Modeling Causal Generalization with Particle Filters

Randall R. Rojas
Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095
rrojas@stat.ucla.edu

Hongjing Lu
Department of Psychology
University of California, Los Angeles
Los Angeles, CA 90095
hongjing@ucla.edu

Alan Yuille
Department of Statistics, Psychology, and Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
yuille@stat.ucla.edu

Abstract

Psychological experiments have shown that human performance on traditional causal reasoning experiments can be greatly influenced by different pretraining and postraining conditions. In this paper we present a Bayesian theory of sequential learning that captures observed experimental results [1] . We implement our theory using the particle filter algorithm, and show that model selection and model averaging are able to capture the respective effects of pre- and post- training. In addition, we model the highlighting effect observed in [5] using a particle filter algorithm as an approximation to exact statistical inference, in accord with the limited computational capacity of human cognition. We find that the inferential approximation based on particle filters predicts the highlighting effect.

1 Introduction

Human reasoning is adaptive, as exemplified by the reasoner's ability to generalize abstract cause-effect relations from one situation to another. The challenge in understanding causal generalization is to identify how humans acquire and propagate abstract causal knowledge across domains. Causal knowledge includes several key aspects, including causal structure as captured by graphical models of the existence of cause-effect links ([8], [18]), causal power as reflected in the strength of cause-effect links [4], and causal integration rules that model how to combine multiple cause-effect relations [4]. In this paper, we will consider these three types of causal knowledge in the context of causal generalization.

A variety of "rational" models of causal learning have taken probabilistic approaches to explain how people acquire causal knowledge from covariational observations presented in the format of summarized contingency data ([8], [4], and [16]). Despite their success in capturing a variety of causal learning phenomena, these models leave open the question of how a learner can cope with non-summary data. In everyday life, people often receive observations incrementally. For such naturalistic learning situations, sequential models are required to account for the influence of the order of data presentation.

To address this issue, a Bayesian sequential model can be used as an inference engine to capture the propagation of causal knowledge over time. Kalman filtering[9] [10] has been successfully applied in sequential causal learning, and has been used to explain various experimental phenomenon in animal conditioning ([6] and [5]). However, a limitation of previous work using Kalman filtering involves the assumption of Gaussian distributions with a linear-sum causal integration rule to combine multiple cause-effect relations. Many empirical studies have shown that the linear-sum rule is

not able to account for human causal learning with binary variables ([3] and [14]). Accordingly, a more flexible inference model is required to account for a broader range of learning situations. In the present paper, we present a model of sequential causal learning based on particle filtering ([19]), a sequential method used for approximate probabilistic inference over time. This model is used to explore how different causal integration rules can be selected, and how causal knowledge can be propagated with increasing certainty as the number of sequential observations increases.

An additional critical issue concerns how to model the generalization of causal knowledge from one context to another. In the laboratory, researchers have designed behavioral experiments to measure causal generalizability in controlled environments. Beckers et al. (2005) [1] first trained human subjects with certain cue-outcome pairs, such as bacon (cue G) and eggs (cue GH) each paired with a moderate allergic reaction. The combination of of the two cues, bacon and eggs (cue GH) was paired with either a moderate or a strong allergic reaction. The subjects were then transferred to a classic forward blocking paradigm with unrelated cues, such as cheese (cue A) paired with moderate allergy, and cheese and nuts (cue AX) also paired with moderate allergy. Finally, subjects were tested on how likely nuts alone (cue X) was to cause allergy. Human participants provided different ratings on the transfer test for cue X depending on whether cue combination GH has been paired with moderate or strong allergy during the pretraining. Beckers et al.'s (2005) study [1] provided empirical evidence that different pretaining conditions using unrelated causal cues can alter the reasoner's assumptions, and thereby change their subsequent causal inferences. From a computational perspective, the influence of pretraining conditions can be explained in terms of a Bayesian process of model selection, which operates to identify one of the most important aspects of causal knowledge, the causal integration rule, and transfer it to the subsequent inference task. An alternative Bayesian procedure would be model averaging [21], which accepts several causal integration rules with different calculated probabilities, and then averages the inference results across all the possible integration rules. In the final section, we will explore the possibility that particle filtering could serve as an approximation to rational inference while allowing for the limitations on human computational capacity [19].

In this paper, we describe our computational theory in section 2. Section 4 compares human results with model predictions in three experiments. Last, we show how the particle filter approach is able to explain highlighting phenomenon, which has been a challenge to Bayesian sequential learning model using Kalman filters with exact inference.

