Notes On Modeling Memory With Student Feedback

Michael C. Mozer
Department of Computer Science
University of Colorado
Boulder, CO 80309-0430 USA
mozer@colorado.edu

In cognitive psychology, theories of human learning aim to describe the basic laws of acquisition and forgetting. For elegance and parsimony, theories focus on predicting average learning performance for a population of individuals (e.g., college students) and a collection of homogenous items (e.g., foreign language vocabulary to be remembered). Because of the focus on aggregate data, theories typically do an underwhelming job of predicting the memory strength of a specific individual for a particular item. Individual variability is generally treated as noise. (In these notes, we focus on fact recall, but the approach could equally well be applied to cognitive skill acquisition.)

To develop AI systems that can guide and schedule practice for optimal learning, one must be fundamentally concerned with individual variability. Given a limited amount of time for study, the system may have to choose whether to practice item A or item B, which in turn depends on specific predictions for the individual items and a given student. These specific predictions are facilitated by estimates of a student's ability or an item's intrinsic difficulty, but another key source of information for modeling memory is feedback from the student. If a student is asked to recall an item, feedback from the student—success or failure on the task—can provide critical information to adapt predictions in the future.

The field hasn't had a great deal of clarity with regard to what performance or retrieval success implies. In Bayesian Knowledge Tracing (Corbett, 1995), the feedback influences the likelihood that the student has already reached mastery—an *inferential* effect on knowledge state. In human memory models, recall success boosts memory strength more than recall failure (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009; Raaijmakers, 2003), or recall success slows subsequent memory decay more than recall failure (Pavlik & Anderson, 2005; Mooney, Sun, & Bomgardner, 2017)—both are *causal* effects on knowledge state. This document explores the implications of the fact that feedback affects knowledge state *both* inferentially and causally.

A memory model that disentangles the inferential and causal effects of feedback

We propose a model of memory that characterizes the time course of learning and forgetting. We call the model CIMM for *Causal and Inferential Memory Modulations*. CIMM is informed by observations of student performance via testing.

The scenario is that a student practices an item i over a series of trials $k \in \{1, 2, ..., K\}$. On trial k, occurring at time t^i_k, the student may fail or succeed, i.e., respond incorrectly or correctly, as denoted by $c^i_k \in \{0, 1\}$. The data set to be modeled thus consists of sequences of tuples,

$$\mathcal{D}^i_K = \{(t^i_1, c^i_1), (t^i_2, c^i_2), \ldots (t^i_K, c^i_K)\},$$

ordered such that $t^i_k \leq t^i_{k+1}$. The goal is to use CIMM to predict memory state at some future time t^i_{K+1}, i.e., $Pr(c^i_{K+1}|\mathcal{D}^i_K, t^i_{K+1})$. CIMM is intended to encompass both fact retrieval and cognitive skill application. We use the term "recall" as a shorthand to describe either. We make the simplifying assumption that items do not interact with one another. Because of this independence, we can simplify notation and drop the explicit index of item i.

CIMM treats the time-varying memory strength, $s \in [0, 1]$, as a random variable. The strength predicts the probability of success,

$$\hat{c} \sim \text{Bernoulli}(s). \tag{1}$$

As a simple instantiation of CIMM, suppose that memory strength decays exponentially in time such that the strength at the start of trial k, denoted s_k, can be expressed in terms of the strength immediately following the previous trial, $s_{(k-1)^+}$:

$$s_k = e^{-\lambda_{k-1}(t_k - t_{k-1})} s_{(k-1)^+}, \tag{2}$$

where $\lambda_{k-1} > 0$ is a random variable indicating the memory decay rate between trials $k-1$ and k. The rate is constant from one trial to the next, but may change as a consequence of experience, in the same spirit as the models of Pavlik and Anderson (2005, 2008) and Mooney et al (2017). The initial memory decay rate and strength are unknown with priors

$$\lambda_0 \sim \text{Gamma}(\mu, \nu) \quad \text{and} \quad s_0 \sim \text{Beta}(\alpha, \beta).^{1} \tag{3}$$

Adapting a proposal of the RPL model (Mooney et al., 2017), CIMM posits that the item's post-trial decay rate is deterministically modulated both by the current strength and recall success:

$$\lambda_k = \lambda_{k-1} \left[1 + \epsilon c_k (1 - s_k)^{\gamma}\right]^{-1}, \tag{4}$$

where ϵ and γ are non-negative free parameters of the model.[2] This rule embodies two mechanistic claims about memory. First, recall success can have causal consequences, consistent with the testing effect. Second, decay slows more following a successful spaced trial (long inter-trial lag) than a massed trial (short inter-trial lag), consistent with the distributed-practice effect. In Equation 4, as the inter-trial lag $t_k - t_{k-1}$ increases, s_{k+} will decrease monotonically, leading to a larger modulation of λ_k via $(1 - s_k)^{\gamma}$.

