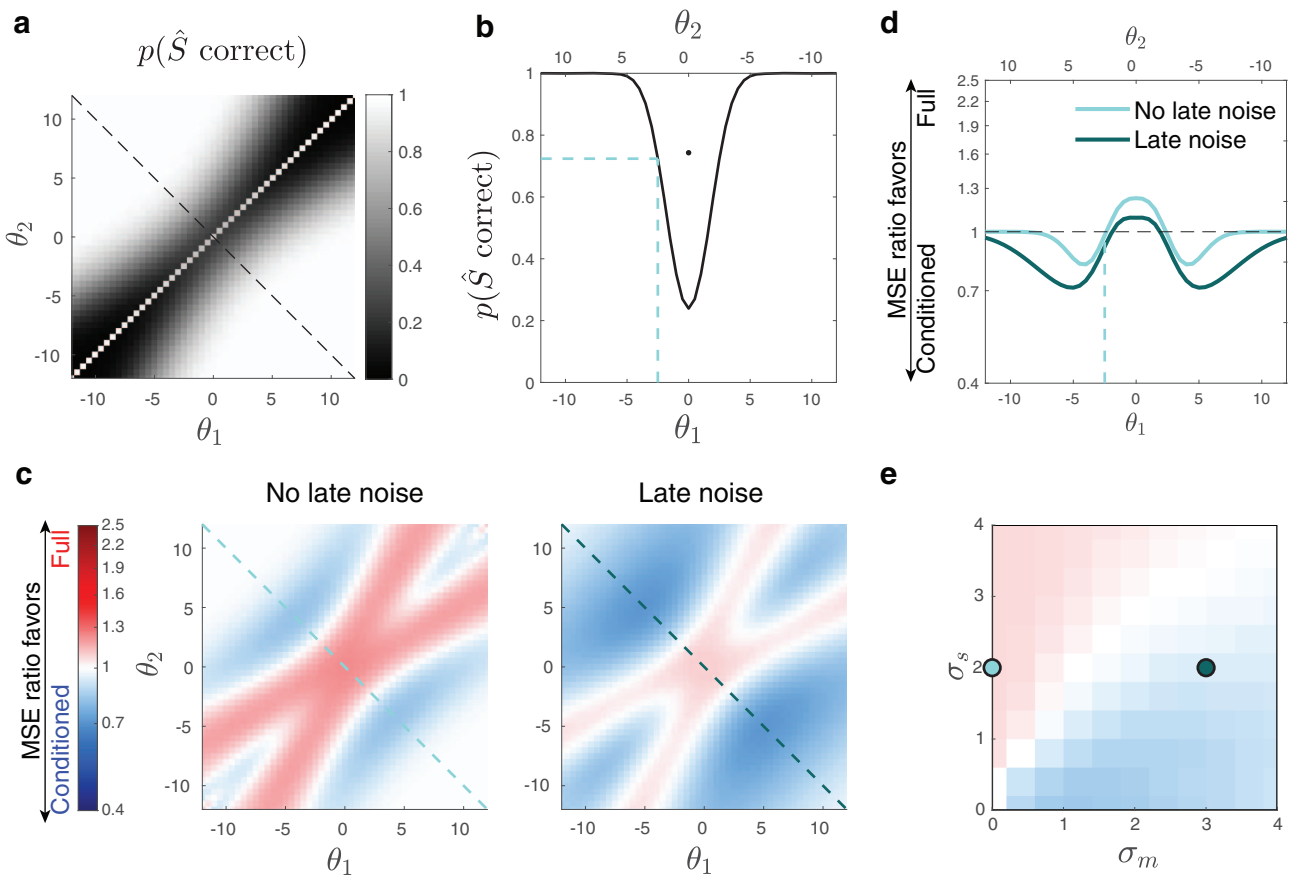

Figure 5. Relative performance between conditioned and full inference for causal cue combination. (a) Probability of making a correct structural assignment $\hat{S}$ as a function of feature values. When θ_1 and θ_2 are far apart, they most likely represent independent sources, whereas when they are 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 blue 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). MSE = mean squared-error. See the online article for the color version of this figure.

Discussion

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 although overall optimal, considering all possible high-level interpretations does not consistently provide better accuracy across the entire range of feature values. Committing to a single interpretation is actually the better strategy for feature values for which the observer has a reasonably high chance of being correct in their high-level interpretation. That is, the penalty of conditioned inference is limited to a relatively small range around the feature values that correspond to the decision boundaries in the high-level interpretations (e.g., the category boundaries). This performance pattern is general and robust to variations of the considered hierarchical models such as the distributions and number of categories as well as other symmetric error metrics (see the Appendix).

The pattern suggests a potential explanation for why humans apply a conditioned inference strategy. While we focused our analysis on comparing inference accuracy only at the feature level (i.e., a loss function only including θ), human cognition is a *holistic process* that involves simultaneous and cojoint inference processes at all hierarchical levels. Thus, it is likely that human inference strategies have evolved with regard to error metrics that are jointly defined across all levels of a hierarchical representation. For example, in the case of category-based perception (see Figure 2a), such error metric would include not only errors at the feature but also at the category level. Furthermore, one can easily make

