Journal of Mathematical Psychology 112 (2023) 102728

Contents lists available at ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

A statistical foundation for derived attention[☆]

Samuel Paskewitz [a,*], Matt Jones [b]

[a] Department of Psychiatry, Children's Hospital, Anschutz Medical Campus, University of Colorado Denver, United States of America
[b] Department of Psychology and Neuroscience, University of Colorado Boulder, United States of America

ARTICLE INFO

Article history:
Received 9 May 2022
Received in revised form 10 November 2022
Accepted 13 November 2022
Available online 8 December 2022

Keywords:
Bayesian cognitive modeling
Attention
Learning
Category learning
Derived attention

ABSTRACT

According to the theory of derived attention, organisms attend to cues with strong associations. Prior work has shown that – combined with a Rescorla–Wagner style learning mechanism – derived attention explains phenomena such as learned predictiveness, inattention to blocked cues, and value-based salience. We introduce a Bayesian derived attention model that explains a wider array of results than previous models and gives further insight into the principle of derived attention. Our approach combines Bayesian linear regression with the assumption that the associations of any cue with various outcomes share the same prior variance, which can be thought of as the inherent importance of that cue. The new model simultaneously estimates cue–outcome associations and prior variance through approximate Bayesian learning. A significant cue will develop large associations, leading the model to estimate a high prior variance and hence develop larger associations from that cue to novel outcomes. This provides a normative, statistical explanation for derived attention. Through simulation, we show that this Bayesian derived attention model not only explains the same phenomena as previous versions, but also retrospective revaluation. It also makes a novel prediction: inattention after backward blocking. We hope that further development of the Bayesian derived attention model will shed light on the complex relationship between uncertainty and predictiveness effects on attention.

1. Introduction

Selective attention is a key part of learning theory. Many experimental results can be explained by supposing that organisms pay attention to some cues and ignore others, and that attention changes as a result of experience. For decades, researchers have attempted to mathematically model this interplay between selective attention and memory formation. One prominent class of models assumes that organisms form direct associations between cues and predicted outcomes (e.g. food, shock, category labels); selective attention acts to re-scale cues and/or control learning rates (e.g. Esber & Haselgrove, 2011; Frey & Sears, 1978; Kruschke, 2001) In this paper, we shall focus on a sub-class of such models based on the principle of *derived attention*.

Derived attention theories assume that the attention paid to a cue is proportional to the size of its association weights. A cue with large association weights (whether positive or negative) will thus be attended to, while one with small association weights will be ignored. Attention is thus *derived* from existing associations. Derived attention models have been proposed in several forms (Esber & Haselgrove, 2011; Frey & Sears, 1978; Le Pelley, Mitchell, Beesley, George, & Wills, 2016), but a review article by Le Pelley et al. (2016) has brought the theory into particular prominence.

Le Pelley et al. (2016) show how derived attention can explain many important learning results, despite its simplicity compared to other attention learning rules (e.g. Kruschke, 2001). For example, consider the learned predictiveness effect (Le Pelley & McLaren, 2003; Lochmann & Wills, 2003). During initial training, certain cues (A, B, C, and D) are correlated with category labels, while others (V, W, X, and Y) are not. In a later transfer stage, people pay more attention to the previously relevant cues than the previously irrelevant ones, even though all cues perfectly predict the new categories (see Fig. 2(a)). Derived attention explains this result by noting that the predictive cues develop larger associations in the first stage and hence greater attention during the second. Similar reasoning explains why people pay a great deal of attention to cues associated with large monetary rewards (Anderson, Laurent, & Yantis, 2011; Le Pelley, Mitchell, & Johnson, 2013, see Fig. 4(a)) and little attention to redundant (blocked) cues (Beesley & Le Pelley, 2011; Kruschke & Blair,

[☆] This research was supported by grant R21MH120741 from the National Institute of Mental Health, United States of America and grant 2020906 from the National Science Foundation, United States of America.

[*] Corresponding author.
E-mail addresses: samuel.paskewitz@cuanschutz.edu (S. Paskewitz), mcj@colorado.edu (M. Jones).

https://doi.org/10.1016/j.jmp.2022.102728
0022-2496/© 2022 Elsevier Inc. All rights reserved.

2000, see Fig. 3(a)). However, not all attentional phenomena are explained by derived attention (Medin & Edelson, 1988; Swan & Pearce, 1988).

In this paper we offer a normative foundation for derived attention by reformulating it in terms of Bayesian inference, and show how this significantly expands the scope of the theory. The new model is based on an insight of Le Pelley et al. (2016): "The idea that attention toward a cue increases to the extent that it predicts a high-value outcome – attention is determined by associative strength – is very intuitive, and is consistent with the idea that attention goes to cues that are known to be significant" (page 1129). We develop this insight into a probabilistic generative model of the organism's environment, and then derive an online variational Bayesian regression algorithm. The resulting algorithm resembles previous derived attention models and explains the same experimental results, including learned predictiveness (Lochmann & Wills, 2003), inattention after blocking (Beesley & Le Pelley, 2011; Kruschke & Blair, 2000), and value-based attention (Anderson et al., 2011; Le Pelley et al., 2013).

The new Bayesian derived attention model can also explain retrospective revaluation effects (which are characterized by learning about absent cues), a class of phenomena that Le Pelley et al.'s (2016) derived attention model cannot handle. Backward blocking is one example of retrospective revaluation (Kruschke & Blair, 2000; Shanks, 1985, see Fig. 5(a)). In a backward blocking task, participants receive paired cue training (A.X → I, B.Y → II) followed by further training with only one cue from each pair (A → I, B → II). This continued single cue training weakens the associative strength of the dropped cues (X and Y) even though they are not present during it. Le Pelley et al.'s (2016) model cannot explain this or other retrospective revaluation effects because its learning rule (based on Rescorla & Wagner, 1972), only updates the value of cues that are present during a trial. It is thus not adequate to describe backward blocking or other retrospective revaluation phenomena. However, the new Bayesian derived attention model produces these effects through an explaining away mechanism: if further training shows that A and B are sufficient to explain the outcomes, then X and Y's weights decrease toward zero (Dayan & Kakade, 2001). Moreover, casting derived attention into a Bayesian framework produces a novel prediction that goes beyond both Le Pelley et al.'s (2016) version of derived attention and Dayan and Kakade's (2001) Bayesian regression model: cues suffer a loss of attention after being subject to backward blocking.

2. Step by step toward a Bayesian derived attention model

Because the new Bayesian derived attention model is somewhat complex, we shall build up to it step by step by describing a series of simpler models. At each stage of the process we shall first describe a generative model of the learner's environment, i.e. a set of implicit probabilistic assumptions held by the learner (computational level model). We then describe a corresponding inference algorithm suitable for modeling learning (algorithmic level model). In this way we shall gradually build up the necessary machinery for the Bayesian derived attention model.

2.1. Basic linear regression

The first step toward building our new learning model is to describe the learner's environment statistically. Assume that there are n_x cues (i.e. stimulus features). The cue vector for trial t is denoted $x_{1,t}, \ldots, x_{n_x,t} = x_t$. These cues ($x_{i,t}$) can be either continuously valued (e.g. brightness of a light) or coded as $\{0, 1\}$ to indicate the presence or absence of a discrete stimulus or property. The experiments described in this paper use discrete cues. We also have n_y outcomes, denoted $y_{1,t}, \ldots, y_{n_y,t} = y_t$. In category learning experiments such as those described in this paper, each outcome ($y_{j,t}$) corresponds to a category label and is coded as 1 if category j is the correct answer for the current stimulus and 0 otherwise. Each response option corresponds to a category label (outcome) and only one response is correct on each trial, so feedback is complete on each trial and does not depend on the learner's choices. We use $\text{lrn}_{j,t}$ as an indicator of whether learning about outcome j should occur on trial t; $\text{lrn}_{j,t}$ equals zero during test stages without feedback or when j is not a possible outcome in the current stage. Similarly, $\text{psb}_{j,t}$ indicates whether outcome j is possible during the current trial ($= 1$ if yes and $= 0$ if no). The same models can be applied to Pavlovian conditioning, with the difference that there is typically only one outcome – the unconditioned stimulus – and hence only one set of association weights ($n_y = 1$; e.g. Rescorla & Wagner, 1972), with the outcome always considered possible with full feedback (i.e. $\text{psb}_t = 1$ and $\text{lrn}_t = 1$).[1]

An example will be help to explain the notation. Consider the simplified learned predictiveness design described in Fig. 2(a), which is based on Lochmann and Wills (2003) and Le Pelley, Suret, and Beesley (2009). The nature of the category labels and cues depends on the experiment's cover story. For example in Lochmann and Wills (2003) participants pretended to be allergists, with the categories corresponding to different allergic reactions and the cues to different foods a hypothetical patient might eat. Let us label the elements of the cue (x_t), outcome (y_t), outcome possibility (psb_t), and outcome feedback (lrn_t) vectors in the following manner:

$$x_t = [\text{cue A; cue B; cue X; cue Y}] \tag{1}$$

$$y_t = [\text{outcome I; outcome II; outcome III; outcome IV}] \tag{2}$$

$$\text{psb}_t = [\text{outcome I; outcome II; outcome III; outcome IV}] \tag{3}$$

$$\text{lrn}_t = [\text{outcome I; outcome II; outcome III; outcome IV}] \tag{4}$$

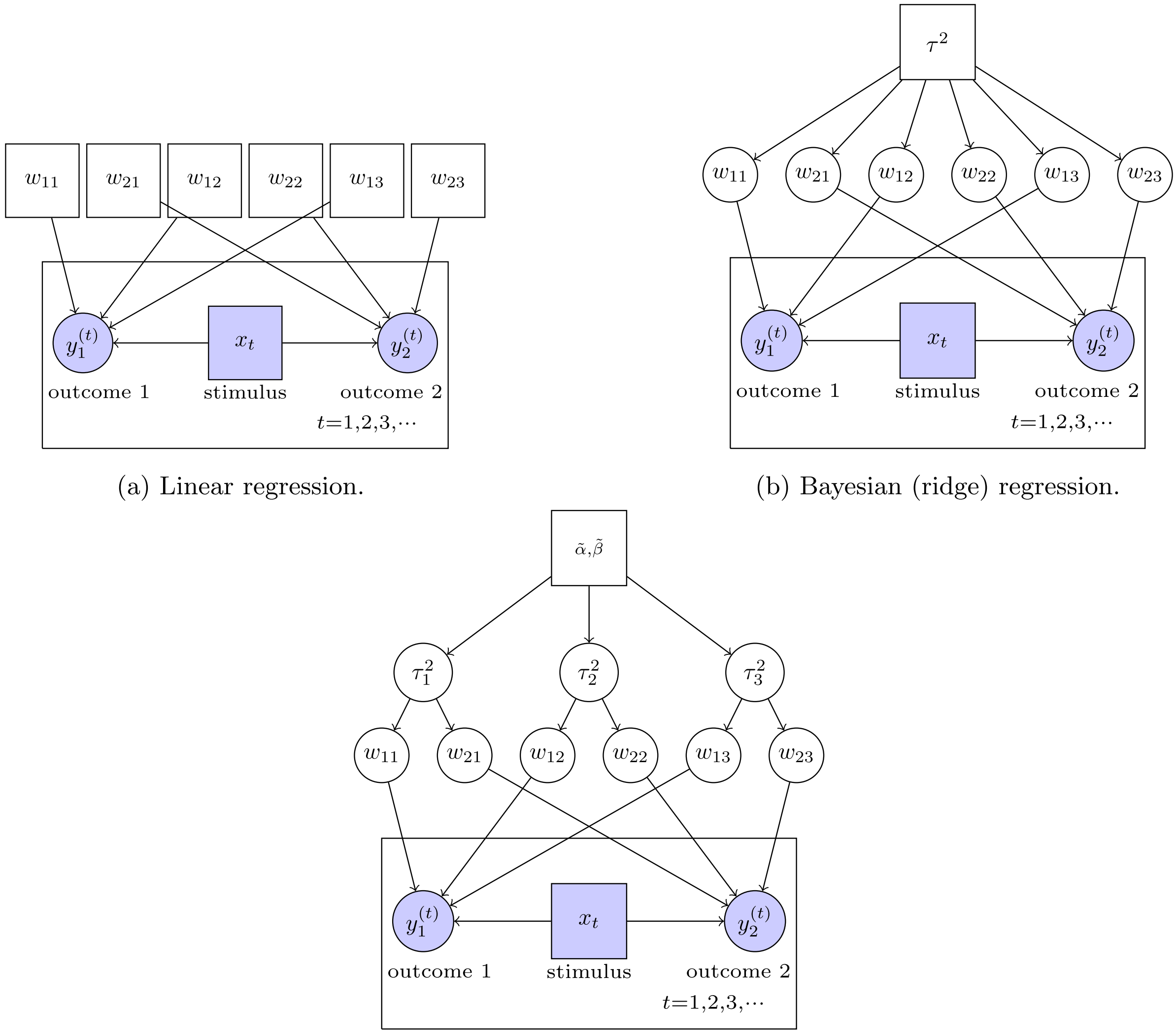
On A.X trials, we have $x_t = [1; 0; 1; 0]$. In the initial stage (Relevance) on A.X trials $y_t = [1; 0; 0; 0]$ (category I is correct), while in the second stage (Transfer) on A.X trials $y_t = [0; 0; 1; 0]$ (category III is correct). On B.X trials, $x_t = [0; 1; 1; 0]$ and $y_t = [0; 1; 0; 0]$ (category II is correct). The vectors x_t and y_t are defined similarly for the other trial types. The outcome possibility indicator (psb_t) is equal to $[1; 1; 0; 0]$ for all trials in the Relevance stage and to $[0; 0; 1; 1]$ throughout the Transfer and Test stages. The outcome feedback indicator (lrn) is equal to $[1; 1; 0; 0]$ in the Relevance stage, to $[0; 0; 1; 1]$ in the Transfer stage; and to $[0; 0; 0; 0]$ in the Test stage. Hopefully this example will clarify the notation to the reader.