If corrective feedback is provided on each study trial, the studied item should have full strength—at least momentarily—following the trial. Thus, CIMM assumes that

$$s_{k^+} = 1. \tag{5}$$

The graphical model that represents CIMM dynamics is presented in Figure 1. The red arrows indicate the inferential process in which an observation (success or failure at recall) updates beliefs about memory decay, and the green arrows indicate the causal process by which an observation influences subsequent decay.

Comparison of CIMM and RPL

CIMM's mechanism for updating memory decay is simpler than the mechanism in RPL (Mooney et al., 2017). The simplification arises because CIMM teases apart causal and inferential effects of feedback. RPL proposes that the decay rate can *rise* (i.e., memory decays faster) following a response failure when s_k is large. One might view this rule as approximately capturing the inference that the decay-rate estimate must be too low if the model overpredicts performance. We eliminate this rule of RPL in CIMM because the inference is already a consequence of the Bayesian update $Pr(\lambda_{k-1}|c_k)$.

[1] These parameters are item specific.

[2] These parameters are potentially student specific.

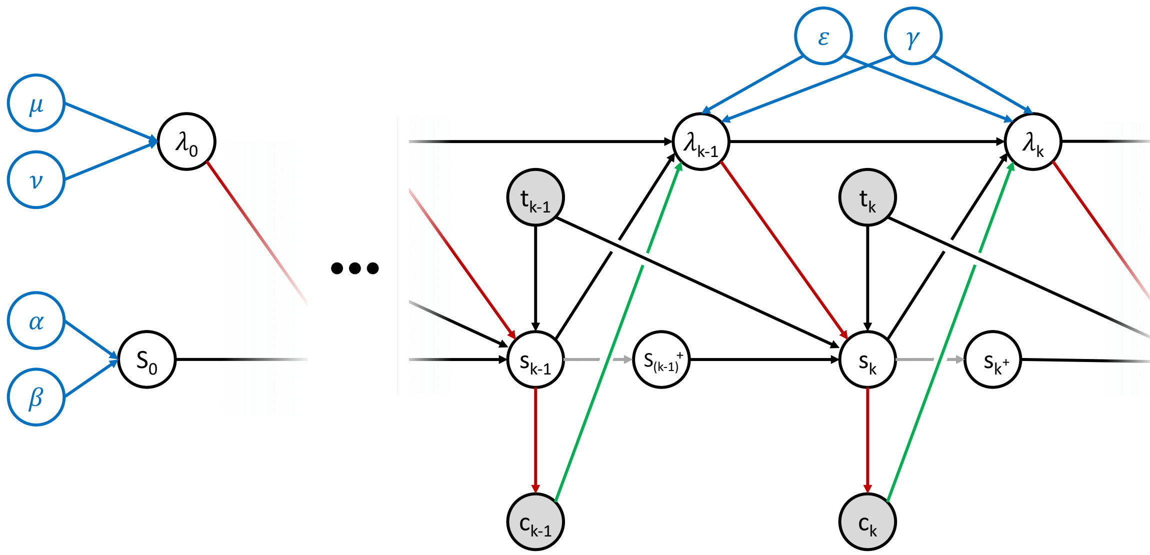

Figure 1: Dynamics of CIMM. Red arrows indicate inferential processes from observations to decay rates, and the green arrow indicates a causal process from observations to subsequent decay rates. The blue nodes are model parameters. The link from s_k to s_{k+} is greyed out because the basic version of the model in the text does not carry activation forward from one trial to the next (Equation 5).

RPL posits that forgetting follows a power law, CIMM posits an exponential. The advantage of the exponential is that it has only one free parameter. The apparent advantage of the power law is that it is consistent with current characterizations of forgetting by a population of students or by a single student for a population of items (Wixted & Carpenter, 2007). However, CIMM does predict power-law forgetting for populations based on the fact that each student-item can have its own exponential decay rate, and a mixture of exponentials well approximates a power-law function (Mozer et al., 2009).