2 The Computational Theory

This section describes our computational theory. We specify two alternative models which compete to explain the data by model selection or cooperate to explain the data by model averaging, see subsection (2.1). We implement the theory using particle filters as described in subsection (2.2).

2.1 The Models

The experiments specify a sequence of input and output pairs $(\vec{x}_1, d_1), ..., (\vec{x}_t, d_t)$. The input $\vec{x} = (x_1, x_2)$ specifies which cause is present: (i) cause 1 if $x_1 = 1, x_2 = 0$ and, (ii) cause 2 if $x_1 = 0, x_2 = 1$. The output d is a continuous variable. We use capital variables as shorthand for these sequences so that $\vec{X}_t = (\vec{x}_1, ..., \vec{x}_t)$ and $D_t = (d_1, ..., d_t)$.

Both models are parameterized by weight variables $\vec{\omega} = (\omega_1, \omega_2)$ which indicate the strength of the causes x_1, x_2 for causing the effect. We specify a prior $P(\vec{\omega})$ on the weights which is a Gaussian with zero mean and large covariance (making weak assumptions about the initial values of the weights). We specify a temporal prior $P(\vec{\omega}_{t+1}|\vec{\omega}_t)$ which allows the weights to change over time and means that the model is most influenced by the most recent data. The prior and temporal prior are specified by:

$$P(\vec{\omega}_1) = \frac{1}{2\pi\sqrt{|\mathbf{\Sigma_1}|}} \exp\{-(1/2)\vec{\omega}_1^T \mathbf{\Sigma}_1^{-1} \vec{\omega}_1\}, \tag{1}$$

$$P(\vec{\omega}_{t+1}|\vec{\omega}_t) = \frac{1}{2\pi\sqrt{|\mathbf{\Sigma}_2|}} \exp\{-(1/2)(\vec{\omega}_{t+1} - \vec{\omega}_t)^T \mathbf{\Sigma}_2^{-1} (\vec{\omega}_{t+1} - \vec{\omega}_t)\}, \tag{2}$$

where $\mathbf{\Sigma}_1 = \sigma_1^2\mathbf{I}$ and $\mathbf{\Sigma}_2 = \sigma_2^2\mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Hence the weights ω_1, ω_2 are decoupled in the priors. The respective σ_1 and σ_2 values are 0.05 and 0.1.

The likelihood functions are of form:

$$P(d|\vec{\omega}, \vec{x}, M) = \sum_{\vec{R}} P(d|\vec{R}, M)P(\vec{R}|\vec{\omega}, \vec{x}), \tag{3}$$

where $\vec{R}$ are the states of hidden units and M indicates the model.

We define $P(\vec{R}|\vec{\omega}, \vec{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\{-(1/2)(\vec{R} - \vec{\omega} \bigotimes \vec{x})^T \Sigma^{-1} (\vec{R} - \vec{\omega} \bigotimes \vec{x})\}$, where $\vec{\omega} \bigotimes \vec{x} = (\omega_1 x_1, \omega_2 x_2)$.

The models $P(d|\vec{R})$ are of form:

$$P(d|\vec{R}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(d - F(\vec{R}, M))^2/(2\sigma^2)\}, \tag{4}$$

where $F(\vec{R}, M = 1) = R_1 + R_2$ for the first model and $F(\vec{R}, M = 2) = R_1 \frac{e^{R_1/T}}{e^{R_1/T} + e^{R_2/T}} + R_2 \frac{e^{R_2/T}}{e^{R_1/T} + e^{R_2/T}}$ for the second model. The first model is the linear-sum model [6] and the second model is a noisy-max model [16] which is a generalization of the noisy-or model [4]. Figure 1 illustrates the two generative models, which employ different combination rules.

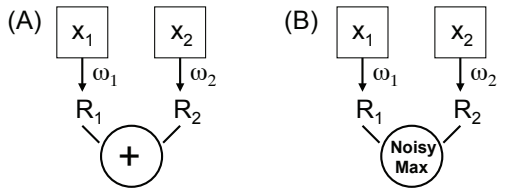

Figure 1: An illustration of the generative models. The different models combine R1 and R2 in different ways, a linear-sum (A) or a noisy-max (B), to yield the output effect R.

