Locally Bayesian Learning

John K. Kruschke (kruschke@indiana.edu)
Department of Psychological and Brain Sciences; Indiana University
Bloomington IN 47405 USA

Abstract

This article is concerned with trial-by-trial, online learning of cue-outcome mappings. In models structured as successions of component functions, an external target can be back-propagated such that the lower layer's target is the input to the higher layer that maximizes the probability of the higher layer's target. Each layer then does locally Bayesian learning. The resulting parameter updating is not globally Bayesian, but can better capture human behavior. The approach is implemented for an associative learning model that first maps inputs to attentionally filtered inputs, and then maps attentionally filtered inputs to outputs. The model is applied to the human-learning phenomenon called highlighting, which is challenging to other extant Bayesian models, including the rational model of Anderson, the Kalman filter model of Dayan and Kakade et al., the noisy-OR model of Tenenbaum and Griffiths et al., and the sigmoid-belief networks of Courville et al. Further details and applications are provided by Kruschke (in press); the present article reports new simulations of the Kalman filter and rational model.

Cognition Modeled as a Succession of Transformations

Cognitive models are often conceived to be successions of transformations from an input representation, through various internal representations, to an output or response representation. Each transformation is a formal operation, typically having various parameter values that are tuned by experience. A well-know example is Marr's (1982) modeling of vision as a succession from a representation of image intensity to a "primal sketch" to a "$2\frac{1}{2}$-D sketch" to a 3-D model representation.

Globally Bayesian Learning

In Bayesian approaches to cognitive modeling, each transformation in the hierarchy takes an input and generates a distribution of possible outputs. Figure 1 shows the input x_ℓ at layer ℓ being transformed into the output y_ℓ, which has a probability distribution $p(y_\ell)$. The input at the first layer is denoted x_1, and the output at the last layer is denoted y_L. The specifics of the distribution are governed by the values of the parameters θ_ℓ.

Each value of the parameters θ_ℓ represents a particular hypothesis about how inputs (stimuli) and outputs (outcomes or responses) are related. The combinations of all possible values of θ_ℓ span the possible beliefs of the model. The core ontological notion in Bayesian approaches is that knowledge consists of the degree of belief in each possible value of the parameters θ_ℓ. That distribution of beliefs in each layer is denoted $p(\theta_\ell)$.

The system starts with some prior distribution of belief over the joint hypotheses, $p(\theta_L,\ldots,\theta_1)$. That distribution is updated each time that an input-output datum is experienced. For input x_1, suppose that the correct outcome, as observed in the environment, is t_L. Bayes' theorem indicates that the appropriate beliefs after witnessing the item $\langle t_L, x_1\rangle$ are

$$p(\theta_L,\ldots,\theta_1|t_L,x_1) = \frac{p(t_L|\theta_L,\ldots,\theta_1,x_1)\; p(\theta_L,\ldots,\theta_1)}{\int d\theta_L\ldots d\theta_1\; p(t_L|\theta_L,\ldots,\theta_1,x_1)\; p(\theta_L,\ldots,\theta_1)} \quad (1)$$

The probability of the outcome given the input, $p(t_L|\theta_L,\ldots,\theta_1,x_1)$, is determined by the particular functions in each layer. The updating of the belief distribution over the joint parameter space is referred to as globally Bayesian learning.

Locally Bayesian Learning

An alternative approach comes from considering the local environment of each layer. Each layer only has contact with its own input and output. If a layer had a specific target and input, then the layer could apply Bayesian updating to its own parameters, without worrying about the other layers.