We now describe a probabilistic model of the organism's environment. We begin with the standard assumption that learning is a linear regression problem. The cues combine linearly with association weights (w_{ji}) to produce each outcome j:

$$y_{j,t} = \sum_i x_{i,t} w_{j,i} + \epsilon_{j,t} \tag{5}$$

$$= x_t^T w_j + \epsilon_{j,t} \text{ for each outcome } (j) \text{ such that } \text{psb}_{j,t} = 1 \tag{6}$$

[1] A note on notation: the value of outcome j on time step t is denoted $y_{j,t}$ the total outcome vector for time step t is denoted y_t, the vector of outcome j values across all time steps is denoted y_j, and the totality of all observed outcome values across both outcomes and time steps is denoted y. Notation for cues (x) is similar. All vectors such as x_t, y_t, x_i, y_j etc. should be considered *column* vectors.



(a) Linear regression.

(b) Bayesian (ridge) regression.

(c) Bayesian derived attention.

Fig. 1. Generative statistical models discussed in the text. Square nodes indicate fixed variables and circular nodes indicate random variables. Filled nodes represent variables that are directly observed by the learner, while unfilled nodes represent variables that are not directly observed.

where $\epsilon_{j,t} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ is Gaussian noise. In other words, $y_{j,t} \sim \mathcal{N}(x_t^T w_j, \sigma^2)$. This is simply a linear regression model repeated n_y times, once for each outcome j.[2] If we assume that the true association weights (w_{ji}) are fixed but unknown quantities then this model is ordinary frequentist regression, as typically taught in university statistics courses (see Fig. 1(a) for a graphical depiction). We can use $\hat{w}$ to denote the learner's estimates of w from observed data, and $\hat{y}$ to denote the learner's outcome predictions given a set of cues (x_t). Given $\hat{w}$, $\hat{y}$ is computed as follows:

$$\hat{y}_{j,t} = \sum_i x_{i,t} \hat{w}_{j,i} \tag{7}$$

$$= x_t^T \hat{w}_j \text{ for each outcome } (j) \text{ such that } \mathrm{psb}_{j,t} = 1 \tag{8}$$

2.1.1. The Rescorla–Wagner model

In order to obtain an online estimate of w that can be constantly revised as new observations are made, we can use stochastic gradient ascent on the likelihood to obtain an approximate maximum likelihood estimate. We denote these estimated association weights as $\hat{w}$; they represent the associations as they exist in the learner's mind. The resulting inference algorithm is the commonly used Delta Rule/Rescorla–Wagner model (Gluck & Bower, 1988; Rescorla & Wagner, 1972, see Algorithm 1 in Appendix A for pseudocode):

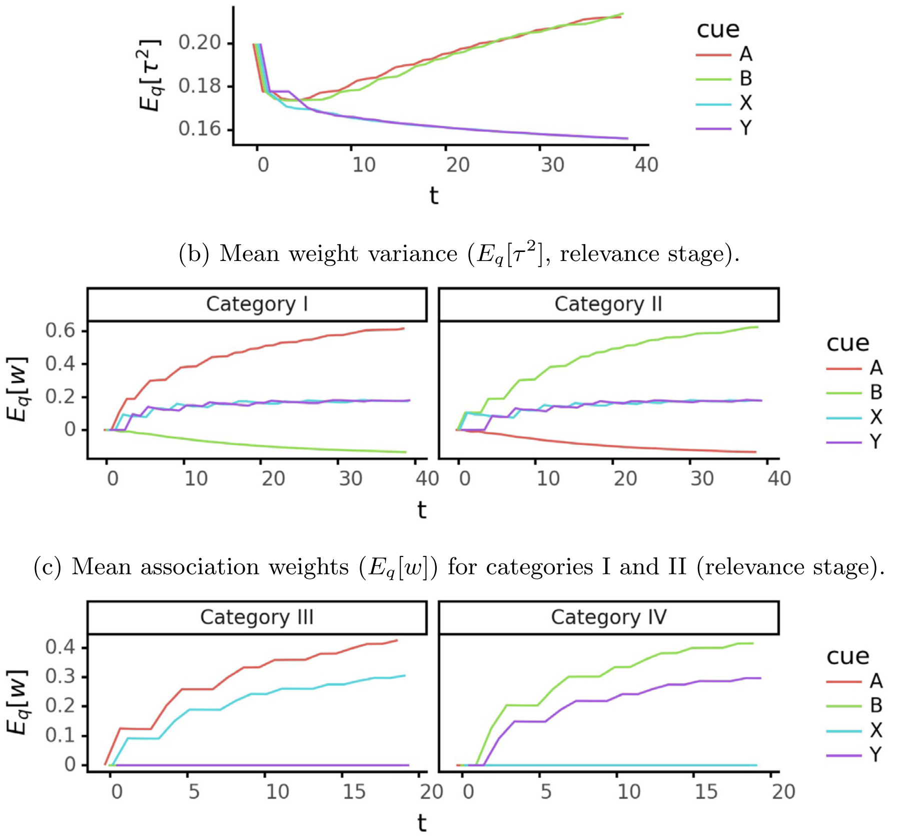
$$\hat{w}_{j,i} \leftarrow \hat{w}_{j,i} + \lambda_i (y_{j,t} - \hat{y}_{j,t}) \text{ for } j \text{ such that } \mathrm{lrn}_{j,t} = 1 \tag{9}$$

where $\lambda_i = \lambda_{\mathrm{par}} x_i$ is the learning rate (λ_{par} is a fixed model parameter) and $\hat{y}_{j,t} = x_t^T \hat{w}_j$ is the predicted outcome. Learning is driven by prediction error, i.e. the difference between the outcome observed ($y_{j,t}$) and the outcome predicted ($\hat{y}_{j,t}$). When prediction error is small ($\hat{y}_{j,t} \approx y_{j,t}$) then not much learning occurs. This explains blocking, i.e. the fact that very little is learned about a redundant cue is introduced when the outcome is already predicted by previous conditioning to another cue (Kamin, 1968). However the basic Rescorla–Wagner model cannot explain

[2] Because the outcomes in category learning experiments are nominal (category labels), strictly speaking it would be more appropriate to use something like probit regression than linear regression as described here. We have in fact implemented a version of probit regression (using a mean field approximation of the latent variable) in the full Bayesian derived attention model described below: it did not produce substantially different results. Thus for the sake of simplicity we shall use linear regression, which after all is consistent with existing models of human classification learning (e.g. Gluck & Bower, 1988; Kruschke, 1996).

Relevance		Transfer		Test		Simulated Test Responses	
Cues	Category	Cues	Category	Cues	Response Pattern	p(III)	p(IV)
A.X	I	A.X	III	A.Y	$p(\text{III}) > p(\text{IV})$	0.65	0.35
A.Y	I	B.Y	IV	B.X	$p(\text{IV}) > p(\text{III})$	0.35	0.65
B.X	II						
B.Y	II						

(a) A simple learned predictiveness design (c.f. Le Pelley et al., 2009; Lochmann & Wills, 2003).



(b) Mean weight variance ($E_q[\tau^2]$, relevance stage).

(c) Mean association weights ($E_q[w]$) for categories I and II (relevance stage).

(d) Mean association weights ($E_q[w]$) for categories III and IV (transfer stage).

Fig. 2. Simulation 1 (learned predictiveness).

selective attention or retrospective revaluation phenomena such as those described above.

2.1.2. The Derived Attention Model of Le Pelley et al. (2016)

The derived attention model of Le Pelley et al. (2016) is a modified version of the Rescorla–Wagner model (see Algorithm 2 in Appendix A for pseudocode). Each cue has its own learning rate (λ_i) that changes with time, representing selective attention. The attention paid to each cue is proportional to the current size of its estimated association weights ($\sum_j |\hat{w}_{j,i}|$). Thus, cues with large associations receive more attention. This model explains a great deal of data, but does not explicitly provide any normative justification for derived attention, i.e. explain why it might be useful to organisms. In the following sections we develop a Bayesian derived attention model that derives a similar attention mechanism through probabilistic inference.

2.2. Bayesian regression (with conjugate prior)

We shall now make the leap from frequentist to Bayesian statistics. Instead of viewing each association weight (w_{ji}) as unknown but fixed, w_{ji} is now a random variable. Rather than computing a single estimate ($\hat{w}_{ji}$), the learner's beliefs about the value of w_{ji} are represented by a probability distribution which is updated based on new information. For example if $w_{ji} \sim \mathcal{N}(m, s^2)$, then the learner believes that m is the most likely value of w_{ji} and that there is a 95% probability that w_{ji} lies in the interval $(m - 1.96s, m + 1.96s)$. With one additional assumption (a conjugate prior), the resulting Bayesian inference algorithm closely resembles the Rescorla–Wagner/Delta Rule model but has additional capabilities.

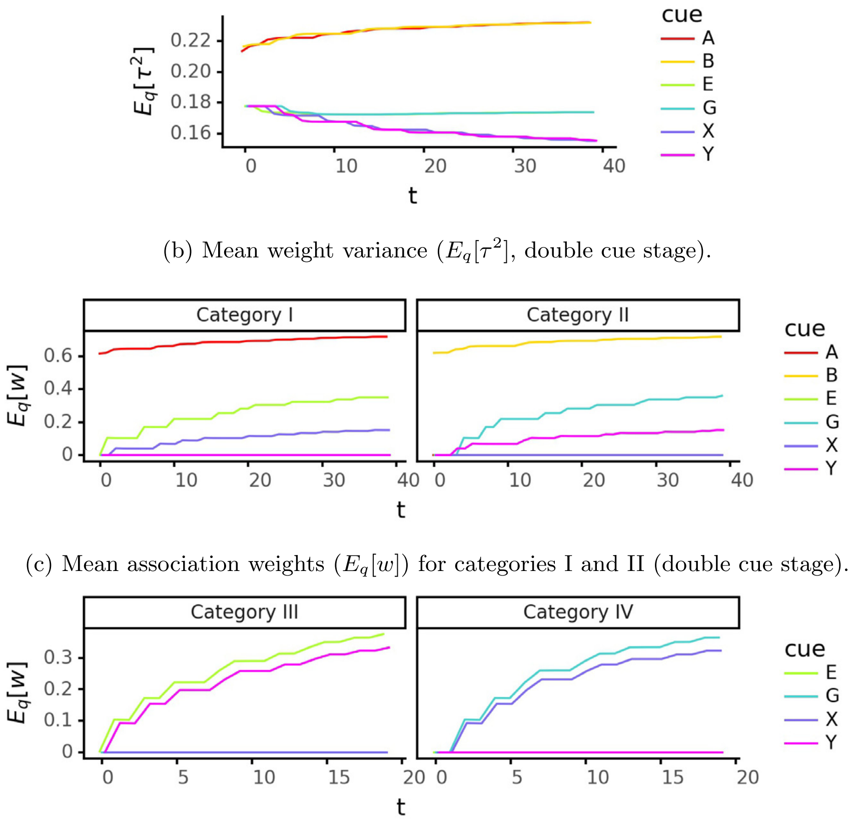
We must give w_j a prior distribution, which represents the learner's beliefs previous to observing any actual data (i.e. before starting the learning task). Choosing a conjugate (in this case multivariate normal) prior makes inference very simple, as we explain below. Thus

$$w_j \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma}) \text{ for each outcome } (j) \tag{10}$$

where $\tilde{\mu}$ and $\tilde{\Sigma}$ are the prior mean and covariance matrix respectively. In particular, we shall choose $\tilde{\mu} = 0$ and $\tilde{\Sigma} = \tau^2 I$, where τ^2 is the prior variance for each weight. We can interpret this choice of prior in the following terms: before the task begins, the environment produces each association weight (w_{ji}) via an independent draw from a normal prior distribution with mean 0

Single Cue		Double Cue		Transfer		Test		Simulated Test Responses	
Cues	Category	Cues	Category	Cues	Category	Cues	Response Pattern	p(III)	p(IV)
A	I	A.X	I	E.Y	III	E.X	$p(\text{III}) > p(\text{IV})$	0.55	0.45
B	II	B.Y	II	G.X	IV	G.Y	$p(\text{IV}) > p(\text{III})$	0.45	0.55
		E.F	I						
		G.H	II						

(a) A simple inattention after blocking design (c.f. Le Pelley et al., 2007).



(b) Mean weight variance ($E_q[\tau^2]$, double cue stage).

(c) Mean association weights ($E_q[w]$) for categories I and II (double cue stage).

(d) Mean association weights ($E_q[w]$) for categories III and IV (transfer stage).

Fig. 3. Simulation 2 (inattention after blocking).

(so that it is *a priori* equally likely to be positive or negative) and variance τ^2. In other words, $w_{ji} \sim \mathcal{N}(0, \tau^2)$. Fig. 1(b) represents this generative model graphically.

Prior variance (τ^2) affects how large the model's posterior weight estimates will be; a smaller value of τ^2 will produce smaller estimated weights, an effect called *shrinkage*. Because the prior mean is zero, prior variance (τ^2) represents the model's expectation for the size (squared magnitude) of the average weight:

$$\begin{aligned}\tau^2 &= V[w_{j,i}] = E[(w_{j,i} - E[w_{j,i}])^2] \\ &= E[w_{j,i}^2] = \text{average size of cue–outcome associations}\end{aligned} \tag{11}$$

When τ^2 is small and the true weights are large, the model will require more data to overcome its prior bias toward small weights. Conversely, if τ^2 accurately reflects the average weight size than the model will learn accurate weight estimates more quickly.