2.2 Inference by Particle Filtering

We use Bayes-Kalman to update the probabilities $P(\vec{\omega}_t|D_t, \vec{X}_t, M)$ when we receive new data. In the rest of the section we will drop the dependencies on $\vec{X}_t$ and M to simplify the notation. Bayes-Kalman specifies that we update weights by a prediction stage followed by a measurement stage:

$$P(\vec{\omega}_{t+1}|D_t) = \int d\vec{\omega}_t P(\vec{\omega}_{t+1}|\vec{\omega}_t)P(\vec{\omega}_t|D_t) \quad \text{(prediction)} \tag{5}$$

$$P(\vec{\omega}_{t+1}|D_{t+1}) = \frac{P(d_{t+1}|\vec{\omega}_{t+1})P(\vec{\omega}_{t+1}|D_t)}{P(d_{t+1}|D_t)} \quad \text{(measurement)}. \tag{6}$$

We implement these equations using particle filters ([7], [15]). Previous methods in the cognitive science literature are unsuited for this problem. Dayan and Kakade [6] used Kalman's algebraic update equations for the means and covariance of $P(\vec{\omega}_t)$, but this cannot be applied to our second model which is non-Gaussian. Lu *et al.* (2008) [17] represented the distributions using a fixed lattice in $\vec{\omega}$ space, but this becomes problematic for the models described here (high accuracy requires a very dense lattice which leads to an extremely slow algorithm). By contrast, particle filters sample the space adaptively and are more efficient. (We validated particle filters by showing that they agree with these other methods when applicable).

Particle filters approximates distributions like $P(\vec{\omega}_t|D_t, M)$ by a set of discrete particles $\{\vec{\omega}_t^\mu : \mu \in \Gamma\}$. This enables us to approximates quantities such as $\int d\vec{\omega} g(\vec{\omega}_t)P(\vec{\omega}_t|D_t)$ by $(1/|\Gamma|) \sum_{\mu\in\Gamma} g(\vec{\omega}_t^\mu)$ for any function $g(.)$.

We initialize by drawing samples $\{\vec{\omega}_1^\mu : \mu \in \Gamma\}$ from the prior distribution $P(\vec{\omega})$. This is straightforward since the prior is a Gaussian.

Then we proceed recursively following the prediction and measurement stages of the Bayes-Kalman filter. Let $\{\vec{\omega}_t^\mu : \mu \in \Gamma\}$ be the set of particles representing $P(\vec{\omega}_t|D_t)$ at time t. Then we sample from the Gaussian distribution $P(\vec{\omega}_{t+1}|\vec{\omega}_t^\mu)$ for each μ to give a new set of particles $\{\vec{\omega}_t^\mu : \mu \in \Gamma\}$ which represents $P(\vec{\omega}_{t+1}|D_t)$.

Next we compute the importance weights $\lambda^\mu = P(d_{t+1}|\vec{\omega}_{t+1}^\mu)$ and normalize them to obtain $\bar{\lambda}^\mu = \lambda^\mu/(\sum_\mu \lambda^\mu)$. Then we re-sample with replacement from the set $\{\vec{\omega}_{t+1}^\mu : \mu \in \Gamma\}$ using probability $\bar{\lambda}^\mu$. This gives new set $\{\vec{\omega}_{t+1}^\nu : \nu \in \Gamma\}$ of particles which represent $P(\vec{\omega}_{t+1}|D_{t+1})$.

To compare to experiments, we need to measure the *model evidence* $P(D_t)$ for each model and to estimate the *mean values* of the weights $\bar{\vec{\omega}}_t = \int d\vec{\omega}_t \vec{\omega}_t P(\vec{\omega}_t|D_t)$. We compute these from the particles as follows.

The mean values are approximated by the average $(1/|\Omega|)\sum_{\mu\in\Gamma}\vec{\omega}_t^\mu$.

The model evidence is expressed as $P(d_t|D_{t-1})P(d_{t-1}|D_{t-2})...P(d_1)$. We evaluate each term $P(d_{t+1}|D_t) = \int d\vec{\omega}_{t+1} P(d_{t+1}|\vec{\omega}_{t+1})P(\vec{\omega}_{t+1}|D_t)$ by $P(d_{t+1}|D_t) = \frac{1}{|\Gamma|}\sum_{\mu\in\Gamma} P(d_{t+1}|\vec{\omega}_{t+1})$.

The simulations are run using 6000 particles since beyond this value, even when performing numerous Monte Carlo runs (over 100 each with 6000 particles), the results do not show any significant variation in their outcomes. Our simulation of highlighting effects is instead performed with 1000 particles. Figure 2 illustrate the change of particle filters in a standard forward blocking paradigm over training trials.

