
A Queueing-Theoretic Foundation for Optimal Spaced Repetition

Siddharth Reddy

Department of Computer Science, Cornell University, Ithaca, NY 14850

SGR45@CORNELL.EDU

Igor Labutov

Department of Electrical and Computer Engineering, Cornell University, Ithaca, NY 14850

IIL4@CORNELL.EDU

Siddhartha Banerjee

School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850

SBANERJEE@CORNELL.EDU

Thorsten Joachims

Department of Computer Science, Cornell University, Ithaca, NY 14850

TJ@CS.CORNELL.EDU

1. Extended Abstract

In the study of human learning, there is broad evidence that our ability to retain a piece of information improves with repeated exposure, and that it decays with delay since the last exposure. This plays a crucial role in the design of educational software, leading to a trade-off between teaching new material and reviewing what has already been taught. A common way to balance this trade-off is *spaced repetition*, which uses periodic review of content to improve long-term retention. Though spaced repetition is widely used in practice, e.g., in electronic flashcard software, there is little formal understanding of the design of these systems. Our paper addresses this gap in three ways. First, we mine log data from spaced repetition software to establish the functional dependence of retention on reinforcement and delay. Second, we use this memory model to develop a stochastic model for spaced repetition systems. In particular, we propose a queueing-network model of the *Leitner system* for reviewing flashcards, and empirically test the validity of our queueing model via a Mechanical Turk experiment. Our experiments verify a key qualitative prediction of our model: the existence of a sharp phase transition in learning outcomes upon increasing the rate of new item introductions. Finally, we propose a heuristic approximation for our queueing model that provides a computationally tractable means of optimizing the design of the Leitner system. We discuss various aspects of the resulting optimal design, and also outline ways of extending our model to study other aspects of designing spaced repetition systems.

1.1. Introduction

The ability to learn and retain a large number of new pieces of information is an essential component of human learning. Scientific theories of human memory, going all the

way back to 1885 and the pioneering work of Ebbinghaus (Ebbinghaus, 1913), identify two critical variables that determine the probability of recalling an item: *reinforcement*, i.e., repeated exposure to the item, and *delay*, i.e., time since the item was last reviewed. Accordingly, scientists have long been proponents of the *spacing effect* for learning: the phenomenon in which periodic, spaced review of content improves long-term retention.

A significant development in recent years has been a growing body of work that attempts to ‘engineer’ the process of human learning, creating tools that enhance the learning process by building on the scientific understanding of human memory. These educational devices usually take the form of ‘flashcards’ – small pieces of information content which are repeatedly presented to the learner on a schedule determined by a *spaced repetition* algorithm (Gwern, 2016). Though flashcards have existed for a while in physical form, a new generation of spaced repetition software such as SuperMemo (Wozniak & Gorzelanczyk, 1994), Anki (Elmes, 2015), Mnemosyne (Bienstman, 2006), Pimsleur (Pimsleur, 1967), and Duolingo (duo, 2011) allow a much greater degree of control and monitoring of the review process. These software applications are growing in popularity (Gwern, 2016), but there is a lack of formal mathematical models for reasoning about and optimizing such systems. In this work, we combine memory models from psychology with ideas from queueing theory to develop such a mathematical model for these systems. In particular, we focus on one of the simplest and oldest spaced repetition methods: the *Leitner system* (Leitner, 1974).

The Leitner system, first introduced in 1970, is a heuristic for prioritizing items for review. It is based on a series of decks of flashcards. After the user sees a new item for the first time, it enters the system at deck 1. The items at each

deck form a first-in-first-out (FIFO) queue, and when the user requests an item to review, the system chooses a deck i according to some schedule, and presents the top item. If the user does not recall the item, the item is added to the bottom of deck $i - 1$; else, it is added to the bottom of deck $i + 1$. The aim of the scheduler is to ensure that items from lower decks are reviewed more often than those from higher decks, so the user spends more time working on forgotten items and less time on recalled items. Existing schemes for assigning review frequencies to different decks are based on heuristics that are not founded on any formal reasoning, and hence, have no optimality guarantees. One of our main contributions is a principled method for determining appropriate deck review frequencies.