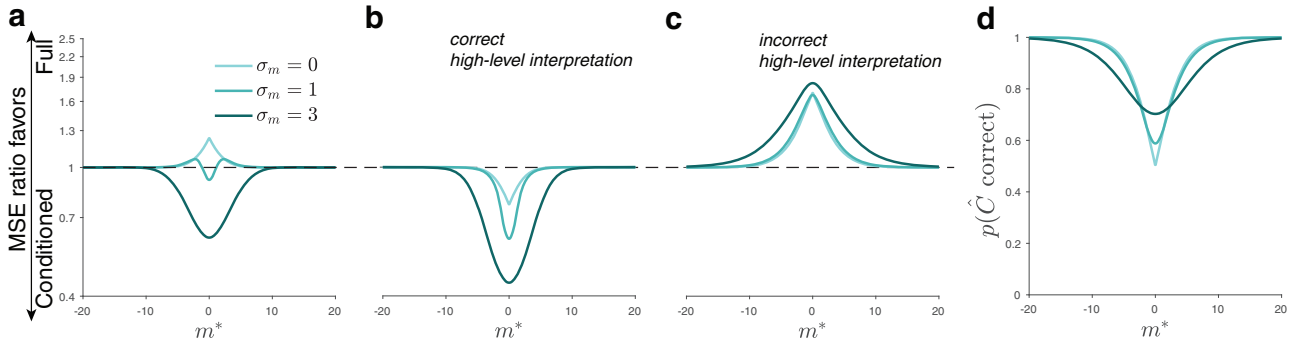

Figure 6. Estimation accuracy as a function of the sensory measurement. (a) Expected relative estimation accuracy (*MSE* ratio) as a function of the sensory measurement m^* for different late noise conditions with sensory noise $\sigma_s = 2$. Under no late noise (m^* equals m for $\sigma_m = 0$), the full inference strategy is always better. We can separately compute relative accuracy for trials in which the high-level categorical interpretation $\hat{C}$ was either (b) correct or (c) incorrect. The total relative estimation accuracy in (a) is the sum of the curves in (b) and (c), weighted by the probability of $\hat{C}$ being correct/incorrect (shown in (d)). For correct trials only, conditioned inference always provides better estimation accuracy even under no late noise conditions. MSE = mean squared-error. See the online article for the color version of this figure.

the case that consistency across the hierarchy is important: Errors at the feature level become irrelevant if at the same time the category assignment of the feature is already wrong (i.e., "Why bother about making an error at the low level if inference is wrong at the top level?"). As we have shown, conditioned inference globally outperforms full inference under such hierarchical error metric (Figure 6b).

We also showed that a conditioned inference strategy becomes increasingly favorable when late noise corrupts the sensory signal m. The underlying assumption we made is that due to its discrete nature (binary in the simplest case) the commitment to the most likely high-level interpretation is a signal robust to late noise. The signal maintains information about the original (uncorrupted) sensory measurement m that the conditioned inference strategy then can exploit during feature inference. Note that the global performance advantage of conditioned inference under late noise conditions is due to this extra information. As we can show, a full inference strategy that also has access to the commitment signal $\hat{C}(m)$ once again globally outperforms conditioned inference under all noise conditions (see Appendix Figure A4). However, adding the signal this way comes with additional costs as it requires an update of the generative model with additional probabilistic information. This information is analog and thus it is questionable whether this strategy is robust to late noise. In contrast, conditioned inference provides a simple and robust way to exploit some of the information about the original measurement m contained in the commitment $\hat{C}(m)$ later during the feature inference process. However, improved accuracy is not the only advantage. We have recently shown that conditioned inference intrinsically avoids inconsistencies of the representations across the hierarchy that may occur under late noise conditions (Luu & Stocker, 2018), and thus may serve as a mechanism to avoid states of cognitive dissonance (Brehm, 1956; Festinger, 1957).

Our results with regard to late noise also have implications for models and theories of working memory formation and recall, and in general, for inference tasks that evolve over time (Gold & Stocker, 2017). In fact, conditioned inference can predict some of the memory biases that have been reported in the recall of color (Bae, Olkkonen, Allred, & Flombaum, 2015) or other low-level visual features (Ding et al., 2017; Luu, Qiu, & Stocker, 2017). It provides a normative motivation for memory biases and potentially other similar forms of confirmation biases (Lange, Chattoraj, Beck, Yates, & Haefner, 2019; Talluri et al., 2018).

Our analysis focused on performance accuracy and did not consider potential differences in terms of computational and representational resource constraints. Naturally, cognitive and perceptual inference processes are subject to such constraints, promoting inference strategies that are commonly referred to as bounded rationality (Gershman, Horvitz, & Tenebaum, 2015; Simon, 1984). Conditioned inference seems a far simpler and computationally less costly strategy than full inference as it does not require marginalization over all potential high-level hypotheses. Furthermore, under many conditions marginalization over all high-level hypotheses is computationally infeasible. Previous studies have suggested that under these conditions humans may apply sampling procedures to approximate full inference (Sanborn, Griffiths, & Navarro, 2010; Vul, Goodman, Griffiths, & Tenebaum, 2014). While formally related—conditioned inference can technically be considered a 1-particle/1-sample particle filter (Brown & Steyvers, 2009)—these sampling models are conceptually quite different as they are intended to approximate full inference. They are averaging models (using samples rather than full distributions) and thus they behave and perform naturally much closer to the full rather than the conditioned inference model discussed here. Future work will be needed to investigate in more detail the overall benefits (or drawbacks) of a conditioned inference strategy with regard to hierarchical error metrics and other costs such as computational and representational complexity. This is of particular interest in light of the ongoing discussion about optimality in human perception and cognition (Rahnev & Denison, 2018; Stocker, 2019).

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. Thus our results may have implications far beyond the particular examples we discussed here. The problem is that conditioned inference does not always lead to clearly identifiable behavioral signatures such as the bimodal estimate distributions in category-based perception (Figure 2b). In some cases behavioral effects may be small and difficult to extract from experimental data as in the case of causal cue combination (Figure 4b, 4c). Other candidates for conditioned inference 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 (Coen-Cagli, Kohn, & Schwartz, 2015; Qiu, Kersten, & Olman, 2013; Schwartz, Sejnowski, & Dayan, 2009), lightness perception affected by perceived surface curvature (Knill & 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.