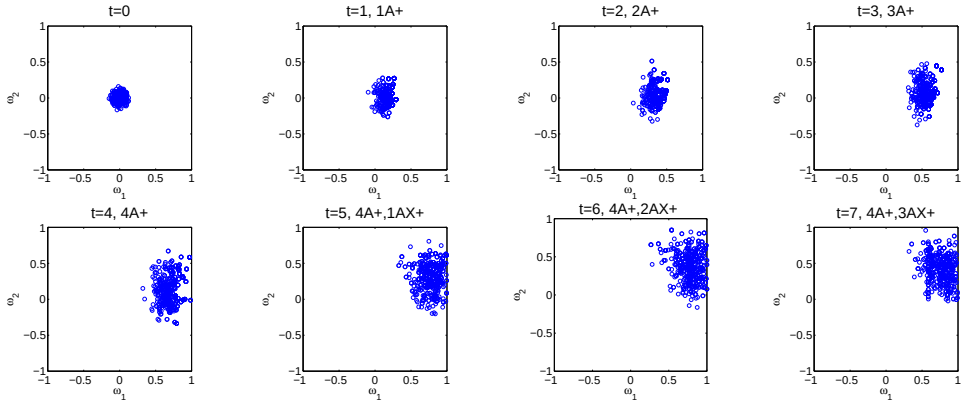

Figure 2: Particle filters in the simulation with forward blocking paradigm as a function of training trials.

3 Comparison of Simulation Results with Experiments

Based on experimental results from Beckers et al. ([1] -see Experiments 2, 3 and 4), we simulate their study of the effect of pre- and post- training on human judgements. Pretraining is simulated for (a) forward blocking (Experiment 2), (b) backward blocking (Experiment 3) and (c) release from overshadowing (Experiment 3). The effect of posttraining is modeled for forward blocking only (Experiment 4). Details of the experimental designs and experiments are given in [1] and previous simulation results for the effect of pretraining in the forward blocking case are discussed in [17].

3.1 Modeling of Pretraining Effect using Model Selection

The experiments conducted by Beckers et al. [1] consist of four different food cues: A, X, K, and L and allergic reactions to these cues are measured as moderate (+) or strong (++). In our notation, G+, H+, and GH+ correspond to $(x_1, x_2) = (1, 0), (0, 1)$, and $(1, 1)$ respectively. The reaction strengths + and ++ correspond to $O = 1$ and $O = 2$ respectively. Human experiment consists of three phases:

(1) pretraining, (2) elemental training and (3) compound training. Pretraining is performed with food cues, namely, G and H, and the subsequent training phases use different food cues. Pretraining conditions can be either additive (G+ $\rightarrow$ H+ $\rightarrow$ GH++) or subadditive (G+ $\rightarrow$ H+ $\rightarrow$ GH+). Cues K and L are only present in phase 3 and therefore serve as control cues.

Table 1 below shows the experimental design for the forward blocking experiment (Exp 2 in [1]) and the backward blocking experiment (Exp 3 in [1]). As discussed in greater detail in [17], using the pretraining trials in Phase 1, we perform model selection, as shown in Figure 3.The simulation results show that the linear-sum model is selected if the pretraining is additive (i.e., G+, H+, GH ++), because the corresponding ratio is below the threshold, whereas the noisy-MAX model is selected if the pretraining is sub-additive (i.e., G+, H+, GH+), because the corresponding ratio is above the threshold. Next, we adopt the Bayesian sequential model to update posterior distributions of the weights ω for each cue presented in Phases 2 and 3. To compare our simulation results with the human ratings, we compute the mean of each ω.

Table 1: Design summary for human pretraining study in Beckers et al. (2005) Experiment 2 & 3. The numerical values indicate the number of trials and + indicates the presence of the outcome effect.