A local updating scheme proceeds as follows. When an input x_1 is presented at the bottom layer, the input is propagated up the layers. The input to layer $\ell+1$ is the expected value of

$$\begin{array}{l} y_{\ell+1} \sim p(y_{\ell+1}|\theta_{\ell+1},x_{\ell+1}) \\ \uparrow\ \theta_{\ell+1} \sim p(\theta_{\ell+1}) \\ x_{\ell+1} \end{array} \qquad \begin{array}{l} y_\ell \sim p(y_\ell|\theta_\ell,x_\ell) \\ \uparrow\ \theta_\ell \sim p(\theta_\ell) \\ x_\ell \end{array}$$

Figure 1. Architecture of successive functions. Vertical arrows indicate a mapping from input to output within a layer, parameterized by θ. The notation "$\theta \sim p(\theta)$" means that θ is distributed according to the probability distribution $p(\theta)$. In the globally Bayesian approach, $x_{\ell+1} = y_\ell$. In the locally Bayesian approach, $x_{\ell+1} = \bar{y}_\ell$.

the output of module ℓ:

$$x_{\ell+1} = \bar{y}_\ell = \int dy_\ell \; y_\ell \; p(y_\ell|x_\ell) \tag{2}$$

Equation 2 is applied recursively up the sequence of layers, so every layer has a specific input.

A target output t_L is provided at the final output layer. The belief probabilities for layer $\ell = L$ are updated according to Bayes theorem,

$$p(\theta_\ell|t_\ell,x_\ell) \;=\; \frac{p(t_\ell|\theta_\ell,x_\ell)\,p(\theta_\ell)}{\int d\theta_\ell \; p(t_\ell|\theta_\ell,x_\ell)\,p(\theta_\ell)} \tag{3}$$

where $x_\ell = \bar{y}_{\ell-1}$ as in Equation 2.

Then a target is selected for the next layer down. This target for the lower layer is the input to the higher layer that maximizes the probability of the higher-layer target. In other words, when the ℓ^{th} layer has a target vector t_ℓ, we choose the next lower target as:

$$\begin{aligned} t_{\ell-1} &= \underset{x^*_\ell}{\text{argmax}} \; p(t_\ell|x^*_\ell) \\ &= \underset{x^*_\ell}{\text{argmax}} \int d\theta_\ell \; p(t_\ell|\theta_\ell,x^*_\ell) \; p(\theta_\ell|t_\ell,x_\ell) \end{aligned} \tag{4}$$

Equation 4 simply states that the target for the lower layer is the input to the upper layer that would maximize the probability of the upper layer's target. The variable x^*_ℓ is given a superscript star to distinguish it from the input value $x_\ell = \bar{y}_{\ell-1}$.

The targets can then be propagated down the layers by recursively applying Equations 3 and 4. For each layer, the beliefs are updated and then a target is determined for the layer below.

An interesting quality of this algorithm is that the target received by a lower layer depends not only on the actual exterior target but also on what the upper layers have learned until that point in training. (As mentioned before, I am assuming trial-by-trial, online learning.) The target for the lower layer is selected to be maximally consistent with what the upper layers have already learned. In this way, the upper layer changes the data to be consistent with its beliefs before the lower layer changes its beliefs to be consistent with the data. As a consequence, the system is not globally Bayesian. Nevertheless, simulations below illustrate that this is an important characteristic for capturing human learning.

A Challenging Behavior: Highlighting

In typical associative learning experiments, people must learn which button to press in response to some simple cues presented on a computer screen. The cues could be simple words, such as "brain" and "world." In a learning trial, the cues are presented, the learner presses the button that s/he thinks is correct, and then the correct response is displayed. The learner studies the cues and correct response and then moves on to the next trial. At first the learner is guessing, but predictive accuracy improves with training.

In the *highlighting* procedure, people are initially trained on cases in which two cues, denoted PE and I, indicate outcome E. Later in training, people are also trained on cases in which a new cue PL along with old cue I indicate a new outcome L. Figure 2 shows the symmetric structure of the cue-outcome relations in highlighting. Notice that cue I is an Imperfect predictor because both outcomes E and L can occur (on different trials) when I occurs. Cue PE is a Perfect predictor of the Earlier trained outcome E, and cue PL is a Perfect predictor of the Later trained outcome L.