References

Adelson, E. (1993). Perceptual organization and the judgement of brightness. *Science, 262,* 2042–2044.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257–262.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General, 144,* 744–763.

Battaglia, P., Hamrick, J., & Tenenbaum, J. (2013). Simulation as an engine of physical scene understanding. *Proceeding of the National Academy of Sciences of the United States of America, 110,* 18327–18332. http://dx.doi.org/10.1073/pnas.1306572110

Brehm, J. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology, 52,* 384–389.

Bronfman, Z. Z., Brezis, N., Moran, R., Tsetsos, K., Donner, T., & Usher, M. (2015). Decisions reduce sensitivity to subsequent information. *Proceedings of the Royal Society of London B: Biological Sciences, 282,* 20150228. http://dx.doi.org/10.1098/rspb.2015.0228

Brown, S., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58,* 49–67. http://dx.doi.org/10.1016/j.cogpsych.2008.09.002

Butler, J. S., Smith, S. T., Campos, J. L., & Bülthoff, H. H. (2010). Bayesian integration of visual and vestibular signals for heading. *Journal of Vision, 10*(11), 23. http://dx.doi.org/10.1167/10.11.23

Chen, S., Ross, B., & Murphy, G. (2016). Eyetracking reveals multiple-category use in induction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 42,* 1050–1067.

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience, 18,* 1648–1655.

Ding, S., Cueva, C., Tsodyks, M., & Qian, N. (2017). Visual perception as retrospective Bayesian decoding from high- to low-level features. *Proceeding of the National Academy of Sciences of the United States of America, 114,* E9115–E9124. http://dx.doi.org/10.1073/pnas.1706906114

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society B, 57,* 45–97.

Duffy, S., Huttenlocher, J., & Crawford, L. (2006). Children use categories to maximize accuracy in estimation. *Developmental Science, 9,* 597–603.

Ernst, M., & Banks, M. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415,* 429–433.

Feldman, N., Griffiths, T., & Morgan, J. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116,* 752–782. http://dx.doi.org/10.1037/a0017196

Festinger, L. (1957). *Theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fetsch, C. R., Turner, A. H., DeAngelis, G. C., & Angelaki, D. E. (2009). Dynamic reweighting of visual and vestibular cues during self-motion perception. *Journal of Neuroscience, 29,* 15601–15612. http://dx.doi.org/10.1523/JNEUROSCI.2574-09.2009

Fleming, S., & Daw, N. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review, 124,* 91–114.

Fritsche, M., & de Lange, F. P. (2019). Reference repulsion is not a perceptual illusion. *Cognition, 184,* 107–118. http://dx.doi.org/10.1016/j.cognition.2018.12.010

Gershman, S., Horvitz, E., & Tenebaum, J. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science, 349,* 273–278.

Gold, J. I., & Stocker, A. A. (2017). Visual decision-making in an uncertain and dynamic world. *Annual Review of Vision Science, 3,* 227–250. http://dx.doi.org/10.1146/annurev-vision-111815-114511

Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences, 14,* 357–364.

Hayes, B. K., & Newell, B. R. (2009). Induction with uncertain categories: When do people consider the category alternatives? *Memory & Cognition, 37,* 730–743. http://dx.doi.org/10.3758/MC.37.6.730

Hedges, J. H., Stocker, A. A., & Simoncelli, E. P. (2011). Optimal inference explains the perceptual coherence of visual motion stimuli. *Journal of Vision, 11*(6), 14.

Helmholtz, H. v. (1867). *Handbuch der physiologischen optik* [Handbook of physiological optics]. Leipzig, Germany: Voss.

Hoeting, J., Madigan, D., Raftery, A., & Volinsky, C. (1999). Bayesian model averaging: A tutorial. *Statistical Science, 14,* 382–417.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of experimental psychology: General, 129,* 220–241.

Jacobs, R. (1999). Optimal integration of texture and motion cues to depth. *Vision Research, 39,* 3621–3629.

Jardri, R., & Deneve, S. (2013). Circular inferences in schizophrenia. *Brain, 136,* 3227–3241. http://dx.doi.org/10.1093/brain/awt257

Jaynes, E. (2003). *Probability theory: The logic of science* (G. Bretthorst, Ed.). New York, NY: Cambridge University Press.

Jazayeri, M., & Movshon, J. (2007). A new perceptual illusion reveals mechanisms of sensory decoding. *Nature, 446,* 912–915.

Kemp, C., & Tenenbaum, J. (2008). The discovery of structural form. *Proceedings of the National Academies of Sciences United States of America, 105,* 10687–10692.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology, 55,* 271–304.

Knill, D. C. (2003). Mixture models and the probabilistic structure of depth cues. *Vision Research, 43,* 831–854.

Knill, D. C. (2007). Robust cue integration: A Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. *Journal of Vision, 7*(7), 5.

Knill, D. C., & Kersten, D. (1991). Apparent surface curvature affects lightness perception. *Nature, 351,* 228–230. http://dx.doi.org/10.1038/351228a0

Knill, D. C., & Richards, W. (Eds.). (1996). *Perception as Bayesian inference*. New York, NY: Cambridge University Press.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2*(9), e943.

Körding, K., & Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature, 427,* 244–247.

Lagnado, D., & Shanks, D. (2003). The influence of hierarchy on probability judgment. *Cognition, 89,* 157–178.

Lange, R. D., Chattoraj, A., Beck, J. M., Yates, J. L., & Haefner, R. M. (2019). A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *bioRxiv*. Advance online publication. http://dx.doi.org/10.1101/440321

Lee, T. S. (2015). The visual system's internal model of the world. *Proceedings of the IEEE, 99,* 1–20. http://dx.doi.org/10.1109/JPROC.2015.2434601

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America, 20,* 1434–1448.