Experiment	**Blocking Paradigm**	**Pretraining Phase 1**	**Compound Phase 2**	**Elemental Phase 3**	**Test**
Exp 2	Forward blocking additive	8G+/8H+/8GH++	8A+	8AX+/8KL	A, X, K, L
	Forward blocking subadditive	8G+/8H+/8GH+	8A+	8AX+/8KL	A, X, K, L
Exp 3	Backward blocking additive	8G+/8H+/8GH++	8AX+/KL+	8A+	A, X, K, L
	Backward blocking subadditive	8G+/8H+/8GH+	8AX+/KL+	8A+	A, X, K, L

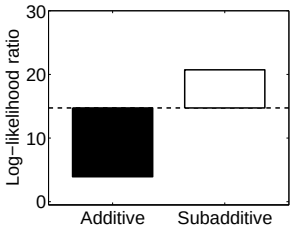

Figure 3: Log-likelihood ratios for the noisy-MAX model relative to the linear-sum model for experiment by Beckers et al. (2005). Black bars indicate the ratio for the additive group; white bars for the sub-additive group. The dashed line indicates the threshold for model selection. These simulation results are in good agreement with experimental findings (see subsection 3.1).

Figure 4 shows the mean causal rating for each cue. In the top panel, the left plot shows the human ratings in forward blocking experiment (Exp 2) by Beckers et al [1], in which black bars indicate the mean rating for additive pretraining group; white bars for sub-additive pretraining group. The right plot shows the predicted ratings based on the selected model for each group. Black bars indicate the mean ω based on the linear-sum model, which gives a good fit for the human means in the additive group. White bars indicate the mean ω based on the noisy-MAX model, which give a good fit for the human means in the sub-additive group. The simulation results are in good agreement with the results for humans. The linear-sum model generates accurate predictions for the additive group: the mean weight for X is much lower than weights for the control cues K and L, indicating blocking of causal learning for cue X. In contrast, the noisy-MAX model gives accurate predictions for the sub-additive group: the mean weight for X is about the same as the weights for the control cues K and L, consistent with absence of blocking for X.

The bottom panel in Figure 4 show the results from human and model in backward blocking experiment (Exp 3) by Beckers et al [1]. Again, the model predictions agree with human performance well. A pretraining effect still preserves for both human and model, although the effect is much weaker than in forward blocking experiment (Exp 2 in [1]).

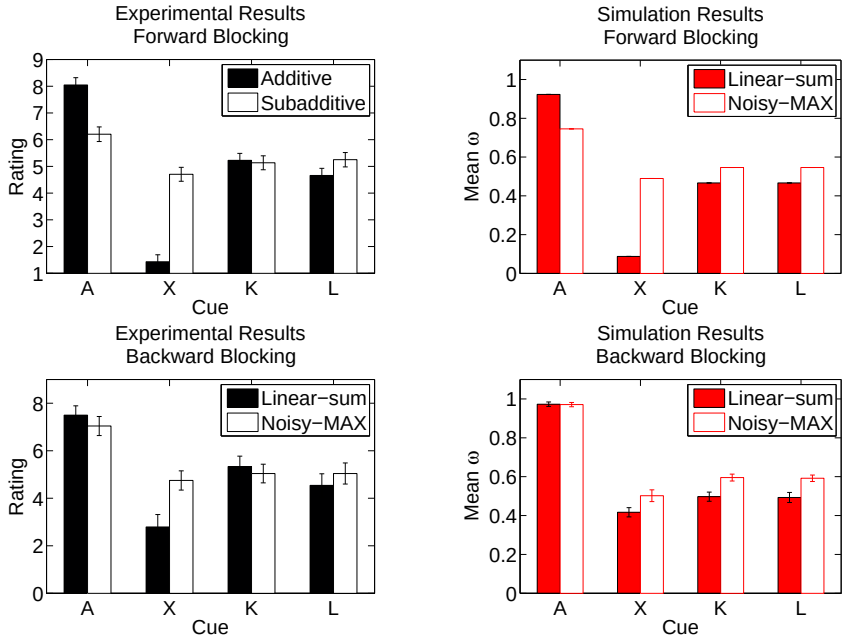

Figure 4: Left, Human ratings by Beckers et al. (2005); Right, Predicted ratings based on the selected model for each group. Top panel, forward blocking experiment (Exp. 2 in [1]); Bottom panel, backward blocking experiment (Exp. 3 in [1]). For further details see subsection 3.1.

3.2 Modeling of Posttraining Effect using Model Averaging

Experiment 4 in the study conducted by Beckers and his colleagues [1] reported that information about outcome additivity have an impact on blocking even if it is presented after the blocking training phases. As shown in Table 2, Phases 1 and 2 correspond to the elemental and compound training phases respectively with cue A and X, but Phase 3 is the posttraining phase with different cues (i.e. cue G and H). After the posttraining phase, human subjects were asked to evaluate the causal power for cue A and X. In the other words, the design in Exp4 is identical as it in Exp1 described in section 3.1, except reversing the order of the actual blocking training and the additivity training, effectively turning the additiveity manipulation in a posttraining instead of a pretraining procedure.