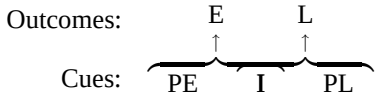

Figure 2. Symmetric structure of cue-outcome relations in the highlighting procedure. Cases of PE.I→E are trained earlier than cases of I.PL→L, but with equal base rates overall.

If people learn the simple underlying symmetry of the cue-outcome correspondences, then when they are tested with cue I by itself, they should choose outcomes E and L equally often. In fact, there is a strong tendency to choose outcome E. This response bias is not a general primacy effect, however, because when people are tested with the pair of cues PE and PL, they prefer outcome L. Apparently, cue PL has been highlighted during learning I.PL→L, so that cue I is not associated strongly with L. But PL apparently is strongly associated with PL, even more than PE is associated with E.

Table 1 shows details of a canonical highlighting design. The learner first sees trials of cues I and PE indicating outcome E, denoted I.PE→E. One "epoch" of trials consists of the items in that phase presented in random order. In the second and third phases of training, trials of I.PL→L are intermixed. The canonical highlighting design equalizes the frequencies of the early and late outcomes. Notice in the table that when $N_3 = N_2 + N_1$, the total number of I.PE→E trials is $3N_1 + 4N_2$, which equals the total number of I.PL→L trials. This equality of base rates distinguishes highlighting from the "inverse base rate effect" reported by Medin and Edelson (1988), which uses only the second phase of Table 1, i.e., $N_1 = 0$ and $N_3 = 0$. The equality of base rates emphasizes that highlighting is an order-of-learning effect, not a base rate effect. Simulations described below show that various Bayesian models of learning predict $p(E|I) = p(E|PE.PL) = .5$, con-

Table 1
Canonical highlighting design.

Phase	# Epochs	Items × Frequency	
First	N_1	I.PE→E ×2	
Second	N_2	I.PE→E ×3	I.PL→L ×1
Third	$N_3 = N_2 + N_1$	I.PE→E ×1	I.PL→L ×3
Test		PE.PL→? (L) I→? (E)	

Note: An item is shown in the format, Cues→Correct Response. In the test phase, typical response tendencies are shown in parentheses.

trary to human behavior.

Highlighting has been obtained in many different experiments using different stimuli, procedures, and cover stories, such as fictitious disease diagnosis (Kruschke, 1996; Medin & Edelson, 1988), random word association (Dennis & Kruschke, 1998; Kruschke, Kappenman, & Hetrick, 2005), and geometric figure association (Fagot, Kruschke, Dépy, & Vauclair, 1998). Many other published experiments have obtained the inverse base rate effect for different relative frequencies and numbers of training blocks (e.g., Juslin, Wennerholm, & Winman, 2001; Medin & Bettger, 1991; Shanks, 1992). I have run several (unpublished) experiments in my lab in which $N_1 = 0$ and $N_2 = N_3$, and in all of these experiments robust highlighting has been obtained.

Predictions of Various Bayesian Models Applied to Highlighting

The remainder of this brief article shows that several Bayesian models of learning cannot accommodate the highlighting effect, but a simple locally Bayesian model does. There is not space here to discuss several other phenomena in human learning that are difficult for globally Bayesian models but which can be addressed by a locally Bayesian model. These other phenomena, and full details of the locally Bayesian model summarized in the next section, are discussed by Kruschke (in press).

Locally Bayesian Learning

An illustrative implementation of the locally Bayesian learning scheme is now presented. Figure 3 shows that the model architecture has two layers of associative weights. Input nodes correspond with stimulus cues, and output nodes correspond to response choices. An essential aspect of the model is that the intermediate ("hidden") nodes represent attentionally modulated copies of the corresponding input cues. The weight from a cue to the corresponding hidden node is constrained to be positive, but weights from cues to non-corresponding hidden nodes can be zero or negative. This allows the network to entertain hypotheses that some cues can inhibit attention to other cues.