Learning consists of updating the prior distribution ($p(w_j)$) with information from each successive observation (y_j) to produce a posterior distribution ($p(w_j|y_j)$) using Bayes' rule:

$$p(w_j|y_j) \propto p(y_j|w_j)p(w_j) \tag{12}$$

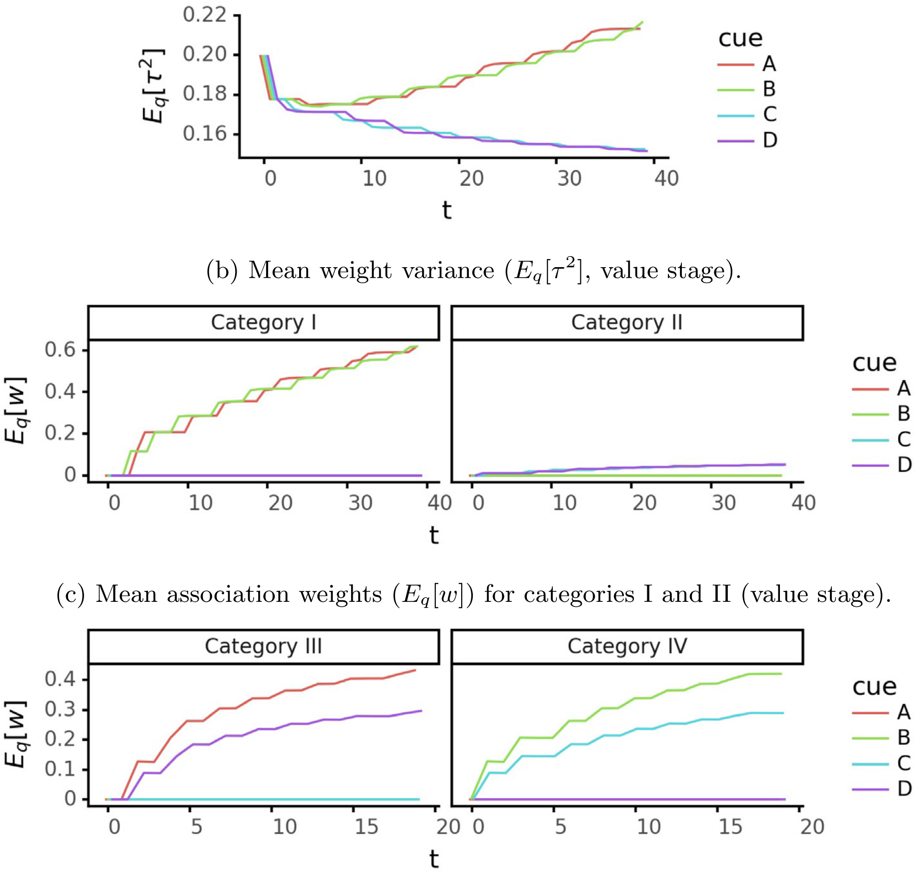
Because our prior is conjugate to the likelihood, the resulting posterior distribution has the same form as the prior (multivariate normal):

$$w_j|y_j \sim \mathcal{N}(\mu_j, \Sigma_j) \tag{13}$$

where μ_j and Σ_j are termed the *hyperparameters* of w_j. The posterior mean (μ_j) represents the learner's best point estimate of w_j. The diagonal elements of the posterior covariance matrix (Σ_j) are posterior variances for each element of w_j, and can be interpreted as the learner's level of confidence in the estimate (smaller posterior variance means more confidence). The off-diagonal elements of Σ_j represent covariance between different cues' weights. When two cues are repeatedly paired together, this covariance will become negative (this can be seen as cue–cue associative learning). This negative covariance between the cues' weights represents the fact that the two cues offer competing explanations for any associated outcomes: the larger one cue's weight is, the smaller the one cue's weight must be (Dayan & Kakade, 2001). Thus, evidence that the first cue has a large weight causes the learner to infer that the second cue has a small weight. The posterior distribution after each observation serves as the prior distribution for the next observation, producing an exact inference algorithm that naturally accommodates trial-by-trial learning.

Value			Transfer		Test		Simulated Test Responses	
Cues	Category	Reward	Cues	Category	Cues	Response Pattern	p(III)	p(IV)
A	I	100	A.D	III	A.C	p(III) > p(IV)	0.66	0.34
B	I	100	B.C	IV	B.D	p(IV) > p(III)	0.34	0.66
C	II	10						
D	II	10						

(a) A simple value effect design (c.f. Le Pelley et al., 2013).



(b) Mean weight variance ($E_q[\tau^2]$, value stage).

(c) Mean association weights ($E_q[w]$) for categories I and II (value stage).

(d) Mean association weights ($E_q[w]$) for categories III and IV (transfer stage).

Fig. 4. Simulation 4 (value effect).

We can write this inference algorithm in two equivalent ways. In Algorithm 3, the updates to the hyperparameters are written explicitly. This form is very similar to the Rescorla–Wagner/Delta Rule model (Algorithm 1); the update for μ_j (analogous to the point estimate $\hat{w}_j$) has the same form. In fact, Algorithm 3 is very similar to the Kalman filter regression algorithm that has been used in psychology to model learning (Dayan & Kakade, 2001). The difference is that in our generative model we assume that w_j remains constant from trial to trial, whereas the Kalman filter assumes that w_j gradually drifts from its original value via a normally distributed random walk. Thus Algorithm 3 is equivalent to Kalman filter regression with a random walk standard deviation of zero.[3]

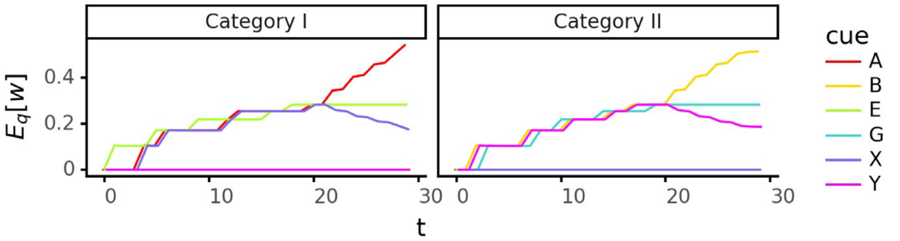
In order to develop the Bayesian derived attention model, it is convenient to write the inference algorithm in a different but equivalent form (Algorithm 4). Here we write the conjugate prior/posterior distribution of w_j in the form of an exponential family, i.e. in terms of its natural hyperparameters ($\psi_{0,j} = \Sigma_j^{-1}\mu_j$ and $\psi_{1,j} = \Sigma_j^{-1}$) and sufficient statistics ($T_{0,j} = \sum_t \frac{x_t y_{j,t}}{\sigma^2}$ and $T_{1,j} = \sum_t \frac{x_t x_t^T}{\sigma^2}$). Learning consists simply of updating the sufficient statistics by summing across trials. To obtain the natural hyperparameters $\psi_{0,j}$ and $\psi_{1,j}$ (and hence the conventional hyperparameters μ_j and Σ_j) for the posterior, we add the sufficient statistics to the prior values for the natural hyperparameters. This way of writing the calculation makes it much simpler to change the prior distribution later in the course of learning than with the conventional form (Algorithm 3), something that will be critical for the algorithmic approximation we propose below for the Bayesian derived attention model. The equivalence between Algorithms 3 and 4 is worked out in Appendix B.

As a model of learning, Bayesian regression has some additional capabilities compared to the Rescorla–Wagner model. Because both Bayesian regression and the Rescorla–Wagner model have additive predictions and learning based on prediction error, both can explain cue additivity effects such as blocking and overshadowing (Rescorla & Wagner, 1972). However, Bayesian regression has two attributes that the Rescorla–Wagner model

[3] So far, we have not been able to incorporate the random walk into our Bayesian derived attention model.

Double Cue		Single Cue		Test		Simulated Test Responses	
Cues	Category	Cues	Category	Cues	Response Pattern	p(I)	p(II)
A.X	I	A	I	E.Y	p(I) > p(II)	0.64	0.36
B.Y	II	B	II	G.X	p(II) > p(I)	0.36	0.64
E.F	I						
G.H	II						

(a) A simple backward blocking design (c.f. Shanks, 1985).



(b) Mean association weights ($E_q[w]$). The single cue stage begins on trial 20.

Fig. 5. Simulation 4 (backward blocking).

lacks: cue-specific learning rates and learning for absent cues. Both of these follow from the fact that the Bayesian model tracks uncertainty in the weights (Σ) in addition to point estimates (μ). Repeated observations of a cue decrease the posterior variance of its weights (a diagonal element of Σ), reducing its learning rate: this *familiarity effect* explains phenomena such as latent inhibition (Gershman, 2015).[4] In addition, Bayesian regression models can explain retrospective revaluation phenomena, because observations about one cue can be informative about others through the off-diagonal elements of Σ, which represent the posterior covariance between cue weights (Dayan & Kakade, 2001). As we shall show below, the Bayesian derived attention model has the same property but can also explain selective attention effects.

2.3. Bayesian derived attention

To turn the Bayesian regression model described above into a Bayesian derived attention model, we make some additional assumptions about τ^2 (the prior variance of w). Our approach somewhat resembles Automatic Relevance Determination (Neal, 1996). Instead of assuming that τ^2 is constant across all weights ($w_{j,i}$), we instead suppose that there is a separate variance (τ_i^2) for all of the weights pertaining to a given cue:

$$w_{j,i} \overset{iid}{\sim} \mathcal{N}(0, \tau_i^2) \text{ for each outcome } (j) \tag{14}$$

τ_i^2 expresses the average size of all of cue i's weights (c.f. Eq. (11)), or in other words the importance or relevance of cue i (larger weights mean more impact on outcomes). Unlike the standard Bayesian regression model described above, we do not treat the values of τ_i^2 as known: τ_i^2 for each cue is a random variable subject to Bayesian inference.[5] In particular, we give each cue's τ_i^2 an inverse gamma prior (or equivalently give the precision τ_i^{-2} a gamma prior):

$$\tau_i^2 \sim \text{InvGamma}(\tilde{\alpha}, \tilde{\beta}) \text{ for each cue } (i) \tag{15}$$

where $\tilde{\alpha}$ and $\tilde{\beta}$ are prior hyperparameters. The complete generative model is represented schematically in Fig. 1(c). The learner must simultaneously infer τ^2 and w. We shall see that this process yields a form of derived attention.

Because we must estimate posterior distributions for both the weights ($w_{j,i}$) and their prior variances (τ_i^2), we no longer have an exact update rule as in the ordinary Bayesian regression model described above. Instead, we must approximate the posterior distribution. For the present paper, we use a streaming version of mean field variational Bayes as our approximate inference algorithm (Broderick, Boyd, Wibisono, Wilson, & Jordan, 2013), because its calculations shed light on the generative model's psychological interpretation. We expect that other computational methods such as Markov Chain Monte Carlo or a particle filter would yield similar simulation results.

What follows is only a brief overview of the variational Bayes algorithm; see Appendix B for more details and Algorithm 5 in Appendix A for pseudocode. The algorithm maintains approximate posterior distributions for w and τ^2 (indicated below by the subscript q) and alternates between using expectations over w to approximate the posterior distribution of τ^2 and using expectations over τ^2 to approximate the posterior distribution of w. The rest of this section explains these approximations, how they are updated to reflect learning after each trial, and how they can be interpreted as derived attention.

The variational Bayes algorithm computes its approximate posterior distribution of w in much the same way as the simple Bayesian regression model described above. Instead of a single known value of τ^2, it substitutes the approximate posterior mean of τ_i^2 for each cue, which we denote $E_q[\tau_i^2]$. Thus in effect the vector of weights pertaining to a particular outcome (j) has the following prior distribution:

$$w_j \sim \mathcal{N}(0, \text{diag}(E_q[\tau_1^2], E_q[\tau_2^2], E_q[\tau_3^2], \ldots)) \text{ for each outcome } (j) \tag{16}$$

Learning, i.e. updating the distribution of w_j following new observations, works the same as in the standard Bayesian regression

[4] As noted above, most psychological models using Bayesian regression have in fact used a Kalman filter, which differs from the plain regression model in assuming that weights randomly drift from trial to trial (Dayan & Kakade, 2001; Gershman, 2015). This behaves similarly to the plain regression model described here, except that posterior weight variance (and hence learning rates) does not converge to zero as it does in the present model.

[5] Technically this makes it incorrect to call τ_i^2 "prior variance" in this model. However, we retain the term because it helps to clarify the Bayesian derived attention model's relationship to simple Bayesian regression, which in turn helps to explain the role that τ_i^2 plays in inference.

model described above, except for the critical difference that $E_q[\tau_i^2]$ varies across cues. Due to the differential effects of shrinkage, this will bias the model toward inferring larger or smaller values (regardless of sign) for each weight ($w_{j,i}$) depending on whether its estimated prior variance ($E_q[\tau_i^2]$) is large or small. We can interpret this as selective attention. Associative learning models (those based on a regression framework, such as those we have considered so far) represent selective attention in various ways. Some use cue-specific learning rates (e.g. Le Pelley et al., 2016), while others also scale the influence of cues on prediction (e.g. Kruschke, 2001). However, all such attentional mechanisms have the basic effect of modulating the influence of each cue based on the attention paid to it.

To estimate τ^2, the variational Bayes algorithm uses the approximate posterior means of w^2 (denoted $E_q[w^2]$). Estimating τ_i^2 is a simple matter of using $E_q[w_{1,i}^2]$, $E_q[w_{2,i}^2]$, . . . to update the conjugate inverse gamma prior (or equivalently gamma prior on the precision τ_i^{-2}):

$$E_q[\tau_i^2] = \frac{\tilde{\beta} + \frac{1}{2}\sum_j E_q[w_{j,i}^2]}{\tilde{\alpha} + \frac{k}{2} - 1} \tag{17}$$

where $\tilde{\beta} + \frac{1}{2}\sum_j E_q[w_{j,i}^2]$ and $\tilde{\alpha} + \frac{k}{2} - 1$ are posterior hyperparameters and k is the number of unique outcomes observed so far. This is exactly the same type of calculation as if $w_{1,i}, w_{2,i}, \ldots$ were directly observed. Recalling that τ_i^2 corresponds to attention to cue i, we see that this expresses the same principle as the derived attention formula used by Le Pelley et al. (2016): pay attention to cues with big weights (Le Pelley et al. use $|w_{j,i}|$ instead of $w_{j,i}^2$ and sum across outcomes rather than averaging; see Algorithm 2 in Appendix A). In the following part of the paper, we shall show that simulations of the Bayesian derived attention model confirm that it displays very similar behavior to Le Pelley et al.'s (2016) derived attention model (and hence to actual human participants) in key experiments.