In the first blocking training phases, we assume that humans update posterior distributions of causal strengths for models, linear-sum and noisy-max. When the posttraining phase is provided, model averaging is performed to combine the estimates of causal strengths from the two models as

$$\langle \omega \rangle = P(D|M_1)\bar{\omega}_{M_1} + P(D|M_2)\bar{\omega}_{M_2} \tag{7}$$

where D is the data, M_1 and M_2 represent the linear-sum and noisy-max gate models respectively, and $P(D|M_i)$ is the 'evidence' for each model from observations in the posttraining phase. $\bar{\omega_{M_i}}$ is the estimated mean value of causal strength using each model from observations in the first two training phases.

Figure 5 shows our results for the posttraining experiment (Exp 4 in [1]). We can see that model averaging is able to capture the posttraining effects qualitatively, and correctly predict a weaker posttraining effect than the pretraining effect described in section 3.1.

4 Highlighting

In this section, we will demonstrate the use of particle filtering to explain another causal learning phenomenon, the highlighting effect reported by Kruschke [11] [12] [13]. Tables 3 describes a

Table 2: Design summary for human posttraining study in Beckers et al. (2005) Experiment 4. The numerical values indicate the number of trials and + indicates the presence of the outcome effect.

Experiment	Group	Elemental Phase 1	Compound Phase 2	Posttraining Phase 3	Test
Exp 4	Forward blocking additive	A+	8AX+/8KL+	8G+/8H+/8GH++	A, X, K, L
	Forward blocking subadditive	A+	8AX+/8KL+	8G+/8H+/8GH+	A, X, K, L

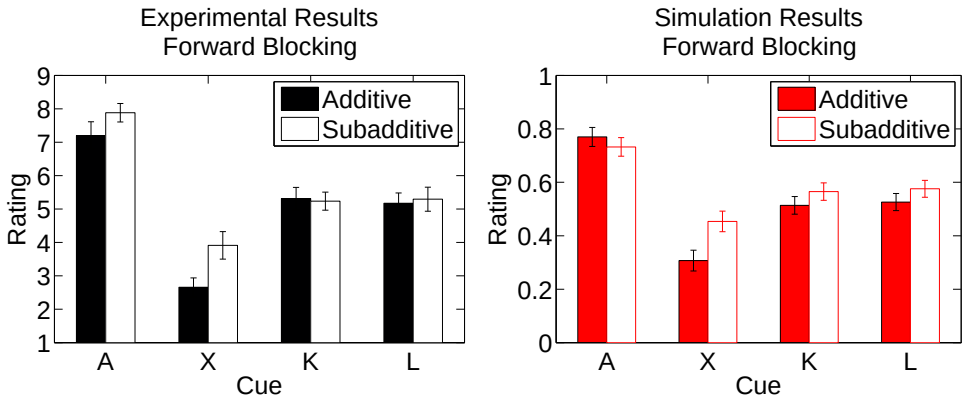

Figure 5: (A) Mean causal rating for each cue based on human subjects in Experiment 4 (Posttraining) of Beckers et al. [1] (see their Figure 5 p247) and (B) mean causal weights produced by the model as a function of different cues in the design of Experiment 4 (Beckers et al., 2005). See section 3.2

canonical experimental design. The learner first see 15 trials of cue A and B associated with outcome O_1, and 5 trials of cue A and C associated with outcome O_2; then in the second phase, the order of the training sessions is reversed. In the subsequent test phase, observers are asked to predict the outcome (O_1 or O_2) when showing cue A only and cue B and C together. The highlighting design equalizes the long-run base rates of the two outcomes, and the frequency of cue-outcome pairs (e.g., 20 trials in total for AC with O_1 and AB with O_2). Humans show a strong tendency to choose outcome O_1 for cue A, but a lower probability to choose O_1 for cue B and C.

Kruschke [12] developed a locally Bayesian learning model to account for the highlighting effect, by combining local Bayesian updating between layers and attention control through back-propagation. In contrast, Daw et al. [5]argued that the highlighting effect could be explained by computation limitations of human cognitive system. Accordingly to their account, human observers conduct inferential learning by an approximation to statistical models, such as Kalman filtering. Daw et al. employed a rational model based on Kalman filtering with a linear-sum rule. With exact statistical inference, this rational model is unable to predict the highlight effect. However, including inferential approximations in their model using reduced-rank approximations was able to explain the highlighting effect.