The weights from the hidden nodes to the outcome nodes can have positive, zero, or negative values. Within each layer,

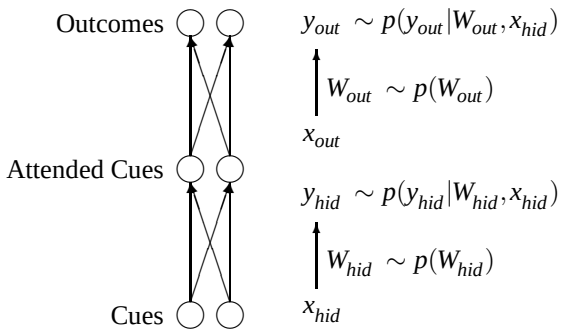

Figure 3. Architecture for the simple model of associative learning. When locally Bayesian, the input to the outcome layer is the mean output of the hidden layer, i.e., $x_{out} = \bar{y}_{hid}$.

a hypothesis is a particular weight matrix, W. The model is supplied with a large number of hypothetical weight matrices. The prior over the hidden weight hypotheses is uniform, and the prior over the output weight hypotheses is Gaussian. The prior therefore is completely neutral and provides no preferential treatment for any cue or outcome.

The upper row of Figure 4 shows the results after training the locally Bayesian model in the highlighting procedure with $N_1 = 1$, $N_2 = 2$ and $N_3 = 3$ in Table 1. The left panel simply lists the training items in the order presented. The right panel shows the choice preference of the model, where it can be seen that the model shows a robust highlighting effect: $p(E|I) > .5$ and $p(E|PE.PL) < .5$.

The panel labeled "Hidden Weights" shows that the model has shifted all its belief to hypotheses in which cue PL inhibits hidden node I: The dotted line marked with a star, and labeled hidI←PL, has all its belief probability loaded over the weight value of -5. But cue PE does *not* symmetrically inhibit hidden node I: The solid line marked with a diamond, and labeled hidI←PE, has all its belief probability loaded over the weight value of 0, not -5.

The panel labeled "Outcome Weights" shows that the model believes in hypotheses for which there is a positive connection from hidden node I to outcome E, but does not believe in hypotheses for which there is a negative connection from hidden node I to outcome E: The line marked with a square and labeled E←hidI has marginal belief probability greater than .4 over weight value $+5$, but has marginal belief probability close to 0 over weight value -5. In other words, the locally Bayesian model has learned to believe in hypotheses that are *not* symmetric across cues.

The locally Bayesian model learns asymmetric beliefs because of the internal targets it generates while learning I.PL→L. Because it has previously learned that cue I indicates outcome E, not the currently correct outcome L, the target at the hidden layer that is most consistent with the target has hidden node I de-activated. The lower layer then learns to believe in hypotheses that suppress hidden node I when cue PL is present.

Globally Bayesian Learning

The simplistic implementation of the locally Bayesian model permits the analogous globally Bayesian model to be exactly implemented. The globally Bayesian model crosses every hidden-weight matrix with every output-weight matrix to create a large joint hypothesis space. If the locally Bayesian model has N^{hid} hidden-weight hypotheses and N^{out} output-weight hypotheses, then it has $N^{\text{hid}} + N^{\text{out}}$ hypotheses altogether. The globally Bayesian model, on the other hand, has $N^{\text{hid}} \times N^{\text{out}}$ hypotheses. The prior on the joint space is also just the product of the local priors, so that the marginal priors on the joint space are identical to the local marginal priors.

The lower row of Figure 4 reveals that the globally Bayesian model shows no highlighting effect whatsoever, and symmetrically distributes its beliefs. The globally Bayesian model believes in hypotheses that have cues PE and PL equally associated with their respective outcomes, and have cue I neutrally or equally associated with both outcomes.