The Bayesian model described here offers us a normative interpretation of derived attention: one can view it as a form of inductive reasoning about cue significance. This is a mathematical formalization of an insight by Le Pelley et al. (2016) quoted in the beginning of this paper. Observe that because the prior mean on weights is zero, τ_i^2 represents the average size (squared magnitude) of cue i's weights (this is a cue-specific version of Eq. (11)):

$$\begin{aligned}\tau_i^2 &= V[w_{j,i}] = E[(w_{j,i} - E[w_{j,i}])^2] \\ &= E[w_{j,i}^2] = \text{average size of cue } i\text{'s associations}\end{aligned} \tag{18}$$

Weight size corresponds to cue significance: a cue is important for making predictions only to the extent that it has large weights. Eq. (14) thus embodies the assumption that a cue that is significant for predicting some outcomes (i.e. has large weights) will tend to be significant for predicting other outcomes. This assumption supports inductive reasoning about cue significance and hence allows us to interpret posterior belief in τ_i^2 as the determiner of selective attention. For example, the Bayesian derived attention model explains learned predictiveness effects (Le Pelley & McLaren, 2003; Lochmann & Wills, 2003, see Fig. 2(a)) as due to participants inferring that cues will be more or less relevant in the transfer stage based on their relevance in stage 1: relevance in stage 1 leads to large estimated weights (w), which leads to large estimated τ_i^2, which leads to larger estimated weights (w) in stage 2. The following simulations illustrate how this happens in practice. The assumption that each cue's value of τ_i^2 (i.e. cue importance or average weight size) is constant across stages is not in fact always valid in the case of experimental tasks, but it may be valid in many natural environments, and if so this could produce an evolutionary drive for organisms to develop derived attention.

3. Simulations

The following simulations demonstrate the explanatory capabilities of the new Bayesian derived attention model. In the first three simulations, we show that it explains the same data as Le Pelley et al.'s (2016) version of derived attention, viz. learned predictiveness (Lochmann & Wills, 2003), inattention after blocking (Beesley & Le Pelley, 2011; Kruschke & Blair, 2000), and attentional capture by high-value cues (Anderson et al., 2011; Le Pelley et al., 2013). In each of these cases, the Bayesian derived attention model infers a higher prior weight variance (τ_i^2) for cues that have large association weights, which leads to greater attention. Simulation 4 shows that the Bayesian derived attention model can also produce retrospective revaluation, specifically backwards blocking (a form of learning about absent cues; Shanks, 1985) in the same manner as Dayan and Kakade's (2001) Bayesian regression model (Kalman filter). Thus the Bayesian derived attention model combines the explanatory capabilities of Le Pelley et al.'s (2016) derived attention model and Dayan and Kakade's (2001) Bayesian regression model under a single mechanism of joint Bayesian inference of weights across different cues and outcomes.

Moreover, casting derived attention into a Bayesian framework allows the new model to make a novel prediction: inattention after backward blocking (Simulation 5). Neither the Le Pelley et al. (2016) or Dayan and Kakade (2001) models make this prediction. Experimental confirmation of inattention after backward blocking would support the Bayesian derived attention model's core assumption, that inference of weights across different cues (as in retrospective revaluation effects) and inference of weights across different outcomes (as in derived attention) are facets of the same mechanism.

3.1. Methods

All simulations were performed using *statsrat*, a Python package written by one of the authors (SP) for simulating psychological learning models. The source code for *statsrat* is available at https://github.com/SamPaskewitz/statsrat, while the code for these particular simulations is available at https://github.com/SamPaskewitz/Bayesian-derived-attention; it allows the reader to replicate all of our results. For ease of presentation, we simulated simplified experimental designs that retain all the essential elements of previously published studies with fewer stimuli. Simulations of the actual experimental designs produced similar results.

The focus of these simulations was to reproduce ordinal response patterns, i.e. a greater probability of choosing one response over another in critical trials (during the test stage). This is a common method for analyzing category learning data, and a similar approach called the ordinal adequacy test (OAT) is often used to evaluate mathematical models (Wills & Pothos, 2012). We used the *statsrat* package's *perform_oat* function for this purpose. Performing an OAT consists of the following steps. First, a behavioral score is defined that captures the empirical pattern of results. For each test trial type, one response is empirically more common than the other. The behavioral score for a simulation consists of the sum across test trials of the simulated probability of the empirically more common response minus that of the less common response. Positive values of this score represent the same ordinal pattern as that observed empirically, while negative values represent the opposite pattern. Next, a function is defined that gives the behavioral score averaged across several random trial sequences (we used 10) as a function of the model's free parameters. Finally, a non-linear optimization function is used to search the model's parameter space to find the maximum and minimum behavioral score values (averaged across trial sequences) that the model can produce.

An OAT has three possible results:

1. The maximum behavioral score produced by the model is positive, and the minimum is negative: this means that the model can reproduce the empirical pattern, but can also produce the opposite.
2. The minimum score is positive: the model strictly predicts the observed pattern (it cannot produce the opposite).
3. The maximum score is negative: the model cannot reproduce the observed pattern.

In all the experimental designs tested with the Bayesian derived attention model we obtained result 2, indicating that the model strictly predicts the observed phenomena. In other words, this aspect of the model's behavior is not parameter dependent. Therefore, we used the same set of parameter values ($\tilde{\phi}_0 = -0.4, \tilde{\phi}_1 = 2.0, \sigma^2 = 1.0, \kappa = 5.0$) for producing simulation graphs.

In all of these simulations we use a softmax function to transform category label predictions ($\hat{y}$) into response probabilities:

$$P(\text{answer category } h) = \frac{\text{psb}_h \exp(\kappa \hat{y}_h)}{\sum_j \text{psb}_j \exp(\kappa \hat{y}_j)} \tag{19}$$

or equivalently in vector form:

$$\text{response probabilities} = \frac{\text{psb} \circ \exp(\kappa \hat{y})}{\sum_j \ \text{psb}_j \exp(\kappa \hat{y}_j)} \tag{20}$$

Here κ is a positive response scaling parameter (recall that $\text{psb}_j = 1$ if category j is a possible response option and 0 otherwise). This softmax response function is commonly used in learning and decision making research. The current simulations aim only to reproduce ordinal patterns in the data, i.e. a greater response probability for one category than another for defined test stimuli. Therefore any other monotonic transformation of $\hat{y}$ into response probabilities would produce exactly the same results as the softmax function we use.

3.2. Simulation 1: Learned predictiveness

Learned predictiveness consists of the fact that organisms pay attention (as assessed by choice data and eye gaze) to cues in accordance with their task relevance and that this attentional bias transfers to later learning (Le Pelley & McLaren, 2003; Lochmann & Wills, 2003). Fig. 2(a) shows a simplified experimental design. In the first stage (Relevance), participants learn to classify stimuli into two categories. Each stimulus consists of two cues, A or B and X or Y. Cues A and B are predictive of category membership: A consistently indicates category I, while every stimulus with B is in category II. In contrast, cues X and Y are irrelevant to the task. In the following Transfer stage, the same cues are associated with two entirely new category labels; all of the cues are now equally relevant. In the final Test stage, ambiguous cue pairs test the relative strength of the cues' associations acquired in the Transfer stage. It has been consistently found that the formerly predictive cues (A and B) form stronger associations in the Transfer stage than the formerly irrelevant cues (X and Y), as indicated by subjects' test responses (Le Pelley & McLaren, 2003; Le Pelley et al., 2009; Lochmann & Wills, 2003; Mitchell, Griffiths, Seetoo, & Lovibond, 2012). Because the two stages use different category labels, this finding suggests that the Relevance stage trained participants to pay more attention to A and B than to X and Y. Eye-tracking results support this attentional interpretation (Mitchell et al., 2012).

The Bayesian derived attention model explains learned predictiveness in the following manner. In the first (Relevance) stage, the predictive cues (A and B) form large associations ($w_{j,i}$) with the categories, while the irrelevant cues (X and Y) have associations that are closer to zero (Fig. 2(c)). By the end of this stage, the model thus infers that τ_i^2 is larger for the predictive cues (A and B) than for the irrelevant ones (X and Y) (Fig. 2(b); see also Eq. (17)). This difference in estimated τ_i^2 values – which we interpret as greater attention to A and B than to X and Y – leads to larger weight estimates for the former cues in the Transfer and Test stages (Fig. 2(d)). This explanation is very similar to that provided by the Le Pelley et al. (2016) derived attention model (compare Fig. 2 with Le Pelley et al.'s Figure 3).

3.3. Simulation 2: Reduced attention to blocked cues

Derived attention also explains inattention to blocked cues. When an outcome is already well predicted, new, redundant cues do not form strong associations. In other words, existing associations *block* new learning. Blocking is widespread in both animal and human learning (Kamin, 1968; Shanks, 1985). Blocking itself is not the primary phenomenon of interest to us here: it can be explained by simple error-correction principles common to every model of the Rescorla–Wagner/linear regression family (including both the Le Pelley et al. and Bayesian versions of derived attention). Rather, we are interested in the fact that cues suffer a loss of attention after being blocked (Beesley & Le Pelley, 2011; Kruschke & Blair, 2000; Le Pelley, Beesley, & Suret, 2007).

We illustrate inattention after blocking using a simplified design (see Fig. 3(a)). In the Single Cue stage, participants learn A → I and B → II associations, followed by the addition of the redundant cues X and Y in the Double Cue stage. Thus, X and Y are blocked by A and B. (The design could easily include a test for blocking but we omit this for simplicity.) The Transfer stage sees the blocked cues (X and Y) paired with control cues (E and G) and uses new category labels. The test stage has pairs of blocked and control cues with opposite associations, allowing us to indirectly compare attention during the Transfer stage. The basic logic of the Transfer and Test stages is the same as in the learned predictiveness experiments described above.

Inattention after blocking is easy for the derived attention theory to explain: blocked cues have small association weights and are thus ignored (Le Pelley et al., 2016). Fig. 3 shows simulation results from the new Bayesian derived attention model. As in Simulation 1 (learned predictiveness), cues X and Y have small association weights (w) prior to the Transfer stage and thus low attention (τ_i^2). This reduces the size of the weights they develop in the Transfer stage, leading to the observed result.

3.4. Simulation 3: Attention toward high value cues (value effect)

Cues associated with large rewards attract more attention than those associated with small rewards. This effect has been found in both visual search (Anderson et al., 2011; Le Pelley, Pearson, Porter, Yee, & Luque, 2018) and category learning experiments (Le Pelley et al., 2013). This value effect on attention is a key support of Le Pelley et al.'s (2016) derived attention theory: high-value cues have larger association weights, and thus receive more attention. We shall show that our new Bayesian derived attention model also captures the value effect.

To simulate the value effect, we use a category learning task based loosely on Le Pelley et al. (2013), but simplified for ease of presentation (see Fig. 4(a)). In the Value stage, correct categorization answers are worth different amounts of reward depending on the cue present: cues A and B are high-value cues (reward of 100) while C and D are low-value cues (reward of 10; both rewards were divided by a factor of 100 in the simulation). This is followed by a Transfer stage similar to those described above, in which the high- and low-value cues from the Value stage are paired together and predict new category labels. The Test stage shows that Transfer stage learning was dominated by A and B (the high-value cues), suggesting that these cues received more attention than did C and D (this is further supported by eye-tracking; Le Pelley et al., 2009). Simulation shows that the Bayesian derived attention model reproduces the empirical result (Fig. 4). In the Value stage, the high-value cues (A and B) develop larger weights ($w_{j,i}$) than do the low-value cues (C and D), because the former are associated with larger outcomes (see Fig. 4(c)). The model thus infers that the high-value cues are more significant (i.e. have greater τ_i^2; Fig. 4(b)), and this affects weight estimates in the Transfer stage similarly to the experiments described above (Fig. 4(d)).

3.5. Simulation 4: Backward blocking (retrospective revaluation)

A cue's ability to control behavior can change even when that cue is not present; this is called *retrospective revaluation*. Backward blocking (Shanks, 1985) is one example. As in the other simulations, we illustrate backward blocking using a simple category learning task (Fig. 5(a)). In the first (Double Cue) stage, various pairs of cues are associated with different outcomes (A.X → I, B.Y → II, E.F → I and G.H → II). In the following Single Cue stage, cues A and B are presented alone and followed by the same outcomes as before. This weakens the control of the cues previously paired with A and B (respectively X and Y) over behavior, as assessed in the Test stage. The design is called *backward* blocking because – compared to ordinary blocking – the order of the Single and Double Cue stages is reversed. Ordinary blocking is easy for all Rescorla–Wagner family models (including Le Pelley et al.'s derived attention model) to explain via prediction error (Rescorla & Wagner, 1972). However, these models cannot explain retrospective revaluation effects such as backward blocking because they do not allow for learning about absent cues. For the sake of thoroughness, we confirmed via simulation (using the *perform_oat* function in *statsrat*) that Le Pelley et al.'s (2016) derived attention model could not produce a backward blocking effect with any combination of free parameter values (maximum behavioral score of 0 across all parameter combinations).

The Bayesian derived attention model easily produces backward blocking, as shown in Fig. 5. This is a consequence of its Bayesian regression machinery, as originally explained by Dayan and Kakade (2001). When cues are presented together, their association weights become negatively correlated in the learner's posterior distribution. Therefore, when the estimate of one of these weights increases during subsequent learning, the other decreases. For example, the weights of cues A and X become negatively correlated in the Double Cue stage because the two cues are presented together; this can be seen as a sort of Hebbian learning about cue relationships. During the following Single Cue stage, further training increases the estimate of cue A's weights, thus decreasing the estimate of cue X's weights and leading to a backward blocking effect (see Fig. 5(b)). Thus retrospective revaluation in the Bayesian derived attention model can be seen as complementary to its selective attention mechanism: the latter produces positive generalization between weights from the same cue to different outcomes due to their prior positive correlation (from shared τ_i^2), whereas the former produces negative generalization between weights from different cues to the same outcome due to their prior negative correlation (from previous joint presentation). The final simulation reported here tests a key prediction of the model that emerges from the combination of these two mechanisms.

3.6. Simulation 5: Reduced attention after backward blocking

By recasting derived attention (Le Pelley et al., 2016) within the context of Bayesian regression (Dayan & Kakade, 2001), the Bayesian derived attention model makes a novel prediction: reduced attention to cues that have undergone *backward* blocking. So far as we are aware, it is the first computational model that has been demonstrated to make this prediction. Fig. 6(a) illustrates a simplified experimental design, which is the same as inattention after (forward) blocking (Simulation 2, Fig. 3(a)) except that the order of the first two stages is reversed. So far as we can tell, this type of experiment has not yet been performed. Kruschke and Blair (2000) did perform a similar experiment. However, their design paired the backwardly blocked cues (analogous to X and Y in our design) with the cues that backwardly blocked them (analogous to A and B) during transfer training trials. Thus, one cannot tell whether the former have weak associations because they lost attention, or because attention was focused on the latter. The design in Fig. 6(a) eliminates this confound by pairing X and Y with cues (E and G) that have not been involved with the backward blocking process.