The problem of deciding how frequently to review different decks in the Leitner system is a specific instance of the more general problem of *review scheduling* for spaced repetition software. The main tension in all settings is that schedules must balance competing priorities of introducing new items and reviewing old items in order to maximize the rate of learning. While most existing systems use heuristics to make this trade-off, our work presents a principled understanding of the tension between novelty and reinforcement.

1.2. Related Work

The scientific literature on memory models is both old (dating back to more than a century) and also highly active. The basic memory model, the so-called *exponential forgetting curve*, was first studied by Ebbinghaus in 1885 (Ebbinghaus, 1913) – it models the probability of recalling an item as an exponentially-decaying function of the time elapsed since previous review and the memory ‘strength’. The exact nature of how strength evolves as a function of the number of reviews, length of review intervals, etc. is a topic of debate, though there is some consensus for the existence of a *spacing effect*, in which spaced reviews lead to greater strength than massed reviews (i.e., cramming) (Dempster, 1989; Cepeda et al., 2006). Recent studies have proposed more sophisticated probabilistic models of learning and forgetting (Pashler et al., 2009; Lindsey et al., 2014), and there is a large body of related work on item response theory and knowledge tracing (Linden & Hambleton, 1997; Corbett & Anderson, 1994). Our work both contributes to this literature (via observational studies on log data from the Mnemosyne software system) and uses it as the basis for our queuing model and scheduling algorithm.

Though used extensively in practice (cf (Gwern, 2016) for an excellent overview), there is very limited literature on the design of spaced repetition software. One notable work in this regard is that of Novikoff et al. (Novikoff et al., 2012), who propose a theoretical framework for spaced

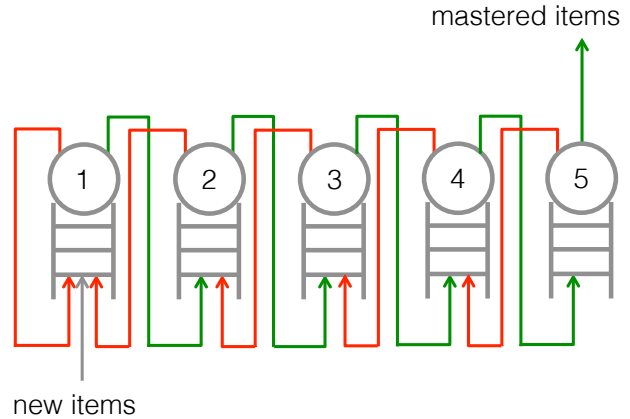


Figure 1. The Leitner Queue Network: Each queue represents a deck in the Leitner system. New items enter the network at deck 1. Green arrows indicate transitions that occur when an item is correctly recalled during review, and red arrows indicate transitions for incorrectly recalled items. Queue i is served (i.e., chosen for review) at a rate μ_i , and selects items for review in a FIFO manner.

repetition, based on a set of deterministic operations on an infinite string of content pieces. In particular, they assume identical items, and design schedules to implement deterministic spacing constraints, which are fixed in advance, based on an intuitive understanding of the effect of memory models on different learning objectives. The focus in (Novikoff et al., 2012) is on characterizing the combinatorial properties, in particular, the maximum asymptotic throughput, of schedules that implement various spacing constraints. Though our work shares the same spirit of formalizing the spaced repetition problem, we improve upon their work in three ways: (1) in terms of empirical verification, as our work leverages both software log-data and large-scale experimentation to verify the memory models we use, and test the predictions made by our mathematical models, (2) in computational terms, wherein, by using appropriate stochastic models and approximations, we formulate optimization problems that are much easier to solve, and (3) in terms of flexibility, as our model can more easily incorporate various parameters such as the user’s review frequency, non-identical item difficulties, different memory models, etc.