Interestingly, it turns out that the globally Bayesian model learns the training items more slowly than the locally Bayesian model. In other words, accuracy on the training

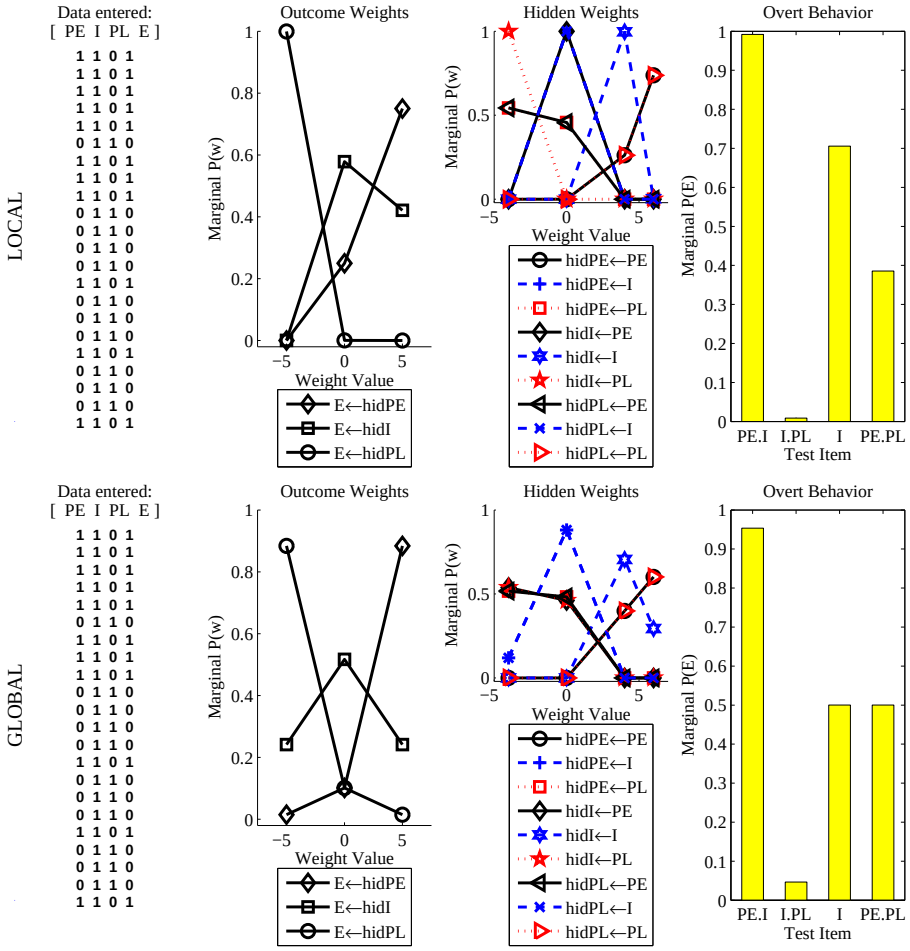

Figure 4. Upper Row: The locally Bayesian model trained in the highlighting procedure (with $N_1 = 1, N_2 = 2, N_3 = 3$ in Table 1). *Lower Row:* The globally Bayesian model trained the same.

items is better in the locally Bayesian model, throughout training. (A hint of this can be seen by comparing the upper and lower rows of Figure 4, but the difference appears to be weak because accuracies are near asymptote by this point in training.) One reason for the relative retardation in the global model is that the global model retains some belief distributed over many candidate hypotheses, and this dilutes performance. A more detailed discussion can be found in Kruschke (in press).

Rational Model

The rational model of category learning, invented by Anderson (1990), is a Bayesian clustering algorithm. Each cluster represents a distribution of beliefs over candidate probabilities of feature values in that cluster. For example, if one stimulus dimension is presence or absence of feature PE, a cluster might have .10 belief that the probability of PE presence is .3, and .15 belief that the probability of PE presence is .4, and so forth. The degree of belief in the conjoined features of a stimulus is simply assumed to be the product of the beliefs in the individual features.