In this section, we show that particle filtering, as an approximation to statistical inference, is able to predict the highlighting effect [19]. The advantage of using particle filtering is that it makes it possible to control the degree of approximation to exact statistical inference, as simulations with a very large number of particles are closer to the rational inference model. To be consistent with the representations used by the two models described in the previous paragraph, we assume that outcome O_1 is indicated when a binary reward value is 1, and outcome O_2 is indicated when the reward value is 0. We thus focus the model on the case of multiple causes and a single effect, rather than extending it to the case of multiple causes and multiple effects. Given that the outcome variable is binary, we adopt two generative models, linear-sum [5] and noisy-logic [4] [20], in the simulation of particle filtering.

The left plot in figures 6 shows the learning curve of causal strengths for each cue and the combination of cue B and C as a function of training trials. Based upon the learned causal strengths for each cue, the model can predict how likely outcome O_1 will be chosen for each cue or cue combination. The highlighting effect is revealed by the difference between Cue A and BC predictions. We find that this difference is reduced with increases in the number of particles employed in the simulation, a result which agrees with the finding in Daw et al. [5], summarized above.

The right plot in figures 6 shows the probability of choosing O_1 for cue A and cue B&C as a function of different generative models, linear-sum and noisy-logic. Both models yield highlighting effects, namely higher $P(O_1)$ for cue A than for cue B&C, although the effect is weaker for the noisy-logic model.

Table 3: Highlighting Design by Daw et al. 2007 [5]

Blocking Paradigm	**Training Phase 1**	**Training Phase 2**	**Test**
Highlighting	$15 \times (\mathrm{AB} \rightarrow O_1)$ $5 \times (\mathrm{AC} \rightarrow O_2)$	$5 \times (\mathrm{AB} \rightarrow O_1)$ $15 \times (\mathrm{AC} \rightarrow O_2)$	$\mathrm{A} \rightarrow ?O_1$ $\mathrm{BC} \rightarrow ?O_2$

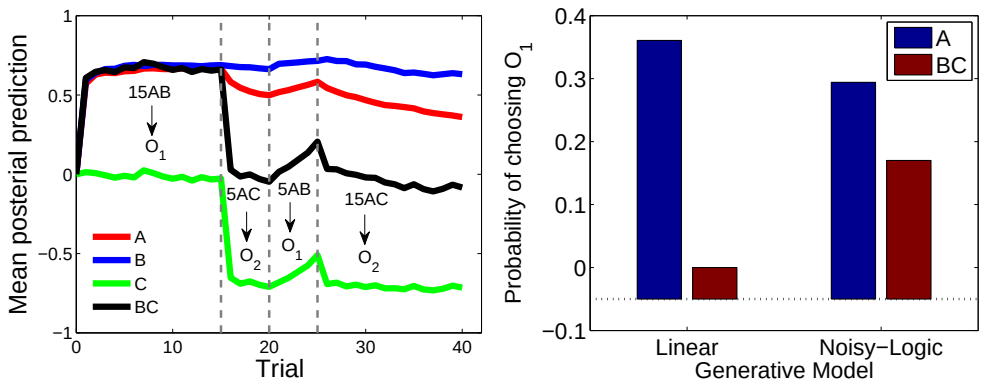

Figure 6: Highlighting results. Left, learning curve of causal strength when using linear-sum model. The highlighting effect is revealed by the difference between Cue A and BC predictions. With the increase of the number of particles, the highlighting effect will reduce. Right, comparison of highlighting effect between linear-sum model and noisy-logic model. See section 4.

5 Conclusions

The Bayesian theory of sequential causal learning described in the present paper provides a unified explanation of important learning phenomena, using the implementation of a particle filter as an approximation to exact statistical inference. In particular, the theory accounts for influences of pretraining on subsequent learning as well as influences of posttraining on previous learning, with completely different stimuli. The key assumption is that learners have available multiple generative models, each reflecting a different integration rule for combining the influence of multiple causes. In particular, when the outcome is a continuous variable, humans have tacit knowledge that multiple effects may have a summative impact on the outcome (linear-sum model). Alternatively, the outcome may be effectively saturated at a level approximated by the weight of the strongest individual cause (noisy-MAX). Using standard Bayesian model selection, the learner selects the model that best explains the pretraining data, and then continues to favor the most successful model during subsequent learning with different cues. In other situations, the learner uses both models to perform causal learning, but is able to retrospectively re-evaluate the estimations from different models when extra information about integration rules is provided by post-training with different cues. This post-training effect can be explained by model averaging.