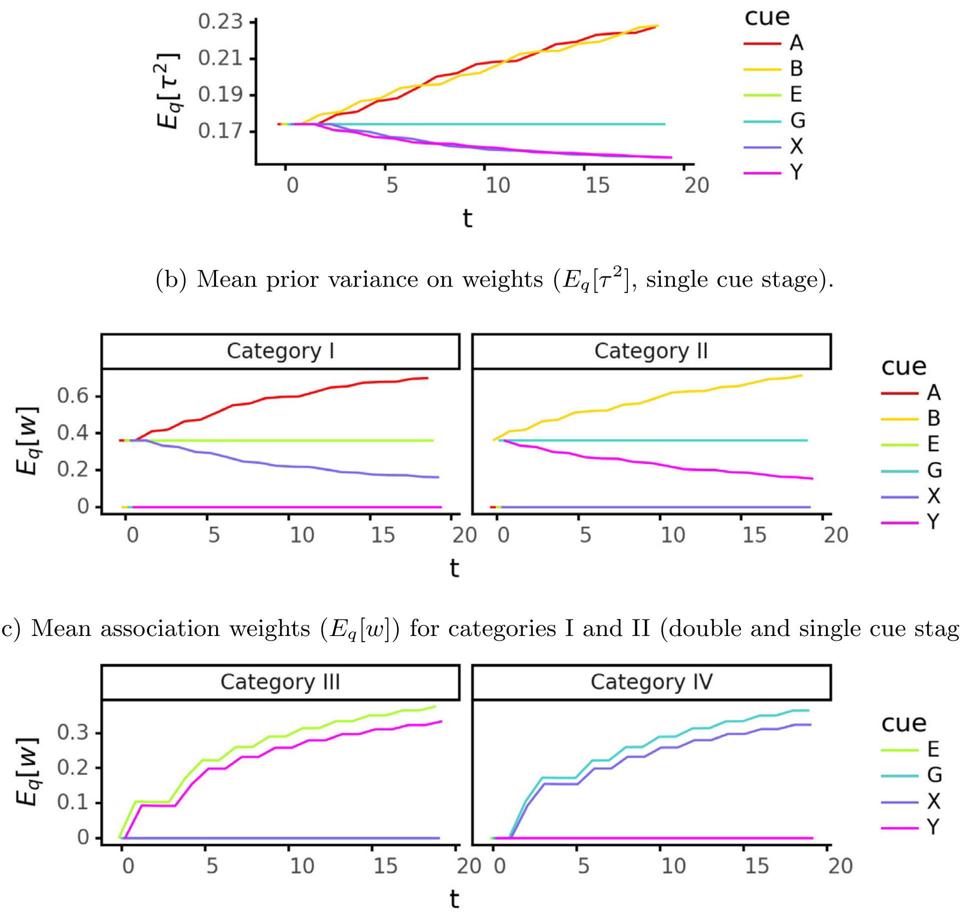
The Bayesian derived attention model predicts inattention after backward blocking. Its Bayesian regression mechanism produces a decrease in the estimated weights of cues X and Y as a result of backward blocking (Fig. 6(c), just as in Simulation 4). This leads to less attention to these cues, i.e. smaller estimated τ_{X}^2 and τ_{Y}^2 (Fig. 6(b), compare to Simulation 2) and hence smaller estimated weights during the Transfer stage (Fig. 6(d)). Thus learning about $w_{\mathrm{I,A}}$ during the Single Cue stage leads to a decreased expectation for the magnitude of $w_{\mathrm{IV,X}}$ prior the Transfer stage, which is possible only because the model performs inference over weights for all cues and all outcomes jointly. Ordinal adequacy tests confirm that neither the Le Pelley et al. (2016) derived attention model nor Dayan and Kakade (2001) Bayesian regression model can predict this result (maximum behavioral score of 0 for both models across all parameter combinations). The former model does not produce any form of backward blocking and hence of changes to the blocked cues' weights, while the latter model has no mechanism for learning about the first set of outcomes (I and II) to affect learning about the second set of outcomes (III and IV).

4. Discussion

We interpret derived attention as Bayesian inference over a generative model that assumes each cue's association weights to various outcomes have a common prior variance (τ_i^2). The estimated size of τ_i^2 influences the estimated size of cue *i*'s weights through the mechanism of shrinkage; τ_i^2 thus can be interpreted as selective attention. Attention learning is thus inductive inference about cue significance, i.e. average weight size. The Bayesian derived attention model explains the same set of attentional phenomena as the Le Pelley et al. (2016) version of derived attention, viz. learned predictiveness (Le Pelley et al., 2009; Lochmann & Wills, 2003), inattention after blocking (Beesley & Le Pelley, 2011; Kruschke & Blair, 2000; Le Pelley et al., 2007), and the value effect (Anderson et al., 2011; Le Pelley et al., 2013). This represents a theoretical advance over previous derived attention

Double Cue		Single Cue		Transfer		Simulated Test Responses		
Cues	Category	Cues	Category	Cues	Category	Cues	p(III)	p(IV)
A.X	I	A	I	E.Y	III	E.X	0.55	0.45
B.Y	II	B	II	G.X	IV	G.Y	0.45	0.55
E.F	I							
G.H	II							

(a) A simplified design for testing inattention after backward blocking (novel design).



(b) Mean prior variance on weights ($E_q[\tau^2]$, single cue stage).

(c) Mean association weights ($E_q[w]$) for categories I and II (double and single cue stages).

(d) Mean association weights ($E_q[w]$) for categories III and IV (transfer stage).

Fig. 6. Simulation 5 (inattention after backward blocking).

models (Esber & Haselgrove, 2011; Frey & Sears, 1978; Le Pelley et al., 2016): we have explained *why* one might expect animals to use derived attention. If inductive inference about cue significance is valid in organisms' environments, then natural selection would tend to produce derived attention. The Bayesian derived attention model also leverages Bayesian regression (Dayan & Kakade, 2001) to explain phenomena that the Le Pelley et al. (2016) model cannot, viz. retrospective revaluation effects such as backward blocking (Shanks, 1985). Finally, the Bayesian derived attention model predicts a result not produced by either derived attention (Le Pelley et al., 2016) or Bayesian regression (Dayan & Kakade, 2001) alone: reduced attention after backward blocking.

The Bayesian derived attention model ultimately explains all of these phenomena in terms of the same process: inference about association weights (w_{ij}). Derived attention involves inference across different weights for the same predictor (large w_{ji} indicates large $w_{j'i}$) by the intermediate step of estimating τ_i^2. Retrospective revaluation, on the other hand, involves inference across different weights for the same outcome (large w_{ji} indicates small $w_{ji'}$). The model's prediction of inattention after backward blocking comes from the combination of these two mechanisms.

We see several directions for future work with the Bayesian derived attention model. The first is extending the model to explain the phenomenon of learned helplessness. Whereas the Bayesian derived attention model deals with inductive reasoning about cue importance, the same approach could be extended to modeling a learner's beliefs about the efficacy of its own actions. In learned helplessness experiments, participants who first are

Table 1
The inverse base rate effect (Kruschke, 1996, Experiment 1). The Category column indicates correct responses, while the Prop. column shows the proportion of each trial type. The rightmost two columns show some of the empirical response proportions in the test stage. Due to the symmetry of the design, the test results for I_1 are averaged with I_2, PC_1 with PC_2, etc. The critical finding is that participants tend to choose the rare outcome (R_1 or R_2) on $PC + PR$ test trials: this is the inverse base rate effect.

Training			Test		
Cues	Category	Prop.	Cues	C Responses	R Responses
$PC_1.I_1$	C_1	1/8	I	.746	.174
$PC_2.I_2$	C_2	1/8	PC	.933	.031
$PR_1.I_1$	R_1	3/8	PR	.040	.911
$PR_2.I_2$	R_2	3/8	$PC.PR$	.353	.612

given an impossible task to perform are less likely to solve a possible task (Hiroto & Seligman, 1975). This can be interpreted as inductive reasoning about the potency of one's own actions: if one's actions were ineffective previously, then one infers that they will be ineffective in the future. This could probably be modeled using a variation of the Bayesian derived attention model featuring both association weights that reflect the learner's actions and association weights that do not reflect the learner's actions. If it is assumed that all of the action weights have a shared variance, we would expect to see learned helplessness effects in a manner analogous to the learned inattention simulations presented above (low τ_i^2 for action weights). Similarly, high confidence in one's ability to affect outcomes could be modeled as a high τ_i^2 for action weights. This approach might be useful for understanding individual differences in perceived self-efficacy, e.g. illusory sense of control over gambling outcomes (Joukhador, Maccallum, & Blaszczynski, 2003).

The Bayesian derived attention model might also be fruitfully applied to Pavlovian conditioning. Formally, there is not much difference between category learning experiments such as those described here and a Pavlovian conditioning experiment. Instead of category labels, the outcomes (y) correspond to unconditioned stimuli such as shocks or food (with $\text{lrn} = 1$ and $\text{psb} = 1$ on each time step) while the cues (x) correspond to unconditioned stimuli. The conditioned response (e.g. freezing, approach) is a monotonically increasing function of the unconditioned stimulus prediction ($\hat{y}$).

In particular, the Bayesian derived attention model might yield insights about the persistence of conditioned responses following extinction. Conditioned responses are not forgotten following extinction, but persist and can be revealed by manipulations such as changing background stimuli (renewal, Bouton & Bolles, 1979) or simply waiting before test (spontaneous recovery, Estes & Skinner, 1941). This closely resembles the fact that in clinical psychological practice, while exposure to fear-provoking stimuli (e.g. spiders, trauma reminders) tends to reduce fear, that fear often returns over time. Modified versions of the Rescorla–Wagner model can explain some of these results when simulated correctly (Delamater & Westbrook, 2014; Paskewitz, Stoddard, & Jones, 2022). During extinction or exposure therapy, background stimuli (i.e. the context) develop conditioned inhibition, becoming safety signals that prevent the total erasure of the fear association. We have recently shown that certain attentional mechanisms allow Rescorla–Wagner based models to explain a wider range of such phenomena than previously thought (Paskewitz et al., 2022). It might well be the case that a modified version of the Bayesian derived attention model, which is similar to the Rescorla–Wagner model, might be able to explain a still broader array of results relating to the return of fear.

The Bayesian derived attention model – like Le Pelley et al.'s (2016) version – cannot explain an important category learning result known as the inverse base rate effect. This is a phenomenon in which people under certain circumstances predict a rare outcome as opposed to a common outcome when presented with conflicting cues equally associated with both (Kruschke, 1996; Medin & Edelson, 1988, see Table 1). The result is counterintuitive because, if there is equal evidence for the rare and common outcomes, one ought to choose the common one. The inverse base rate effect and related phenomena are explained by learning models in which cues compete for attention on the basis of their predictiveness (Kruschke, 2001; Paskewitz & Jones, 2020). However, derived attention cannot produce the same effect: the common cues should always receive more attention and hence have more control over behavior (Le Pelley et al., 2016). Ordinal adequacy tests confirm that the Bayesian version of derived attention, like the (Le Pelley et al., 2016) model, fails to produce the inverse base rate effect.

One way to obtain the inverse base rate effect might be to limit attentional capacity in the Bayesian derived attention model. People can only pay attention to a limited number of stimuli at once (Landauer, 1986). Limiting attentional capacity allows other regression-based category learning models to explain the inverse base rate effect in terms of cue competition (Kruschke, 2001; Paskewitz & Jones, 2020). Future iterations of Bayesian derived attention theory might be able to model limited capacity attention by changing the form of prior distribution on association weights (w). We can interpret limited attention as sparsity in the matrix of association/regression weights: most cues are ignored and have a weight of zero, while a few cues are attended and have non-zero weights. Other forms of prior distribution on w used in Bayesian regression enforce sparsity more stringently than the normal priors used here (Kuo & Mallick, 1998; Mitchell & Beauchamp, 1988), and might thus be useful for constructing models of limited capacity attention.

Another idea that might allow the Bayesian derived attention model to produce the inverse base rate effect would be to incorporate change points into the generative model (Wilson, Nassar, & Gold, 2013). Instead of assuming that the true association weights remain constant across time, the model would assume that these weights are occasionally re-drawn from a generative prior ($w_{j,i} \sim \mathcal{N}(0, \tau_i^2)$) at unsignaled change points. This would tend to increase the learning rate (relative importance of new vs. old observations) whenever the inference algorithm believes a change point has occurred. This might increase the learning rate in particular during rare cue/outcome trials due to the high prediction error, somewhat like how the high prediction error draws attention toward the rare cues in other models (Kruschke, 1996; Paskewitz & Jones, 2020, see Table 1). If changepoint detection really does this, a suitably revised version of Bayesian derived attention might produce an inverse base rate effect. We plan to investigate this possibility in future work.

Incorporating change points into the Bayesian derived attention model is attractive for another reason. Historically, models of selective attention in learning have tended to focus on two competing principles: predictiveness (Mackintosh, 1975) and uncertainty (Pearce & Hall, 1980). The first principle says that organisms should attend to cues that are known to be good predictors, while the second says that organisms should attend to cues whose meaning is uncertain. Derived attention is an expression of the predictiveness principle. However, the Bayesian framework naturally represents the principle of uncertainty in the form of posterior weight variance. In the current Bayesian derived attention model, this causes the effective learning rate for a cue to decrease the more times it is observed, which is a very limited expression of the uncertainty principle. Adding

Table 2
Key to mathematical symbols.