Extensions to CIMM

Like RPL, CIMM assumes that memory goes to full strength immediately after a trial with corrective feedback. This assumption is valid only if the student paid attention to the feedback, and a more complete version of either CIMM or RPL might include a boolean variable that indicates whether the feedback was absorbed. Students will differ in how reliably they absorb the feedback.

Unlike RPL, CIMM does not include a correction for guessing. CIMM can incorporate this correction in domains where chance accuracy is significantly nonzero (e.g., where the response options are limited).

CIMM is deterministic except for the uncertainty in λ_0 and s_0. And the uncertainty in s_0 probably does not matter much. MLE could be used to determine the parameters, $\{\alpha, \beta, \mu, \nu, \epsilon, \gamma\}$. It would be advantageous to treat these parameters as random variables, leading a hierarchical model with student-item specific λ_0 and s_0, and student-specific ϵ_* and γ_* (see Figure 2.

References

Corbett, J., A. & Anderson. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, *4*, 253-278.

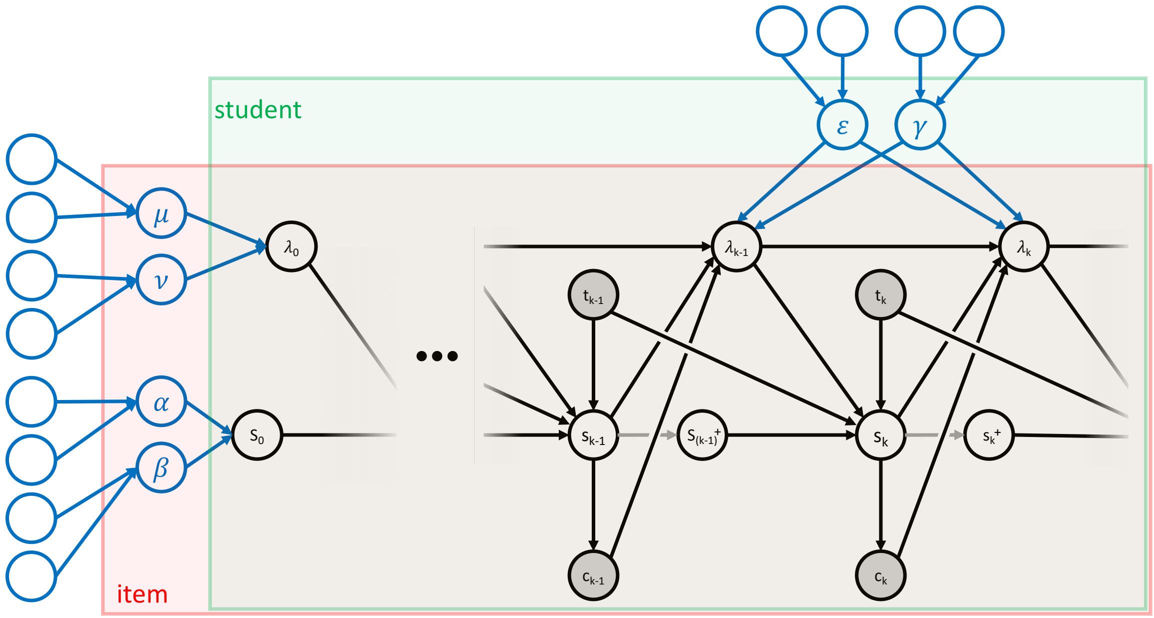

Figure 2: A view of CIMM across students and items. Using hierarchical Bayesian inference, it should be possible to model the dynamics of forgetting for individual items as well as the dynamics of learning for individual students.

Mooney, S., Sun, K., & Bomgardner, E. (2017). Optimizing study: Adaptively predicting recall probability of facts. In *Nips 2017 education workshop.* Long Beach, CA.

Mozer, M., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 22, p. 1321-1329).

Pavlik, P. I., & Anderson, J. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, *14*, 101-117.

Pavlik, P. I., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559-586.

Raaijmakers, J. G. W. (2003). Spacing and repetition effects in human memory: application of the SAM model. *Cognitive Science*, *27*, 431–452.

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science*, *18*, 133–134.