Finally, we propose that particle filter simulation could be a good candidate to mimic the limitations of computational capacity (in particular, working memory resources) in the human cognitive system. Accordingly, particle filtering may serve as a computationally realistic approximation to rational inference. This model of approximate inference is able to explain the highlighting effect. In future work, we hope to extend the model to more complex causal networks that include multiple causes and multiple effects.

References

[1] Beckers, T. and J. De Houwer. "Outcome Additivity and Outcome Maximality Infleunce Cue Competition in Human Causal Learning". Journal of Experimental Psychology. Learning, Memory and Cognition. Vol. 11. No. 2. pp 238-249. 2005.

[2] Beckers, T., R. R. Miller, J. De Houwer, and K. Urushihara."Reasoning Rats: Forward Blocking in Pavlovian Animal Conditioningis Sensitive to Constraints of Causal Inference". Journal of Experimental Psychology. General. Vol. 135. No. 1. pp 92-102. 2006.

[3] Buehner, M. J., Cheng, P. W., & Clifford, D. (2003). From covariation to causation: A test of the assumption of causal power. Journal of Experimental Psychology: Learning, Memory, and Cognition, 29, 1119-1140.

[4] Cheng., P. W. (1997) From covariation to causation: A causal power theory. Psychological Review, 104, 367405.

[5] Daw, N., Courville, A. C., & Dayan, P. (2007). Semi-rational Models of Conditioning: The Case of Trial Order. In M. Oaksford and N. Chater (Eds.). *The probabilistic mind: Prospects for rational models of cognition*. Oxford: Oxford University Press.

[6] Dayan, P. & Kakade, S. (2000). Explaining away in weight space. In T. K. Leen et al., (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 451-457). Cambridge, MA: MIT Press.

[7] Doucet, A., de Freitas, N., & Gordon, N. J. (2001) *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, New York, NY, 2001.

[8] T. L. Griffiths, and J. B. Tenenbaum. Structure and strength in causal induction. Cognitive Psychology, 51, 334-384, 2005.

[9] Kalman, R. E. A new approach to linear filtering and prediction problems. *Transactions of the ASME-Journal of Basic Engineering*, 82:35-45, 1960

[10] Kamin, L.J. (1969). Predictability, surprise, attention, and conditioning. In B. A. Campbell & R. M. Church (Eds.), *Punishment and aversive behavior*. New York: Appleton-Century-Crofts, 279-296.

[11] Kruschke, J. K. (2001). Cue competition in function learning: Blocking and highlighting. Presented at the 3rd International Conference on Memory, July 2001, Valencia, Spain.

[12] Kruschke, J. K., (2006). Locally Bayesian Learning with Applications to Retrospective Revaluation and Highlighting. Psychol Rev, Vol. 113, No. 4, pp. 677-699

[13] Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. Learning & Behavior, 36(3), 210-226.

[14] Liljeholm, M., & Cheng, P. W. (2007). When is a cause the same? Coherent generalization across contexts. Psychological Science, 18, 1014-1021

[15] Liu, J. S. (2001) *Monte Carlo strategies in scientific computing.* Springer-Verlag, New York, NY.

[16] Lu, H., Yuille, A., Liljeholm, M., Cheng, P. W., & Holyoak, K. J. (2008). Bayesian generic priors for causal learning. Psychological Review, 115(4), 955-984.

[17] Lu, H., Rojas, R. R., Beckers, T., & Yuille, A. (2008). Sequential causal learning in humans and rats. Proceedings of the Twenty-ninth Annual Conference of the Cognitive Science Society.

[18] Pearl, J. (2000). Causality. Cambridge, UK: Cambridge University Press.

[19] Wood, F., & Griffiths, T. L. (2007). Particle filtering for nonparametric Bayesian matrix factorization. *Advances in Neural Information Processing Systems 19.*

[20] Yuille, A. L., & Lu, H. (2007). The noisy-logical distribution and its application to causal inference. Advances in neural information processing systems, Vol. 20. Cambridge, MA: MIT Press.

[21] Erven, T. V. , Grnwald, P. D. & Rooij, S. de. (2007). Catching up Faster in Bayesian Model Selection and Model Averaging. Advances in Neural Information Processing Systems (NIPS), 20.