The belief distributions on each feature are continuous and parameterized as Dirichlet distributions. These distributions have one parameter per feature value. A convenient characteristic of these distributions is that when a stimulus is added to a cluster and the cluster's beliefs are updated, the posterior distribution again has the form of a Dirichlet distribution. The parameters of the posterior distribution are simply the prior parameters incremented by 1 wherever a feature was present in the added stimulus.

When a stimulus is presented to the rational model, the model computes the probability of each cluster given the stimulus. One of the candidate clusters is always the novel cluster which has a uniform prior. The stimulus is added to whichever cluster has highest probability. If it is added to the until-then novel cluster, a new novel cluster is recruited for subsequent trials.

Predictions regarding missing features are determined by computing, in each cluster, the probabilities of the values of the missing feature, and adding those probabilities across clusters, weighted by the probability of the cluster given the presented features. This is the normative Bayesian approach: The prediction is the average of the predictions of each hypothesis, weighted by the degree of belief in each hypothesis.

Figure 5 shows the results of applying the rational model to the highlighting procedure with $N_1 = 1$, $N_2 = 2$, $N_3 = 3$

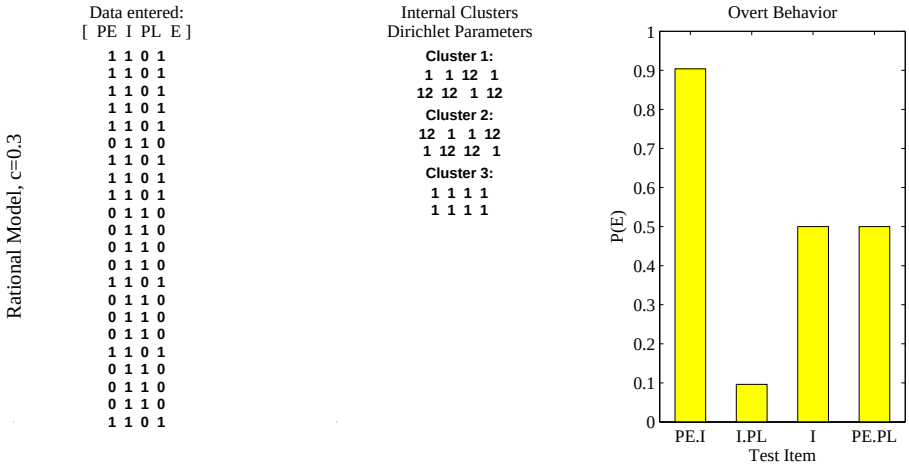

Figure 5. Rational model (Anderson, 1990) trained in the highlighting procedure (with $N_1 = 1$, $N_2 = 2$, $N_3 = 3$ in Table 1).

in Table 1. The right panel reveals that the model shows no highlighting effect. The middle panel shows the state of the cluster nodes at the end of training. The model has recruited two clusters. One cluster represents all the I.PE→E items, and the other cluster represents all the I.PL→L items. (The third cluster is the omnipresent novel cluster.) Because the clusters are completely symmetric with respect to the cues, the predicted behavior is also.[1]

The Kalman Filter

The top layer of the simplistic locally Bayesian model is closely related to a Kalman filter, which was introduced to associative learning researchers by Sutton (1992) and has been used to model some aspects of attention in learning by Dayan, Kakade and collaborators (e.g., Dayan & Kakade, 2001; Dayan et al., 2000; Dayan & Yu, 2003; Kakade & Dayan, 2002). In a Kalman filter, continuous-scale outcomes are computed as a weighted sum of input cues. The weighting coefficients have prior distributions defined as multivariate normal. The Kalman filter uses Bayesian updating to adjust the probability distribution on the weights (Meinhold & Singpurwalla, 1983). Because the model is linear, the posterior distributions on the weights are also multivariate normal, and the Kalman filter equations elegantly express the posterior mean and covariance as a simple function of the prior mean and covariance. One difference between the models is that the Kalman filter can add uncertainty to the weight distributions on every trial. Because of the accumulation of noise across trials, the Kalman filter can exhibit some trial order effects. Typically the amount of uncertainty added is a constant.