Lieder, I., Adam, V., Frenkel, O., Jaffe-Dax, S., Sahani, M., & Ahissar, M. (2019). Perceptual bias reveals slow-updating in autism and fast-forgetting in dyslexia. *Nature Neuroscience, 22,* 256–264. http://dx.doi.org/10.1038/s41593-018-0308-9

Luu, L., Qiu, C., & Stocker, A. A. (2017). High- to low-level decoding does not generally improve perceptual performance. *bioRxiv*. Advance online publication. http://dx.doi.org/10.1101/240390

Luu, L., & Stocker, A. A. (2018). Post-decision biases reveal a self-consistency principle in perceptual inference. *eLife, 7,* e33334. http://dx.doi.org/10.7554/eLife.33334

Mamassian, P. (2016). Visual confidence. *Annual Review of Vision Science, 2,* 459–481. http://dx.doi.org/10.1146/annurev-vision-111815-114630

Murphy, G. L., & Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology, 27,* 148–193. http://dx.doi.org/10.1006/cogp.1994.1015

Murphy, G. L., & Ross, B. H. (2005). The two faces of typicality in category-based induction. *Cognition, 95,* 175–200. http://dx.doi.org/10.1016/j.cognition.2004.01.009

Newell, B., Paton, H., Hayes, B., & Griffiths, O. (2010). Speeded induction under uncertainty: The influence of multiple categories and feature conjunctions. *Psychonomic Bulletin & Review, 17,* 869–874. http://dx.doi.org/10.3758/PBR.17.6.869

Nickerson, R. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2,* 175–220.

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., . . . Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour, 1,* 0139. http://dx.doi.org/10.1038/s41562-017-0139

Qamar, A. T., Cotton, R. J., George, R. G., Beck, J. M., Prezhdo, E., Laudano, A., . . . Ma, W. J. (2013). Trial-to-trial, uncertainty-based adjustment of decision boundaries in visual categorization. *Proceedings of the National Academy of Sciences, 110,* 20332–20337. http://dx.doi.org/10.1073/pnas.1219756110

Qiu, C., Kersten, D., & Olman, C. A. (2013). Segmentation decreases the magnitude of the tilt illusion. *Journal of Vision, 13*(13), 19. http://dx.doi.org/10.1167/13.13.19

Qiu, C., Luu, L., & Stocker, A. A. (2019, May). *Is "confirmation bias" always a bad thing?* Poster session presented at Vision Science Society VSS Conference, St. Pete Beach, FL.

Rahnev, D., & Denison, R. N. (2018). Suboptimality in perceptual decision making. *Behavioral and Brain Sciences, 41,* 1–107. http://dx.doi.org/10.1017/S0140525X18000936

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society of London B: Biological Sciences, 273,* 2159–2168. http://dx.doi.org/10.1098/rspb.2006.3578

Sanborn, A. N., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117,* 1144–1167. http://dx.doi.org/10.1037/a0020511

Schneegans, S., & Bays, P. M. (2018). Drift in neural population activity causes working memory to deteriorate over time. *Journal of Neuroscience, 38,* 4859–4869. http://dx.doi.org/10.1523/JNEUROSCI.3440-17.2018

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2009). Perceptual organization in the tilt illusion. *Journal of Vision, 9*(4), 19. http://dx.doi.org/10.1167/9.4.19

Simon, H. (1984). *Models of bounded rationality*. Cambridge, MA: MIT Press.

Stocker, A. A. (2019). Credo for optimality. *Behavioral and Brain Sciences, 41,* 38–39. http://dx.doi.org/10.1017/S0140525X18001346

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9,* 578–585. http://dx.doi.org/10.1038/nn1669

Stocker, A. A., & Simoncelli, E. P. (2007). A Bayesian model of conditioned perception. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems NIPS 20* (pp. 1409–1416). Cambridge, MA: MIT Press.

Summerfield, C., & Tsetsos, K. (2012). Building bridges between perceptual and economic decision-making: Neural and computational mechanisms. *Frontiers in Neuroscience, 6,* 1–20. http://dx.doi.org/10.3389/fnins.2012.00070

Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M., & Donner, T. H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology, 28,* 1–8. http://dx.doi.org/10.1016/j.cub.2018.07.052

Van den Berg, R., Vogel, M., Josic, K., & Ma, W. (2012). Optimal inference of sameness. *Proceeding of the National Academy of Sciences of the United States of America, 109,* 3178–3183.

Vul, E., Goodman, N., Griffiths, T., & Tenebaum, J. (2014). One and done? optimal decisions from very few samples. *Cognitive Science, 38,* 599–637. http://dx.doi.org/10.1111/cogs.12101

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. *Experimental Brain Research, 158,* 252–258.

Wolpert, D. (2007). Probabilistic models in human sensorimotor control. *Human Movement Science, 26,* 511–524. http://dx.doi.org/10.1016/j.humov.2007.05.005

Wu, S., Lu, H., & Yuille, A. (2009). Model selection and parameter estimation in motion perception. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 1793–1800). Cambridge, MA: MIT Press.

Zamboni, E., Ledgeway, T., McGraw, P., & Schluppeck, D. (2016). Do perceptual biases emerge early or late in visual processing? Decision-biases in motion perception. *Proc. of Royal Society of London B, 283,* 1–9. http://dx.doi.org/10.1098/rspb.2016.0263

(*Appendix follows*)

Appendix

Category-Based Perception

The analysis in the main text is limited to comparing optimal estimators under a L_2-norm loss function (MSE estimates) and symmetric top-level priors for two categories. In order to probe the generality of our results, we extended our comparison to estimators according to a L_1-norm loss function (measured with regard to absolute error, correspondingly) as well as for asymmetric category priors $p(C)$ and multiple categories (for L_2-norm). In addition, we considered a full inference strategy with knowledge of $\hat{C}(m)$ under late noise conditions (effectively, bottom-up conditioning), and compared it with the top-down conditioned inference.