Symbol	Meaning
x_i	indicates if cues i was present
y_j	indicates if outcome j occurred (e.g. category label)
psb_j	indicates if outcome j is possible during the current stage of the task, i.e. if category j is a response option
lrn_j	indicates whether feedback was given about outcome j and hence whether the organism will learn about it
fb_j	indicates if feedback was given about outcome j
$\hat{y}_j$	prediction of outcome j
δ_j	prediction error for outcome j
$\mathrm{softmax}(\hat{y}; \kappa)$	response probabilities, equal to $\frac{\mathrm{psb} \circ \exp(\kappa \hat{y})}{\sum_j \ \mathrm{psb}_j \exp(\kappa \hat{y}_j)}$
$u \circ v$	component-wise multiplication of vectors or matrices u and v (Hadamard product)
κ	softmax response scaling parameter
Basic Rescorla–Wagner model	
λ_{par}	fixed learning rate parameter
$\hat{w}_{ji}$	point estimate of association between cue i and outcome j
Derived attention model of Le Pelley et al.	
λ_i	learning rate for cue i
$\hat{w}_{ji}$	point estimate of association between cue i and outcome j
λ_{min}	minimum learning rate
Bayesian Regression	
w_{ji}	association between cue i and outcome j (random variable)
τ^{-2}	unknown prior precision of $w_{1,1}, w_{2,2}, \dots, w_{1,2}, \dots$ (fixed and known, equal to $\frac{1}{\tau^2}$)
σ^2	variance of y (fixed and known)
μ_j	conventional hyperparameter for w_j (mean vector of w_j)
Σ_j^{-1}	conventional hyperparameter for w_j (precision matrix, i.e. inverse covariance matrix of w_j)
$\psi_{0,j}$	natural hyperparameter for w_j $(= \Sigma_j^{-1} \mu_j)$
$\psi_{1,j}$	natural hyperparameter for w_j $(= \Sigma_j^{-1})$
$T_{0,j}, T_{1,j}$	sufficient statistics for w_j
Bayesian derived attention model	
w_{ji}	association between cue i and outcome j (random variable)
τ_i^{-2}	unknown prior precision of $w_{1,i}, w_{2,i}, \dots$ (random variable, equal to $\frac{1}{\tau_i^2}$)
σ^2	variance of y (fixed and known)
q	variational distribution, $\approx$ posterior distribution of w and τ^2
$E_q[w_{ji}]$	variational expectation (mean) of w_{ji}
$V_q[w_{ji}]$	variational variance of w_{ji}
$E_q[w_{ji}^2]$	variational expectation of w_{ji}^2
$E_q[\tau_i^{-2}]$	variational mean of τ_i^{-2}
$\psi_{0,j}, \psi_{1,j}$	variational hyperparameters for w_j (respectively equal to $\Sigma_j^{-1} \mu_j$ and Σ_j^{-1})
$\phi_{0,i}, \phi_{1,i}$	variational hyperparameters for τ_i^{-2} (respectively equal to $-\beta_i$ and $\alpha_i - 1$)
$\tilde{\phi}_0, \tilde{\phi}_1$	prior variational hyperparameters for τ^{-2}
$T_{0,j}, T_{1,j}$	sufficient statistics for w_j

change point detection would allow that posterior weight variance to increase again when expectations are violated, providing a fuller expression of the uncertainty principle. Thus, a version of Bayesian derived attention with change point detection might be able to integrate the two attentional principles in a principled manner through its generative model and thereby explain a wide range of learning data.

Appendix A. Pseudocode

Note: in order to keep the notation uncluttered, we suppress the t subscript for time steps in the algorithm pseudocode. Thus y, x etc. denote y_t, x_t etc. Table 2 explains the notation used.

Appendix B. Mathematical details

B.1. Exponential distribution forms

It will helpful to first re-write the multivariate and univariate normal distributions – along with the gamma distribution – in exponential family form, i.e.:

$$p(u) = \exp\Big(\sum_k \langle \theta_k, T_k(u) \rangle - f(u) - g(\theta)\Big) \tag{21}$$

where $\theta_0, \theta_1, \dots$ are the *natural parameters* and $T_0(u), T_1(u), \dots$ are the accompanying *sufficient statistics* of u. Each natural parameter (v_k) can be a scalar, vector or square matrix; the corresponding sufficient statistic ($T_k(u)$) has the same dimensions (e.g. an $m \times 1$ vector if v_k is an $m \times 1$ vector). The inner product operation $\langle v_k, T_k(u) \rangle$ is defined as follows for the various types of natural parameter and sufficient statistic:

$$\langle v_k, T_k(u) \rangle$$
$$= \begin{cases} v_k T_k(u) & \text{scalars} \\ \sum_i \theta_{k,i} T_{k,i}(u) = v_k^T T_k(u) & \text{vectors} \\ \sum_i \sum_j \theta_{k,i,j} T_{k,i,j}(u) & \text{matrices (Frobenius inner product)} \end{cases} \tag{22}$$

Writing distributions in exponential family forms greatly simplifies Bayesian learning (conjugate prior updating).

B.1.1. Univariate normal distribution

Suppose $U \sim \mathcal{N}(\mu, \sigma^2)$ with support $u \in \mathbb{R}$. Then

$$p(u) = (2\pi\sigma^2)^{-1/2} \exp\big(-\tfrac{1}{2} \tfrac{(u-\mu)^2}{\sigma^2}\big) \tag{23}$$

Algorithm 1: The Rescorla–Wagner model (Rescorla & Wagner, 1972) applied to category learning. See Table 2 for an explanation of symbols.

```
input: simulation parameters ($\lambda_{\text{par}}, \kappa$), sequences of variables
  defining the experiment ($x, y$, psb, lrn)
output: sequence of simulated response probabilities
begin
    initialize weight estimates;
    for $j \in 1 : n_y$ do
        $\hat{w}_j \leftarrow 0$ ($n_x \times 1$ column vector);
    while task continues do
        prediction and response;
        for $j \in 1 : n_y$ do
            $\hat{y}_j \leftarrow \text{psb}_j x^T \hat{w}_j$;
        response probabilities $\leftarrow$ softmax($\hat{y}; \kappa$);
        association learning;
        $\lambda \leftarrow \lambda_{\text{par}} x$;
        for $j | lrn_j = 1$ do
            $\hat{w}_j \leftarrow \hat{w}_j + \lambda(y_j - \hat{y}_j)$;
```

Algorithm 2: The derived attention model of Le Pelley et al.. See Table 2 for an explanation of symbols.

```
input: simulation parameters ($\lambda_{\text{min}}, \kappa$), sequences of variables
  defining the experiment ($x, y$, psb, lrn)
output: sequence of simulated response probabilities
begin
    initialize weight estimates;
    for $j \in 1 : n_y$ do
        $\hat{w}_j \leftarrow 0$ ($n_x \times 1$ column vector);
    while task continues do
        prediction and response;
        for $j \in 1 : n_y$ do
            $\hat{y}_j \leftarrow \text{psb}_j x^T \hat{w}_j$;
        response probabilities $\leftarrow$ softmax($\hat{y}; \kappa$);
        cue-specific learning rates (attention);
        for $i \in 1 : n_x$ do
            $\lambda_i \leftarrow x_i \min\left(\max(\sum_j |\hat{w}_{j,i}|, \lambda_{\text{min}}), 1\right)$;
        association learning;
        for $j | lrn_j = 1$ do
            $\hat{w}_j \leftarrow \hat{w}_j + \lambda(y_j - \hat{y}_j)$;
```

$$= \exp\left(-\tfrac{1}{2}\tfrac{(u-\mu)^2}{\sigma^2} - \tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(\sigma^2)\right) \tag{24}$$

$$= \exp\left(\tfrac{\mu}{\sigma^2}u + \tfrac{1}{\sigma^2}\tfrac{-u^2}{2} - \tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\left(\tfrac{\mu^2}{\sigma^2} + \log(\sigma^2)\right)\right) \tag{25}$$

We can express this in terms of the natural parameters $\theta_0 = \frac{\mu}{\sigma^2}$ and $\theta_1 = \frac{1}{\sigma^2}$.[6] The corresponding sufficient statistics are $T_0(u) = u$ and $T_1(u) = \frac{-u^2}{2}$.

B.1.2. Multivariate normal distribution

Suppose $U \sim \mathcal{N}(\mu, \Sigma)$ with support $u \in \mathbb{R}^k$. Then

$$p(u) = (2\pi)^{-k/2}|\Sigma|^{-1/2}\exp\left(-\tfrac{1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu)\right) \tag{26}$$

$$= \exp\left(-\tfrac{1}{2}(u-\mu)^T\Sigma^{-1}(u-\mu) - \tfrac{k}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma|\right) \tag{27}$$

Algorithm 3: Bayesian regression model (conventional parameters version, equivalent to Algorithm 4). See Table 2 for an explanation of symbols.

```
input: simulation parameters ($\tau^{-2}, \sigma^2, \kappa$), sequences of variables
  defining the experiment ($x, y$, psb, lrn)
output: sequence of simulated response probabilities
begin
    initialize distribution for $w$;
    for $j \in 1 : n_y$ do
        $\mu_j \leftarrow 0$ ($n_x \times 1$ column vector);
        $\Sigma_j^{-1} \leftarrow \tau^{-2}\mathbb{I}$ ($n_x \times n_x$ matrix);
    while task continues do
        prediction and response;
        for $j \in 1 : n_y$ do
            $\hat{y}_j \leftarrow \text{psb}_j x^T \mu_j$;
        response probabilities $\leftarrow$ softmax($\hat{y}; \kappa$);
        learning (update $\mu$ and $\Sigma$);
        for $j | lrn_j = 1$ do
            $\lambda_j \leftarrow \frac{\Sigma_j x}{x^T \Sigma_j x + \sigma^2}$;
            $\mu_j \leftarrow \mu_j + \lambda_j(y_j - \hat{y}_j)$;
            $\Sigma_j \leftarrow (\Sigma_j^{-1} + \frac{xx^T}{\sigma^2})^{-1}$;
```

Algorithm 4: Bayesian regression model (exponential family version, equivalent to Algorithm 3). See Table 2 for an explanation of symbols.

```
input: simulation parameters ($\tau^{-2}, \sigma^2, \kappa$), sequences of variables
  defining the experiment ($x, y$, psb, lrn)
output: sequence of simulated response probabilities
begin
    initialize sufficient statistics;
    for $j \in 1 : n_y$ do
        $T_{0,j} \leftarrow 0$ ($n_x \times 1$ column vector);
        $T_{1,j} \leftarrow 0$ ($n_x \times n_x$ matrix);
    while task continues do
        distribution of $w$;
        for $j | psb_j = 1$ do
            $\psi_{0,j} \leftarrow T_{0,j}$;
            $\psi_{1,j} \leftarrow \tau^{-2}\mathbb{I} + T_{1,j}$;
            $\mu_j \leftarrow \psi_{1,j}^{-1}\psi_{0,j}$;
        prediction and response;
        for $j \in 1 : n_y$ do
            $\hat{y}_j \leftarrow \text{psb}_j x^T \mu_j$;
        response probabilities $\leftarrow$ softmax($\hat{y}; \kappa$);
        learning (update sufficient statistics);
        for $j | lrn_j = 1$ do
            $T_{0,j} \leftarrow T_{0,j} + \frac{xy_j}{\sigma^2}$;
            $T_{1,j} \leftarrow T_{1,j} + \frac{xx^T}{\sigma^2}$;
```

$$= \exp\left(u^T\Sigma^{-1}\mu - \tfrac{1}{2}u^T\Sigma^{-1}u - \tfrac{k}{2}\log(2\pi) - \tfrac{1}{2}(\mu^T\Sigma^{-1}\mu + \log|\Sigma|)\right) \tag{28}$$

$$= \exp\left(\langle\Sigma^{-1}\mu, u\rangle + \langle\Sigma^{-1}, -\tfrac{1}{2}uu^T\rangle - \tfrac{k}{2}\log(2\pi) - \tfrac{1}{2}(\mu^T\Sigma^{-1}\mu + \log|\Sigma|)\right) \tag{29}$$

[6] Some sources give $-\frac{1}{2}\frac{1}{\sigma^2}$ instead of $\frac{1}{\sigma^2}$ as the second natural parameter. We find it more convenient to assign the $-\frac{1}{2}$ factor to the sufficient statistic (T_1) instead, so that θ_1 is intrepretable as the precision (inverse variance).

Algorithm 5: Bayesian derived attention model. See Table 2 for an explanation of symbols.

```
input: simulation parameters $(\tilde{\phi}_0, \tilde{\phi}_1, \sigma^2, \kappa)$, sequences of variables
  defining the experiment $(x, y, \text{psb}, \text{lrn})$
output: sequence of simulated response probabilities
begin
    initialize sufficient statistics for $w$;
    for $j \in 1 : n_y$ do
        $T_{0,j} \leftarrow 0$ ($n_x \times 1$ column vector);
        $T_{1,j} \leftarrow 0$ ($n_x \times n_x$ matrix);
    while task continues do
        compute $E_q[\tau_i^{-2}]$;
        for $i \in 1 : n_x$ do
            $E_q[\tau_i^{-2}] \leftarrow \frac{\phi_{1,i}+1}{-\phi_{0,i}}$;
        update variational distribution of $w$;
        for $j | psb_j = 1$ do
            $\psi_{0,j} \leftarrow T_{0,j}$;
            $\psi_{1,j} \leftarrow \text{diag}(E_q[\tau^{-2}]) + T_{1,j}$;
            $E_q[w_j] \leftarrow \psi_{1,j}^{-1} \psi_{0,j}$;
            $V_q[w_j] \leftarrow$ the diagonal of $\psi_{1,j}^{-1}$;
            $E_q[w_j^2] \leftarrow V_q[w_j] + E_q[w_j]^2$ (the squares in the second
              term are component-wise);
        update variational distribution of $\tau^{-2}$;
        for $i \in 1 : n_x$ do
            $\phi_{0,i} \leftarrow \tilde{\phi}_0 - \frac{1}{2} \sum_j E_q[w_{ji}^2]$;
            $\phi_{1,i} \leftarrow \tilde{\phi}_1 + \frac{k}{2}$ ($k$ = number of outcomes so far);
        prediction and response;
        for $j \in 1 : n_y$ do
            $\hat{y}_j \leftarrow \text{psb}_j x^T E_q[w_j]$;
        response probabilities $\leftarrow \text{softmax}(\hat{y}; \kappa)$;
        learning (update sufficient statistics for $w$);
        for $j | lrn_j = 1$ do
            $T_{0,j} \leftarrow T_{0,j} + \frac{x y_j}{\sigma^2}$;
            $T_{1,j} \leftarrow T_{1,j} + \frac{x x^T}{\sigma^2}$;
```

The natural parameters are thus $\theta_0 = \Sigma^{-1}\mu$ and $\theta_1 = \Sigma^{-1}$; the corresponding sufficient statistics are $T_0(u) = u$ and $T_1(u) = -\frac{1}{2}uu^T$.[7]

Going from the second to last line to the last line we use the following (a special case of the cyclic property of the trace operator):

$$-\tfrac{1}{2}u^T\Sigma^{-1}u = -\tfrac{1}{2}\sum_i u_i(\Sigma^{-1}u)_i \tag{30}$$

$$= -\tfrac{1}{2}\sum_i u_i \sum_j \Sigma^{-1}_{i,j}u_j \tag{31}$$

$$= -\tfrac{1}{2}\sum_i \sum_j \Sigma^{-1}_{i,j}u_i u_j \tag{32}$$

$$= -\tfrac{1}{2}\sum_i \sum_j \Sigma^{-1}_{i,j}(uu^T)_{i,j} \tag{33}$$

$$= \langle \Sigma^{-1}, -\tfrac{1}{2}uu^T \rangle \tag{34}$$

B.1.3. Gamma distribution

Suppose $U \sim \text{Gamma}(\alpha, \beta)$ with support $u \in (0, \infty)$. Then

$$p(u) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\beta u} \tag{35}$$

$$= \exp\big(-\beta u + (\alpha - 1)\log(u) - (\log(\Gamma(\alpha)) - \alpha\log(\beta))\big) \tag{36}$$

The natural parameters of the gamma distribution are thus $\theta_0 = -\beta$ and $\theta_1 = \alpha - 1$; the corresponding sufficient statistics are $T_0(u) = u$ and $T_1(u) = \log(u)$.