Figure 6 shows the behavior of the Kalman filter when applied to highlighting (with $N_1 = 1$, $N_2 = 2$, $N_3 = 3$ in Table 1). The format of the figure matches that used in reports by Dayan et al. The top panel of Figure 6 shows the mean weight (i.e., the mean of the Gaussian distribution of beliefs over possible weight values) on each cue, at the beginning of each epoch of training. The means start unbiased at zero. At the end of training, the mean on cue I is nearly zero, and the means on cues PE and PL are nearly equal (but opposite) magnitude. Therefore, when presented with items I or PE.PL, the model predicts nearly 50-50 outcomes. This behavior can be modulated somewhat by the amount of uncertainty that is added on each trial, but increased uncertainty can be counteracted by longer training.

The lower panel of Figure 6 indicates the "uncertainties" on each cue, which are simply the variances (diagonal elements of the covariance matrix) of the Gaussian belief distribution. As training progresses, uncertainty decreases, which indicates that beliefs sharpen-up over particular weight values. The graph indicates that uncertainties are very nearly symmetric at the end of training.

The locally Bayesian model extends the Kalman-filter approach by pre-pending an attentional learning layer. Whereas the Kalman filter learns about the cues in their totality, the upper layer of the locally Bayesian model learns only about attentionally filtered cues at the hidden layer. The attentional filtration depends on the temporal order of training items. The temporal dependencies of the two models are not incompatible; future extensions of the models could incorporate both the uncertainty accumulation of the Kalman filter model with the attentional selection of the locally Bayesian model.

Other Bayesian Models

Tenenbaum and collaborators (e.g., Sobel, Tenenbaum, & Gopnik, 2004; Tenenbaum & Griffiths, 2003) have developed Bayesian models in which the hypotheses are noisy-OR gates. The models handily address some aspects of rapid learning, but are not able to exhibit highlighting because the models have no time dependencies. That is, all that matters to the model is the overall frequency of the training items, not their training order.

Courville and colleagues (Courville, Daw, Gordon, & Touretzky, 2004; Courville, Daw, & Touretzky, 2005) con-

[1] Anderson (1990) reported that the rational model can capture some aspects of the "inverse base rate effect," which is the procedure of Table 1 with $N_1 = 0$ and $N_3 = 0$. The model works in that situation because the more frequent cluster has a tighter Dirichlet distribution than the less frequent cluster. But with the equal overall frequencies in canonical highlighting, the two clusters have equal variances.

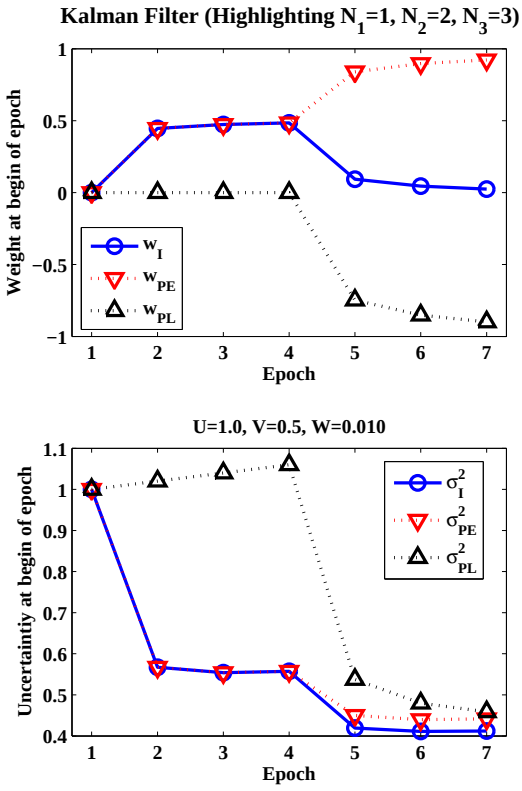

Figure 6. Kalman filter (Dayan et al., 2000) trained in the highlighting procedure (with $N_1 = 1$, $N_2 = 2$, $N_3 = 3$ in Table 1).