More Than Two Categories

Relative estimation accuracy exhibits the same pattern if the high-level interpretation consists in deciding among more than two potential categorical assignments. Figure A1 shows an example for three different categories and their categorical prior distributions (Figure A1a). As with two categories, accuracy deficits of conditioned inference are limited to feature values close to the decision boundaries whereas accuracy is better everywhere else when compared with a full inference strategy (Figure A1b). If sufficient late noise is corrupting the sensory evidence, conditioned inference globally trumps a full inference strategy that does not have access to the committed high-level interpretation $\hat{C}(m)$ (Figure A1d).

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 A2 shows the results for both overlapping and non-overlapping categories. With unequal category priors, the psychometric curves $p(\hat{C} \mid \theta)$ are shifted away from the categorical boundary (e.g., $p(C_1 \mid \theta)$ is shifted away from the more prevalent category C_2). Compared with the equal prior case, the pattern of local relative accuracy maintains, although it is also shifted (Figure A2b, f). The asymmetric categorical priors exaggerate the differences between posteriors $p(C_1 \mid m)$ and $p(C_2 \mid m)$ 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.

Minimizing L1-Norm Loss

Using the same generative model and model parameters as in the main text (Figure 2a) we formulated optimal estimators for each inference strategy as the median of the posterior distributions Equations 3 and 6, respectively, and performed the same accuracy comparison as in the main text with regard to absolute error. As shown in Figure A3, the results are qualitatively quite similar to the L_2-norm case (Figure 3).

Full Inference With Knowledge of $\hat{C}(m)$ Under Late Noise Conditions

Full inference can also benefit from knowledge of $\hat{C}(m)$. However, in contrast to the conditioned inference strategy it does not consider $\hat{C}(m)$ a correct high-level interpretation but rather a constraint on the potential values of m. Such bottom-up conditioning requires an update of the generative model as shown in Figure A4a. Specifically, it defines a likelihood function $p(\hat{C}(m) \mid m)$ over m that is binary (i.e., in the simplest case a step-function) such that it is one for all values of m that are consistent with the given $\hat{C}(m)$ and zero otherwise. We analyzed the estimation accuracy of such full inference strategy with knowledge of $\hat{C}(m)$ for the case of category-based perception with overlapping categories (Figure 3a). While the local MSE ratio is similar to before (see Figure 3c) overall performance is indeed in favor of the modified full inference strategy, independent on the amount of late noise (Figure A4c).

Note that bottom-up conditioning increases the complexity of the inference problem as it requires an update of the generative model with additional probabilistic information about m (Figure A4a). The additional information essentially reflects the values of the optimal decision boundaries, which are analog quantities and likely not robust to late noise.

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 with the example in the main text (Figure A5).

(*Appendix continues*)

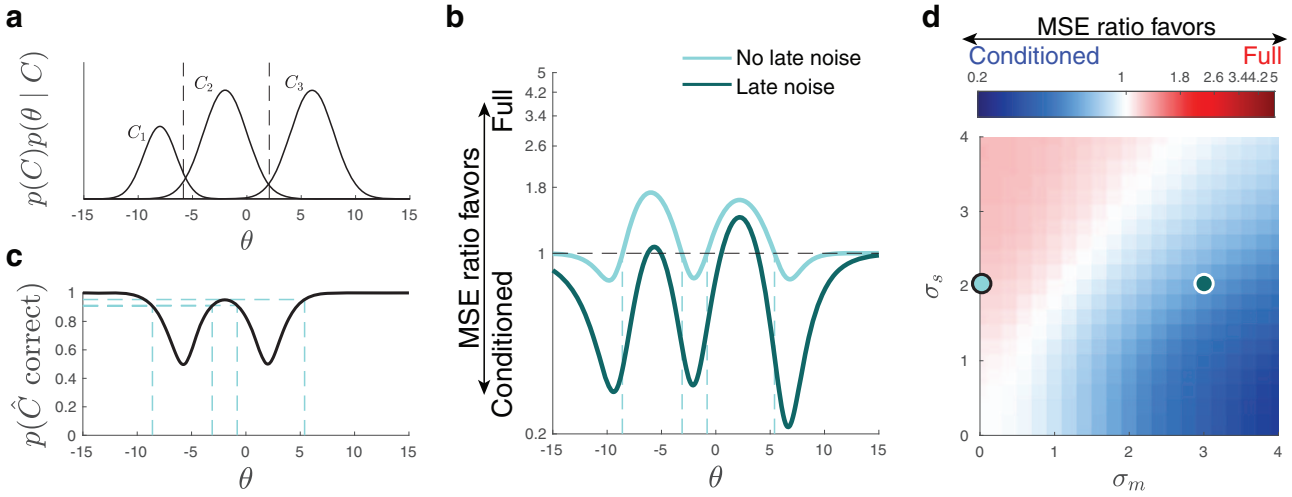

Figure A1. Estimation accuracy for three possible categorical assignments. The panels are similarly organized and labeled as in Figure 3. (a) Categorical feature priors multiplied with the categorical prior for illustration purposes. (b) Local performance comparison between conditioned and full inference. Benefits of conditioned inference emerge where there is reasonably high probability for committing to the correct category. (c) Psychometric curve indicating the probability of assigning θ to the correct category. (d) Global *MSE* ratio as a function of sensory (σ_s) and late noise (σ_m). The two dots correspond to the two noise conditions shown in (b). MSE = mean squared-error. See the online article for the color version of this figure.