1.3. Contributions and Argument Outline

Our key contributions fall into two categories. First, we introduce a principled methodology for designing review scheduling systems with various learning objectives. Second, the models we develop provide qualitative insights and general principles for spaced repetition learning. Our overall argument consists of the following three steps:

1. *Mining large-scale log data to identify human memory models:* First, we perform observational studies on data from Mnemosyne (Bienstman, 2006), a popular flashcard software, to compare different models of retention probability of items as a function of reinforcement and delay. Our results add to the existing literature on memory models, and also provide the empirical foundation upon which we base our mathematical models.
2. *Mathematical modeling of spaced-repetition systems:* Our main contribution lies in embedding the above memory model into a stochastic model for spaced-repetition learning systems, and using this model to optimize the review schedule. Our framework, which we refer to as the *Leitner Queue Network*, is based on ideas from queueing theory and job scheduling. Though conceptually simple and easy to simulate, the Leitner Queue Network does not however provide a tractable way to optimize the review schedule. To this end, we propose a (heuristic) approximate model, which in simulations is close to our original model for low arrival rates, and which leverages the theory of product-form networks (Kelly, 2011) to greatly simplify the scheduling problem. This allows us to quantitatively study the structure of optimal scheduling policies.
3. *Verification of mathematical model in controlled experiments:* Finally, we use Amazon Mechanical Turk (mtu, 2005) to perform large-scale experiments to test our mathematical models. In particular, we verify a critical qualitative prediction of our mathematical model: the existence of a *phase transition* in learning outcomes upon increasing the rate of introduction of new content beyond a maximum threshold. Our experimental results agree well with our model’s predictions, thereby reaffirming the utility of our framework.

Our work thus provides the first mathematical model for spaced repetition systems which is empirically tested and which admits a tractable means for optimizing such systems. In the process, it opens several directions for further research: developing better models for such systems, providing better analysis for the resulting models, and performing more empirical studies to understand these systems. Moreover, our experimental platform can help serve as a testbed for future studies, and to this end, we release all data and software tools to replicate our experiments and facilitate follow-up studies.

1.4. Key Results

The Leitner Queue Network model predicts a sharp decrease in the rate of learning when the rate of new item introductions exceeds a certain threshold. In a controlled experiment on Mechanical Turk, we expose students to dif-