ceptualized both the cues and outcomes as effects to be predicted by latent causes (analogous to the clusters in the rational model). In their approach, a hypothesis is a set of weights from latent causes to cues and outcomes, with the probability of each effect being determined by a sigmoidal function of the summed weights from activated latent causes. The hypothesis space consists of many weight combinations, and Bayesian learning shifts belief probability among the hypotheses. Courville et al. (2004) showed that the approach can account for the dependency of conditioned inhibition on the number of trials of training, by virtue of the prior probabilities being gradually overwhelmed by training data. But the model would not be able to exhibit highlighting because it has no time dependencies.

Conclusion

The locally Bayesian attention model produces highlighting (and other challenging phenomena) by generating internal target data that depend on current beliefs. When learning a cue-outcome correspondence, the model first generates internal representations that are maximally consistent with its current (upper-layer) beliefs before updating its (lower-layer) beliefs. Thus, the locally Bayesian model changes the data to fit its beliefs before changing its beliefs to fit the data. Alas, people seem to behave that way too.

Acknowledgments

Supported in part by grant BCS-9910720 from the National Science Foundation.

References

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Courville, A. C., Daw, N. D., Gordon, G. J., & Touretzky, D. S. (2004). Model uncertainty in classical conditioning. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems* (Vol. 16, pp. 977–984). Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in neural information processing systems* (Vol. 17, p. **). Cambridge, MA: MIT Press.

Dayan, P., & Kakade, S. (2001). Explaining away in weight space. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems* (Vol. 13, pp. 451–457). Cambridge, MA: MIT Press.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

Dayan, P., & Yu, A. J. (2003). Uncertainty and learning. *IETE (Institution of Electronics and Telecommunication Engineers, India) Journal of Research*, *49*, 171–182.

Dennis, S., & Kruschke, J. K. (1998). Shifting attention in cued recall. *Australian Journal of Psychology*, *50*, 131–138.

Fagot, J., Kruschke, J. K., Dépy, D., & Vauclair, J. (1998). Associative learning in baboons (papio papio) and humans (homo sapiens): species differences in learned attention to visual features. *Animal Cognition*, *1*, 123–133.

Juslin, P., Wennerholm, P., & Winman, A. (2001). High level reasoning and base-rate use: Do we need cue competition to explain the inverse base-rate effect? *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 849–871.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review*, *109*, 533–544.

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

Kruschke, J. K. (in press). Locally Bayesian learning with applications to retrospective revaluation and highlighting. *Psychological Review*.

Kruschke, J. K., Kappenman, E. S., & Hetrick, W. P. (2005). Eye gaze and individual differences consistent with learned attention in associative blocking and highlighting. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *31*, 830–845.

Marr, D. (1982). *Vision*. San Francisco: W. H. Freeman.

Medin, D. L., & Bettger, J. G. (1991). Sensitivity to changes in base-rate information. *American Journal of Psychology*, *104*, 311–332.

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.

Meinhold, R. J., & Singpurwalla, N. D. (1983). Understanding the Kalman filter. *American Statistician*, *37*(2), 123–127.

Shanks, D. R. (1992). Connectionist accounts of the inverse base-rate effect in categorization. *Connection Science*, *4*, 3–18.

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303–333.

Sutton, R. S. (1992). Gain adaptation beats least squares? In *Proceedings of the seventh annual Yale workshop on adaptive and learning systems* (p. 161-166). New Haven, CT: Yale University.

Tenenbaum, J. B., & Griffiths, T. L. (2003). Theory-based causal inference. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems* (Vol. 15, pp. 35–42). Cambridge, MA: MIT Press.