(*Appendix continues*)

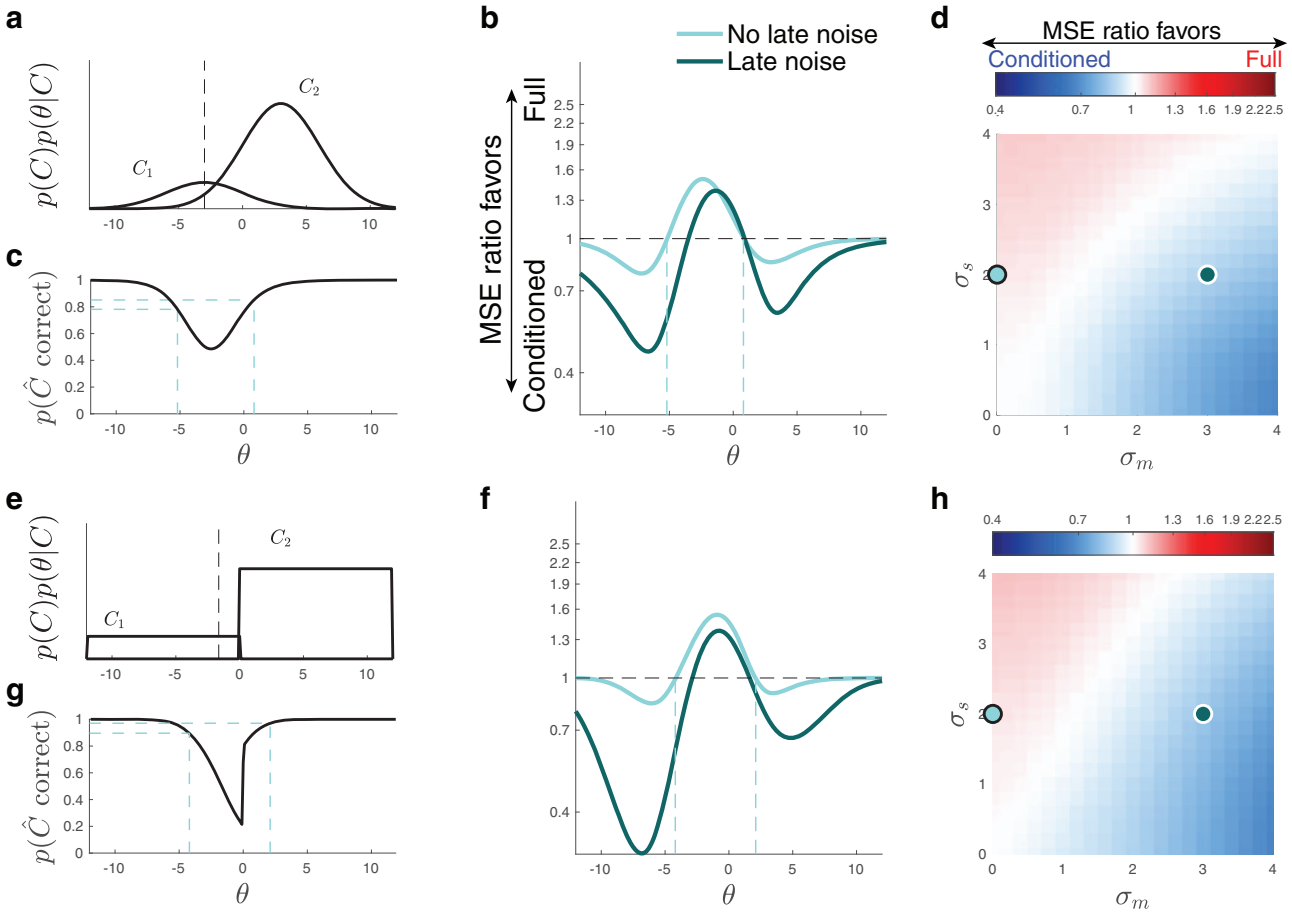

Figure A2. Estimation accuracy for asymmetric category priors. The panels are similarly organized and labeled as in Figure 3. (a) Categorical feature priors multiplied with the categorical prior for illustration purposes. (b) Local performance comparison between conditioned and full inference. Larger benefits are seen when there is reasonable evidence for committing to the less prevalent category. (c) The probability for making a correct categorical assignment $\hat{C}$. (d) Global *MSE* ratio as a function of sensory (σ_s) and late noise (σ_m). (e–h) Same as (a–d) for nonoverlapping categorical feature priors. MSE = mean squared-error. See the online article for the color version of this figure.

(*Appendix continues*)