The gamma distribution is the conjugate prior for the precision (inverse variance) of a normal distribution. If $U \sim \text{Gamma}(\alpha, \beta)$, then $U^{-1} \sim \text{InverseGamma}(\alpha, \beta)$. It is useful to note that $E[U] = \frac{\alpha}{\beta} = \frac{\theta_1+1}{-\theta_0}$ and (assuming $\alpha > 1$) we also have $E[U^{-1}] = \frac{\beta}{\alpha-1} = \frac{-\theta_0}{\theta_1}$. We shall respectively use these formulas to obtain $E_q[\tau_i^{-2}]$ and $E_q[\tau_i^2]$. in the mean field inference algorithm for the Bayesian derived attention model.

B.2. Conjugate prior updating

B.2.1. Bayesian regression

This section shows the updating rule for Bayesian linear regression (with a conjugate prior). The resulting posterior distribution can be used as the prior for the next observation in turn. This step by step updating rule makes the most sense for modeling learning, but it is equivalent to the batch learning rule that is often presented (one simply adds up sufficient statistics).

For simplicity, we suppress the subscript indicating outcome (e.g. we write μ instead of μ_j). We assume that $y_t \sim \mathcal{N}(x_t^T w, \sigma^2)$ and $w \sim \mathcal{N}(\tilde{\mu}, \tilde{\Sigma})$. We use the symbol "$\propto$" to denote "proportional with respect to w", i.e. equal to the previous expression up to multiplication by a value that does not involve w. Such w-less factors can be ignored when computing the posterior distribution because they do not effect the final shape of the curve (which is a probability density function and hence must be normalized in the end).

$$p(w|y) \propto p(y|w)p(w) \tag{37}$$

$$= \Big(\prod_t p(y_t|w)\Big)p(w) \tag{38}$$

$$= \Big(\prod_t \exp\Big(y_t \tfrac{x_t^T w}{\sigma^2} - \tfrac{y_t^2}{2}\tfrac{1}{\sigma^2} - \tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\big(\tfrac{(x_t^T w)^2}{\sigma^2} + \log(\sigma^2)\big)\Big)\Big)\, p(w) \tag{39}$$

$$\propto \exp\Big(\sum_t y_t \tfrac{x_t^T w}{\sigma^2} - \sum_t \tfrac{1}{2}\tfrac{(x_t^T w)^2}{\sigma^2}\Big)p(w) \tag{40}$$

$$= \exp\Big(\sum_t y_t \tfrac{x_t^T w}{\sigma^2} - \sum_t \tfrac{1}{2}\tfrac{(x_t^T w)^2}{\sigma^2}\Big)\exp\Big(w^T\tilde{\Sigma}^{-1}\tilde{\mu} - \tfrac{1}{2}w^T\tilde{\Sigma}^{-1}w - \tfrac{n_x}{2}\log(2\pi) - \tfrac{1}{2}(\tilde{\mu}^T\tilde{\Sigma}^{-1}\tilde{\mu} + \log|\tilde{\Sigma}|)\Big) \tag{41}$$

$$\propto \exp\Big(\sum_t y_t \tfrac{x_t^T w}{\sigma^2} - \sum_t \tfrac{1}{2}\tfrac{(x_t^T w)^2}{\sigma^2} + w^T\tilde{\Sigma}^{-1}\tilde{\mu} - \tfrac{1}{2}w^T\tilde{\Sigma}^{-1}w\Big) \tag{42}$$

$$= \exp\Big(w^T\big(\tilde{\Sigma}^{-1}\tilde{\mu} + \sum_t \tfrac{x_t y_t}{\sigma^2}\big) - \tfrac{1}{2}w^T\big(\tilde{\Sigma}^{-1} + \sum_t \tfrac{x_t x_t^T}{\sigma^2}\big)w\Big) \tag{43}$$

$$= \exp\Big(\langle \tilde{\Sigma}^{-1}\tilde{\mu} + \sum_t \tfrac{x_t y_t}{\sigma^2}, w\rangle + \langle \tilde{\Sigma}^{-1} + \sum_t \tfrac{x_t x_t^T}{\sigma^2}, -\tfrac{1}{2}ww^T\rangle\Big) \tag{44}$$

From this we recognize that $w|y \sim \mathcal{N}(\mu, \Sigma)$ where $\Sigma = (\tilde{\Sigma}^{-1} + \sum_t \frac{x_t x_t^T}{\sigma^2})^{-1}$ and $\mu = \Sigma(\tilde{\Sigma}^{-1}\tilde{\mu} + \sum_t \frac{x_t y_t}{\sigma^2})$. It is much more convenient to keep track of the natural parameters $\psi_0 = \Sigma^{-1}\mu$ and $\psi_1 = \Sigma^{-1}$ instead of μ and Σ. The posterior value for each natural parameter is the sum of its prior value ($\tilde{\psi}_0 = \tilde{\Sigma}^{-1}\tilde{\mu}$

[7] As in the case of the univariate normal distribution, we assign the $-\frac{1}{2}$ factor to T_1 instead of θ_1 so that the latter is defined as the precision matrix (Σ^{-1}).

and $\tilde{\psi}_1 = \tilde{\Sigma}^{-1}$) and a sufficient statistic ($T_0 = \sum_t \frac{x_t y_t}{\sigma^2}$ and $T_1 = \sum_t \frac{x_t x_t^T}{\sigma^2}$).

Formulating the posterior distribution in terms of its natural parameters (ψ_0 and ψ_1) makes it simple to turn the batch learning procedure described above into a step by step learning rule, which is more appropriate for simulating biological learning. One needs only to update the natural parameters after each new observation, and this is simply a matter of adding sufficient statistics. Algorithm 4 expresses this simple learning rule, with the assumptions that $\tilde{\mu} = 0$ and $\tilde{\Sigma} = \tau^2 \mathbb{I}$, or equivalently $\tilde{\psi}_0 = 0$ and $\tilde{\psi}_1 = \tau^{-2}\mathbb{I}$.

Algorithm 3 is the same thing expressed in terms of the conventional parameters (μ and Σ), written in a way that shows the connection between Bayesian regression and the Rescorla–Wagner learning rule. Let μ denote the posterior mean of w after the first n observations, and μ' denote the posterior mean after observation $n + 1$. We wish to compute μ' by updating μ with prediction error $(y - \hat{y})$ times a learning rate vector (λ), as in the Rescorla–Wagner model:

$$\mu' = \mu + \lambda(y - \hat{y}) \tag{45}$$

We computing the learning rate vector (λ) thus:

$$\begin{aligned}
\mu' - \mu &= \big(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2}\big)^{-1}\big(\Sigma^{-1}\mu + \tfrac{xy}{\sigma^2}\big) - \mu && (46)\\
&= \big(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2}\big)^{-1}\big(\Sigma^{-1}\mu + \tfrac{xy}{\sigma^2}\big) \\
&\quad - \big(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2}\big)^{-1}(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2})\mu && (47)\\
&= \big(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2}\big)^{-1}\big(\Sigma^{-1}\mu + \tfrac{xy}{\sigma^2} - \Sigma^{-1}\mu - \tfrac{xx^T}{\sigma^2}\mu\big) && (48)\\
&= \big(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2}\big)^{-1}\tfrac{x}{\sigma^2}(y - x^T\mu) && (49)\\
&= \big(\Sigma^{-1} + \tfrac{xx^T}{\sigma^2}\big)^{-1}\tfrac{x}{\sigma^2}(y - \hat{y}) && (50)\\
&= \big(\sigma^2\Sigma^{-1} + xx^T\big)^{-1}x(y - \hat{y}) && (51)\\
&= \Sigma\Sigma^{-1}\big(\sigma^2\Sigma^{-1} + xx^T\big)^{-1}x(y - \hat{y}) && (52)\\
&= \Sigma\Big(\big(\sigma^2\Sigma^{-1} + xx^T\big)\Sigma\Big)^{-1}x(y - \hat{y}) && (53)\\
&= \Sigma\Big(\big(\sigma^2 + xx^T\Sigma\big)\Big)^{-1}x(y - \hat{y}) && (54)\\
&= \Sigma\Big(\big(\sigma^2 + x^T\Sigma x\big)\Big)^{-1}x(y - \hat{y}) && (55)\\
&= \frac{\Sigma x}{x^T\Sigma x + \sigma^2}(y - x^T\mu) && (56)
\end{aligned}$$

so $\lambda = \frac{\Sigma x}{x^T\Sigma x + \sigma^2}$, as in Algorithm 3.

Note that it is more straightforward to change the prior on cue weight variance ($\tilde{\Sigma} = \text{diag}(\tau^2)$) in Algorithm 4 than in Algorithm 3, even though the algorithms are equivalent. This is because Algorithm 4 explicitly separates the posterior hyperparameter $\psi_1 = \Sigma^{-1}$ into a prior value ($\tilde{\psi}_1 = \text{diag}(\tau^{-2})$) and the effect of observations (T_1). This is the reason that we use Algorithm 4 instead of the more familiar Algorithm 3 in developing the Bayesian derived attention model: the model is constantly re-estimating the variance of weights for each cue (τ_i^2).

B.2.2. Prior weight variance (τ^2)

In the previous section, we developed the conjugate prior learning algorithm for regression weights (w) when the prior variance (τ^2) is known; now we shall do the opposite and develop updates for τ^2 when w is known. In particular, we assume that weights (w) are generated by Eq. (14) (the inductive assumption) and that they are known. In the following section, we shall remove this second, unrealistic assumption by simultaneously estimating weights and prior variance using a variational Bayesian method.

We will derive the result for a single cue (i) with associations to multiple outcomes ($j = 1, \dots, k$), where k is the number of possible outcomes (category labels) in the task so far (e.g. during the first stage of the learned predictiveness design $k = 2$, while in the second and test stages $k = 4$). It is simpler to perform the calculations in terms of prior precision (τ_i^{-2}) instead of prior variance (τ_i^2), so we shall do so. First we assume that τ_i^{-2} has a gamma prior distribution,[8] $\tau_i^{-2} \sim \text{Gamma}(\tilde{\alpha}, \tilde{\beta})$. Recall that $w_{1,i}, w_{2,i}, \dots, w_{k,i}|\tau_i^2 \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_i^2)$. Thus

$$\begin{aligned}
&p(\tau_i^{-2}|w_{1,i}, w_{2,i}, \dots) \\
&\quad\propto p(w_{1,i}, w_{2,i}, \dots|\tau_i^{-2})p(\tau_i^{-2}) && (57)\\
&\quad= \Big(\prod_{j=1}^{k} p(w_{j,i}|\tau_i^{-2})\Big)p(\tau_i^{-2}) && (58)\\
&\quad= \Big(\prod_{j=1}^{k} \exp\big(-\tfrac{w_{j,i}^2}{2}\tau_i^{-2} - \tfrac{1}{2}\log(2\pi) + \tfrac{1}{2}\log(\tau_i^{-2})\big)\Big)p(\tau_i^{-2}) && (59)\\
&\quad\propto \exp\Big(-\tau_i^{-2}\sum_{j=1}^{k}\tfrac{w_{j,i}^2}{2} + \tfrac{k}{2}\log(\tau_i^{-2})\Big) \\
&\qquad\times \exp\big(-\tau_i^{-2}\tilde{\beta} + \log(\tau_i^{-2})(\tilde{\alpha} - 1)\big) && (60)\\
&\quad= \exp\Big(\big(-\tilde{\beta} - \sum_{j=1}^{k}\tfrac{w_{j,i}^2}{2}\big)\tau_i^{-2} + (\tilde{\alpha} + \tfrac{k}{2} - 1)\log(\tau_i^{-2})\Big) && (61)
\end{aligned}$$

Thus we have the following posterior distribution:

$$\begin{aligned}
&\tau_i^{-2}|w_{1,i}, w_{2,i}, \dots \sim \text{Gamma}(\alpha_i, \beta_i) && (62)\\
&\alpha_i = \tilde{\alpha} + \tfrac{k}{2} && (63)\\
&\beta_i = \tilde{\beta} + \sum_{j=1}^{k}\tfrac{w_{j,i}^2}{2} && (64)
\end{aligned}$$

Algorithm 5 performs this calculation in terms of the natural parameters $\phi_{0,i} = -\beta_i$ and $\phi_{1,i} = \alpha_i - 1$.