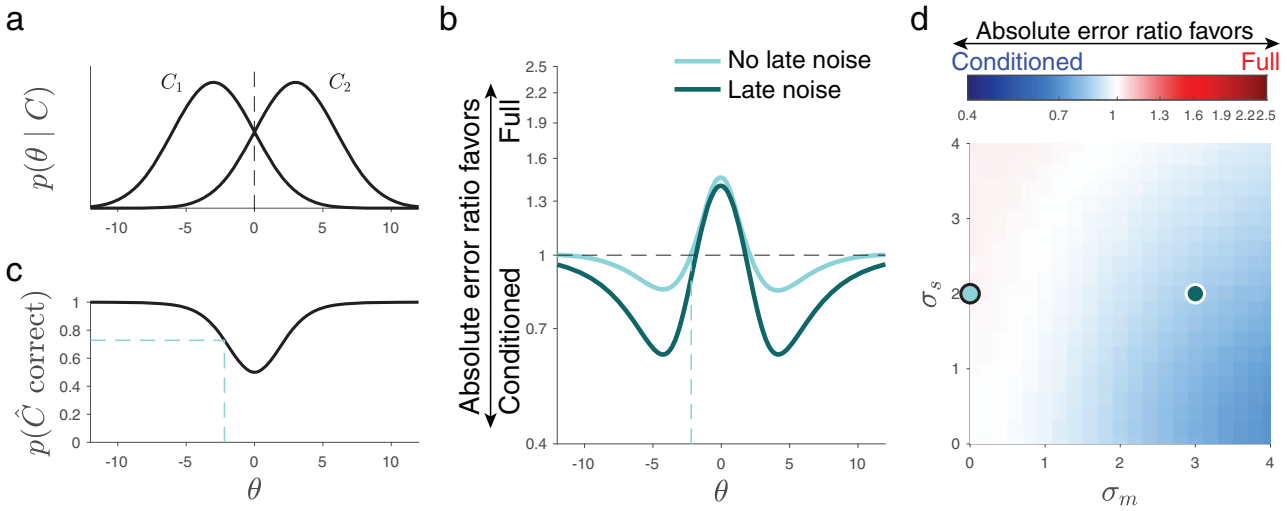

Figure A3. Comparison of the full and conditioned inference strategy when the observer is optimized with regard to a L_1-norm loss function. Accuracy comparison is correspondingly assessed in terms of L_1-norm loss (absolute error). Panels correspond to panels in Figure 3. All other parameters were identical to the example in the main text. MSE = mean squared-error. See the online article for the color version of this figure.

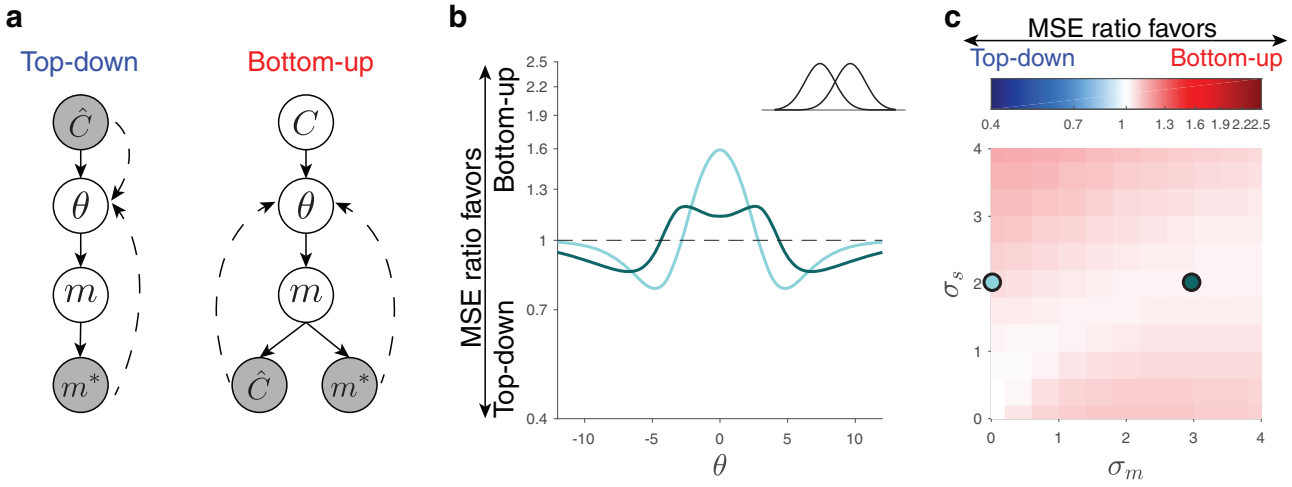

Figure A4. Full inference with knowledge of $\hat{C}(m)$. (a) Graphs for both inference strategies. Conditioned inference (top-down) is the simplest way to extract some of the original sensory measurement information m stored in the categorical interpretation $\hat{C}(m)$ (and thereby protected from corruption by late noise) during the feature inference process. Provided with the same information, one can formulate a modified full inference strategy where the category choice $\hat{C}(m)$ serves as a *bottom-up* conditioning constraint on m. Note, however, that this requires the generative model to be updated with analog information about the decision boundaries. (b) Relative local accuracy between the modified full inference strategy and the conditioned inference strategy shows slightly different behavior compared with the original comparison under late noise (Figure 3b). (c) Bottom-up conditioning allows the full inference strategy to perform globally better for all noise conditions. MSE = mean squared-error. See the online article for the color version of this figure.

(*Appendix continues*)