B.3. Mean field approximation (variational Bayes)

We shall now briefly describe the variational inference algorithm used to simulate the Bayesian derived attention model; see Blei, Kucukelbir, and McAuliffe (2017) and Broderick et al. (2013) for more about the underlying mathematics. Let $\mathcal{D}$ denote the set of stimuli etc. observed by the learner during the experiment, i.e. $\mathcal{D} = \{x_1, \dots, x_t, y_1, \dots, y_t, \text{psb}_1, \dots, \text{psb}_t, \text{lrn}_1, \dots, \text{lrn}_t\}$. We shall approximate the joint posterior distribution over weights (w) and variances (τ^2) using a variational distribution denoted q[9]:

$$p(w_1, \dots, w_{n_y}, \tau_1^{-2}, \dots, \tau_{n_x}^{-2}|\mathcal{D}) \approx q(w_1, \dots, w_{n_y}, \tau_1^{-2}, \dots, \tau_{n_x}^{-2}) \tag{65}$$

For the mean-field approximation, we assume that q factorizes into independent distributions for each w_j and τ_i^{-2}:

$$q(w_1, \dots, w_{n_y}, \tau_1^{-2}, \dots, \tau_{n_x}^{-2}) = \prod_j q_{w_j}(w_j)\prod_i q_{\tau_i^{-2}}(\tau_i^{-2}) \tag{66}$$

We use E_q and V_q to respectively denote expectation and variance according to q.

[8] This is equivalent to giving τ_i^2 an inverse gamma prior.

[9] Recall that $w_j = (w_{j,1}, w_{j,1}, \dots, w_{j,n_x})$, i.e. the full vector of association weights between cues and outcome j. Also, n_y and n_x are respectively the number of outcomes and cues that the learner has observed so far.

The objective is to find the variational distribution (q) that best approximates the exact joint posterior distribution $p(w_1, \ldots, w_{n_y}, \tau_1^{-2}, \ldots, \tau_{n_x}^{-2} | \mathcal{D})$. An iterative algorithm for finding the optimal q consists of cycling through each variable (each w_j and τ_i^{-2}) and setting its variational distribution as follows (Blei et al., 2017):

$$q_{w_j}(w_j) \propto \exp\Big(E_q[\log(p(w_j | w_{-j}, \tau_1^{-2}, \ldots, \tau_{n_x}^{-2}, \mathcal{D}))]\Big) \tag{67}$$

$$q_{\tau_i^{-2}}(\tau_i^{-2}) \propto \exp\Big(E_q[\log(p(\tau_i^{-2} | w_1, \ldots, w_{n_y}, \tau_{-i}^{-2}, \mathcal{D}))]\Big) \tag{68}$$

where w_{-j} denotes the set of all weight vectors except for w_j and similarly τ_{-i}^{-2} denotes the set of all prior weight precisions except for τ_i^{-2}. We use a streaming version of this algorithm (Broderick et al., 2013) in which updates are made whenever the learner observes new data, i.e. on each trial.

If we follow through the calculations, we see that the variational distribution for w_j is the same as the posterior distribution in ordinary Bayesian regression, except that $E_q[\tau_i^{-2}]$ is used instead of a known value of τ_i^{-2}:

$$E_q[\log(p(w_j | w_{-j}, \tau_1^{-2}, \ldots, \tau_{n_x}^{-2}, \mathcal{D}))]$$

$$= E_q[\log(p(w_j | \tau_1^{-2}, \ldots, \tau_{n_x}^{-2}, \mathcal{D}))] \tag{69}$$

$$= E_q[\langle \sum_t \tfrac{x_t y_{j,t}}{\sigma^2}, w_j \rangle + \langle \mathrm{diag}(\tau^{-2}) + \sum_t \tfrac{x_t x_t^T}{\sigma^2}, -\tfrac{1}{2} w_j w_j^T \rangle + C] \tag{70}$$

$$= \langle \sum_t \tfrac{x_t y_{j,t}}{\sigma^2}, w_j \rangle + \langle \mathrm{diag}(E_q[\tau^{-2}]) + \sum_t \tfrac{x_t x_t^T}{\sigma^2}, -\tfrac{1}{2} w_j w_j^T \rangle + C \tag{71}$$

Similarly, the variational distribution for τ_i^{-2} is the same as its posterior distribution if w were known, simply substituting $E_q[w_{j,i}^2]$ for the unknown $w_{j,i}^2$:

$$E_q[\log(p(\tau_i^{-2} | w_1, \ldots, w_{n_y}, \tau_{-i}^{-2}, \mathcal{D}))]$$

$$= E_q[\log(p(\tau_i^{-2} | w_{1,i}, \ldots, w_{n_y,i}))] \tag{72}$$

$$= E_q[(\tilde{\phi}_{0,i} - \tfrac{1}{2} \sum_j w_{j,i}^2) \tau_i^{-2} + (\tilde{\phi}_{1,i} + \tfrac{n_y}{2}) \log(\tau_i^{-2}) + C] \tag{73}$$

$$= (\tilde{\phi}_{0,i} - \tfrac{1}{2} \sum_j E_q[w_{j,i}^2]) \tau_i^{-2} + (\tilde{\phi}_{1,i} + \tfrac{n_y}{2}) \log(\tau_i^{-2}) + C \tag{74}$$

These calculations are the basis for mean field variational Bayes inference of the Bayesian derived attention model (Algorithm 5).

References

Anderson, B. A., Laurent, P. A., & Yantis, S. (2011). Value-driven attentional capture. *Proceedings of the National Academy of Sciences*, *108*(25), 10367–10371. http://dx.doi.org/10.1073/pnas.1104047108, URL http://www.pnas.org.colorado.idm.oclc.org/content/108/25/10367.

Beesley, T., & Le Pelley, M. E. (2011). The influence of blocking on overt attention and associability in human learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(1), 114–120. http://dx.doi.org/10.1037/a0019526, URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0019526.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. http://dx.doi.org/10.1080/01621459.2017.1285773, URL https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1285773.

Bouton, M. E., & Bolles, R. C. (1979). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, *10*(4), 445–466. http://dx.doi.org/10.1016/0023-9690(79)90057-2, URL https://linkinghub.elsevier.com/retrieve/pii/0023969079900572.

Broderick, T., Boyd, N., Wibisono, A., Wilson, A. C., & Jordan, M. I. (2013). Streaming variational bayes. *Advances in Neural Information Processing Systems*, *26*.

Dayan, P., & Kakade, S. (2001). Explaining Away in Weight Space. *Advances in Neural Information Processing Systems*, 451–457.

Delamater, A. R., & Westbrook, R. F. (2014). Psychological and neural mechanisms of experimental extinction: A selective review. *Neurobiology of Learning and Memory*, *108*, 38–51. http://dx.doi.org/10.1016/j.nlm.2013.09.016, URL https://linkinghub.elsevier.com/retrieve/pii/S1074742713001937.

Esber, G. R., & Haselgrove, M. (2011). Reconciling the influence of predictiveness and uncertainty on stimulus salience: a model of attention in associative learning. *Proceedings of the Royal Society B: Biological Sciences*, *278*(1718), 2553–2561. http://dx.doi.org/10.1098/rspb.2011.0836, URL http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2011.0836.

Estes, W. K., & Skinner, B. F. (1941). Some quantitative properties of anxiety. *Journal of Experimental Psychology*, *29*(5), 390–400. http://dx.doi.org/10.1037/h0062283, URL http://content.apa.org/journals/xge/29/5/390.

Frey, P. W., & Sears, R. J. (1978). Model of conditioning incorporating the Rescorla-Wagner associative axiom, a dynamic attention process, and a catastrophe rule. *Psychological Review*, *85*(4), 321–340.

Gershman, S. J. (2015). A unifying probabilistic view of associative learning. In J. Diedrichsen (Ed.), *PLoS Computational Biology*, *11*(11), Article e1004567. http://dx.doi.org/10.1371/journal.pcbi.1004567, URL http://dx.plos.org/10.1371/journal.pcbi.1004567.

Gluck, M. A., & Bower, G. H. (1988). From conditioning to category learning: An adaptive network model. *Journal of Experimental Psychology: General*, *117*(3), 227–247.

Hiroto, D. S., & Seligman, M. E. (1975). Generality of learned helplessness in man. *Journal of Personality and Social Psychology*, *31*(2), 311–327. http://dx.doi.org/10.1037/h0076270, URL http://doi.apa.org/getdoi.cfm?doi=10.1037/h0076270.

Joukhador, J., Maccallum, F., & Blaszczynski, A. (2003). Differences in cognitive distortions between problem and social gamblers. *Psychological Reports*, *92*, 1203–1214.

Kamin, L. (1968). *"Attention-like" processes in classical conditioning* (pp. 9–31). Coral Gables, Florida: University of Miami Press.

Kruschke, J. K. (1996). Base rates in category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 3–26.

Kruschke, J. K. (2001). Toward a unified model of attention in associative learning. *Journal of Mathematical Psychology*, *45*(6), 812–863. http://dx.doi.org/10.1006/jmps.2000.1354, URL http://www.sciencedirect.com/science/article/pii/S0022249600913543.

Kruschke, J. K., & Blair, N. J. (2000). Blocking and backward blocking involve learned inattention. *Psychonomic Bulletin & Review*, *7*(4), 636–645. http://dx.doi.org/10.3758/BF03213001, URL http://www.springerlink.com/index/10.3758/BF03213001.

Kuo, L., & Mallick, B. (1998). Variable selection for regression models. *Sankhyā*, 65–81, Publisher: JSTOR.

Landauer, T. K. (1986). How much do people remember? some estimates of the quantity of learned information in long-term memory how much do people remember? some estimates of the quantity of learned information in long-term memory. *Cognitive Science*, *10*(4), 477–493. http://dx.doi.org/10.1207/s15516709cog1004_4, Retrieved 2022-08-29 http://doi.wiley.com/10.1207/s15516709cog1004_4.

Le Pelley, M. E., Beesley, T., & Suret, M. B. (2007). Blocking of human causal learning involves learned changes in stimulus processing. *Quarterly Journal of Experimental Psychology*, *60*(11), 1468–1476. http://dx.doi.org/10.1080/17470210701515645, URL http://journals.sagepub.com/doi/10.1080/17470210701515645.

Le Pelley, M. E., & McLaren, I. P. L. (2003). Learned associability and associative change in human causal learning. *The Quarterly Journal of Experimental Psychology Section B*, *56*(1b), 68–79. http://dx.doi.org/10.1080/02724990244000179, URL http://journals.sagepub.com/doi/10.1080/02724990244000179.

Le Pelley, M. E., Mitchell, C. J., Beesley, T., George, D. N., & Wills, A. J. (2016). Attention and associative learning in humans: An integrative review. *Psychological Bulletin*, *142*(10), 1111–1140. http://dx.doi.org/10.1037/bul0000064, URL http://doi.apa.org/getdoi.cfm?doi=10.1037/bul0000064.

Le Pelley, M. E., Mitchell, C. J., & Johnson, A. M. (2013). Outcome value influences attentional biases in human associative learning: Dissociable effects of training and instruction. *Journal of Experimental Psychology: Animal Behavior Processes; Washington*, *39*(1), 39, URL http://search.proquest.com/docview/1270551815?pq-origsite=summon&.

Le Pelley, M. E., Pearson, D., Porter, A., Yee, H., & Luque, D. (2018). Oculomotor capture is influenced by expected reward value but (maybe) not predictiveness. *Quarterly Journal of Experimental Psychology*, http://dx.doi.org/10.1080/17470218.2017.1313874, 17470218.2017.1, URL http://journals.sagepub.com/doi/10.1080/17470218.2017.1313874.

Le Pelley, M. E., Suret, M. B., & Beesley, T. (2009). Learned predictiveness effects in humans: A function of learning, performance, or both? *Journal of Experimental Psychology: Animal Behavior Processes*, *35*(3), 312–327. http://dx.doi.org/10.1037/a0014315, URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0014315.

Lochmann, T., & Wills, A. (2003). Predictive history in an allergy prediction task. *Vol. 3*, In *Proceedings of EuroCogSci* (pp. 217–222).

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, *82*(4), 276–298. http://dx.doi.org/10.1037/h0076778, URL http://content.apa.org/journals/rev/82/4/276.

Medin, D. L., & Edelson, S. M. (1988). Problem structure and the use of base-rate information from experience. *Journal of Experimental Psychology: General*, *117*, 68–85.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032, Publisher: Taylor & Francis.

Mitchell, C. J., Griffiths, O., Seetoo, J., & Lovibond, P. F. (2012). Attentional mechanisms in learned predictiveness. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(2), 191–202. http://dx.doi.org/10.1037/a0027385, URL http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027385.

Neal, R. M. (1996). Bayesian learning for neural networks. *Lecture notes in statistics*: *Vol. 118*, New York, NY: Springer New York, http://dx.doi.org/10.1007/978-1-4612-0745-0, URL http://link.springer.com/10.1007/978-1-4612-0745-0,

Paskewitz, S., & Jones, M. (2020). Dissecting EXIT. *Journal of Mathematical Psychology*, *97*, Article 102371. http://dx.doi.org/10.1016/j.jmp.2020.102371, URL https://linkinghub.elsevier.com/retrieve/pii/S0022249620300419.

Paskewitz, S., Stoddard, J., & Jones, M. (2022). Explaining the return of fear with revised Rescorla-Wagner models. *Computational Psychiatry*, *6*(1), Publisher: Ubiquity Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Shanks, D. R. (1985). Forward and Backward Blocking in Human Contingency Judgement. *The Quarterly Journal of Experimental Psychology Section B*, *37*(1b), 1–21. http://dx.doi.org/10.1080/14640748508402082, URL http://journals.sagepub.com/doi/10.1080/14640748508402082.

Swan, J. A., & Pearce, J. M. (1988). The orienting response as an index of stimulus associability in rats. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(3), 292–301. http://dx.doi.org/10.1037/0097-7403.14.3.292, URL http://doi.apa.org/getdoi.cfm?doi=10.1037/0097-7403.14.3.292.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*(1), 102, Publisher: American Psychological Association.

Wilson, R. C., Nassar, M. R., & Gold, J. I. (2013). A mixture of delta-rules approximation to Bayesian inference in change-point problems. In T. Behrens (Ed.), *PLoS Computational Biology*, *9*(7), Article e1003150. http://dx.doi.org/10.1371/journal.pcbi.1003150, URL https://dx.plos.org/10.1371/journal.pcbi.1003150.