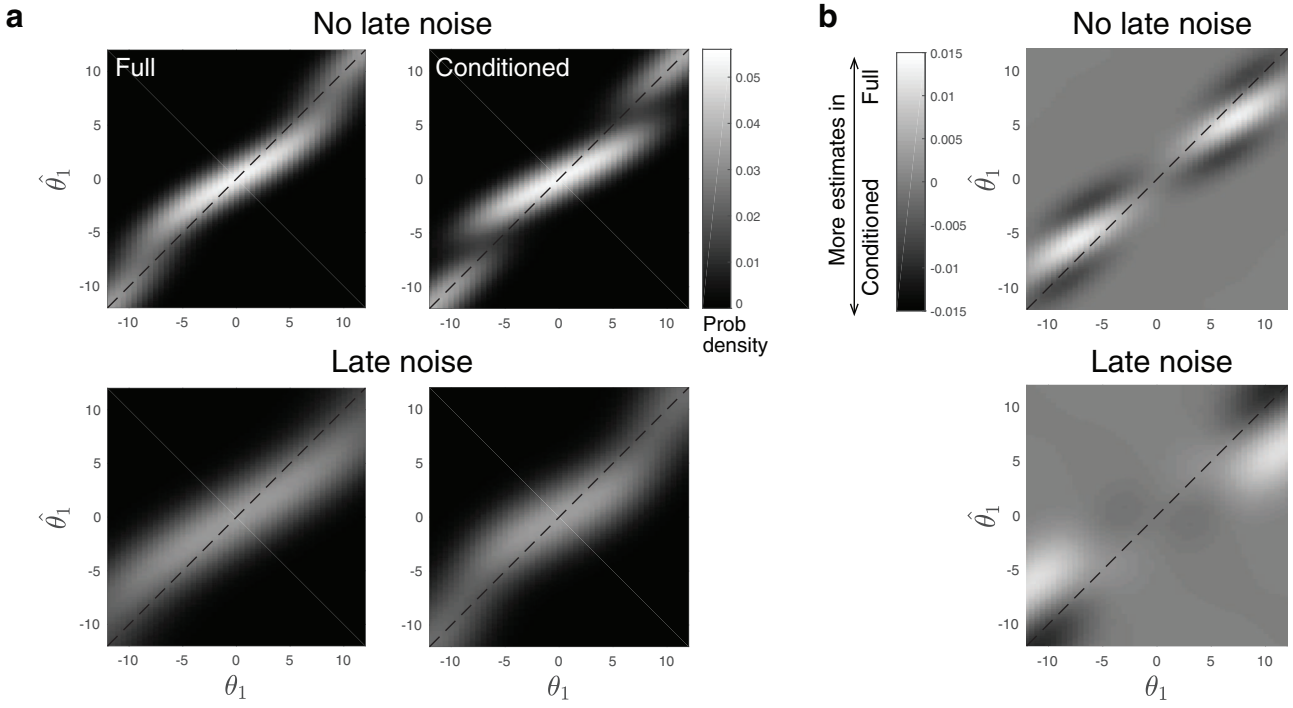

Figure A5. Simulations of causal cue combination with approximately uniform feature prior ($\sigma_p = 100$). (a, b) Estimate distributions and their differences between the two strategies (c–g). Relative *MSE* accuracy. Panels correspond to panels in Figure 5. MSE = mean squared-error. See the online article for the color version of this figure. (*Figure continues on next page.*)

(*Appendix continues*)

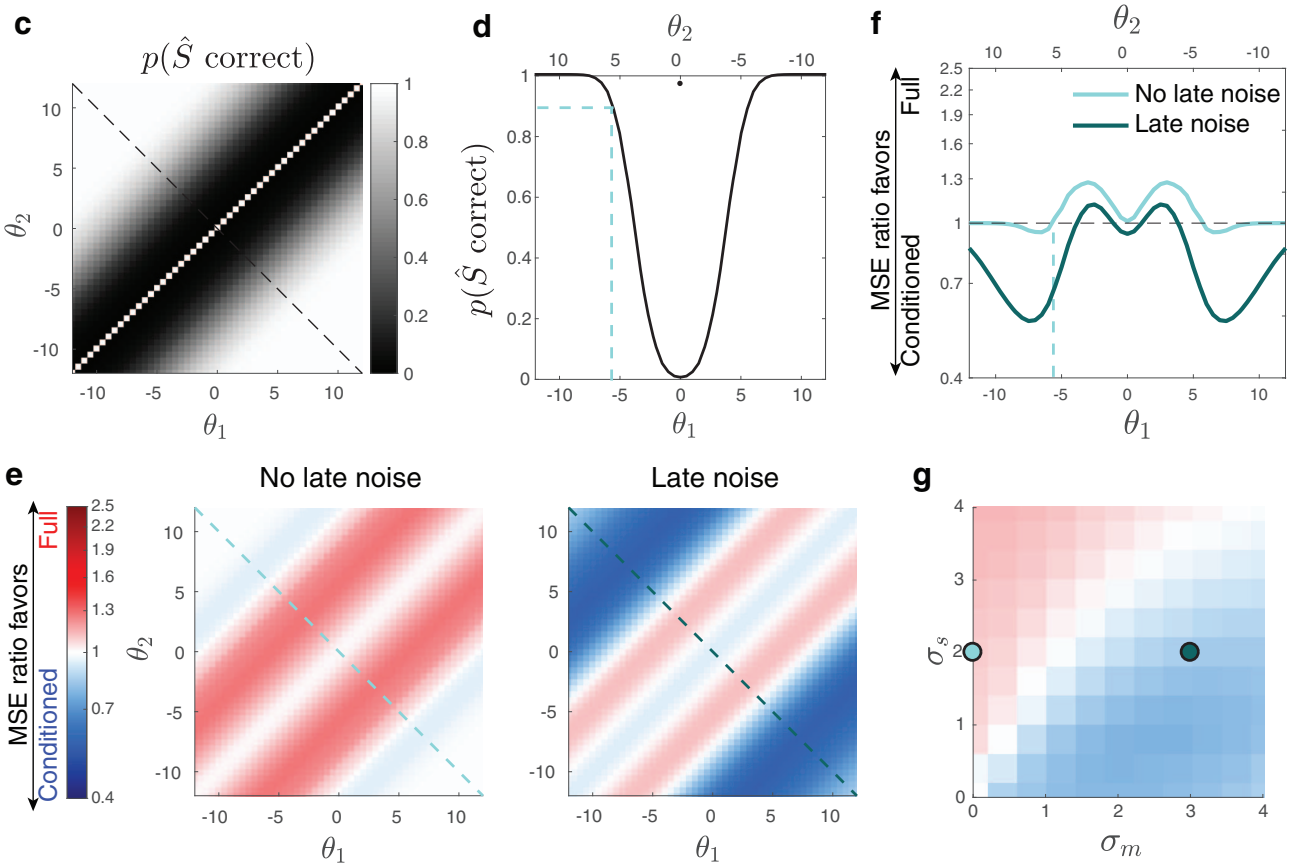

Figure A5. (continued)

Received May 31, 2019
Revision received January 22, 2020
Accepted January 23, 2020 ■