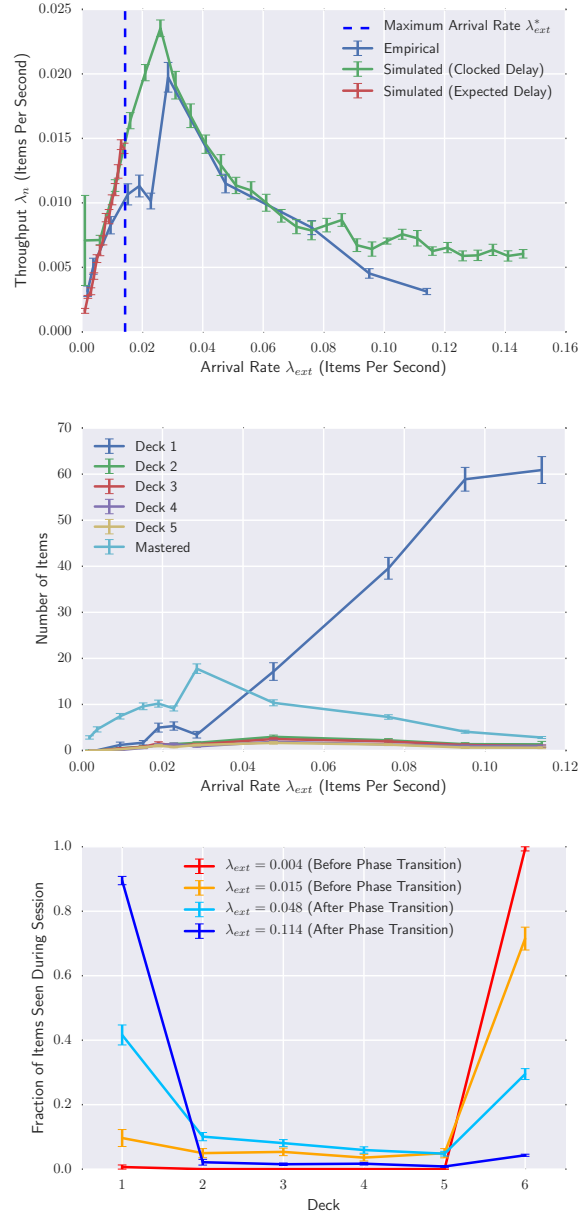


Figure 2. Results of an experiment on Mechanical Turk, where over 300 students spent 15 minutes reviewing flashcards for English-Japanese and English-ASL vocabulary. Each student was randomly assigned to a fixed arrival rate of new items. Plot (a) compares the empirical throughput observed in the Mechanical Turk experiment with simulated throughput and the theoretical phase transition threshold – the empirical data agree well with the simulations. Plots (b-c) examine the final state of each of the Leitner decks in the Mechanical Turk experiment – items tend to pile up in deck 1 when the system is destabilized by a high arrival rate.

ferent arrival rates of new items and observe a phase transition near the threshold predicted by our model. See Figure 2 for details.

1.5. Conclusion

Our work develops the first formal mathematical model for reasoning about spaced repetition systems – in particular, the *Leitner system* – that is tractable to optimize and validated by empirical data. Our formalization suggests the maximum speed of learning as a natural design metric; using techniques from queueing theory, we derive a tractable program for calibrating the Leitner system to optimize the speed of learning. Additionally, the queueing framework opens doors to leveraging an extensive body of work in this area to develop more sophisticated extensions. To inspire and facilitate further research in this direction, we release (1) all model and evaluation code, (2) framework code for carrying out user studies, and (3) the data collected in our Mechanical Turk study. For a complete description of our methodology and results, see our full paper (Reddy et al., 2016). Code and data for replicating our experiments are available online at <http://siddharth.io/leitnerq>.

1.6. Future Work

- The primary direction of future work is to obtain a better understanding of the Leitner Queue Network, and better approximations with rigorous performance guarantees. Doing so will allow us to design better control policies, which, ideally, can optimize the average learning rate in the transient regime. The latter is critical for designing policies for *cramming* (Novikoff et al., 2012), a complementary problem to long-term learning where the number of items to be learnt is of the same order as the number of reviews. We have some ongoing work that focuses on this setting.
- Our approximation of the Leitner Queue Network (which provides a tractable review schedule optimization problem) is only valid at low arrival rates and for moderately-sized review sessions. If a delay-driven Leitner Queue Network is simulated for a longer period, then it exhibits instability due to the presence of a positive feedback loop where a burst of new arrivals leads to large delays which leads to more items re-entering the queue. We are currently investigating ways to control this in theory, and also trying to replicate it via experiments.

1.7. Acknowledgements

We thank Peter Bienstman and the Mnemosyne project for making their data set publicly available. This research was funded in part through NSF Awards IIS-1247637, IIS-1217686, IIS-1513692, the Cornell Presidential Research Scholars Program, and a grant from the John Templeton Foundation provided through the Metaknowledge Network.

References

- Amazon mechanical turk. <https://www.mturk.com>, 2005.
- Duolingo. <https://www.duolingo.com>, 2011.
- Bienstman, Peter. The mnemosyne project. <http://mnemosyne-proj.org>, 2006.
- Cepeda, Nicholas J, Pashler, Harold, Vul, Edward, Wixted, John T, and Rohrer, Doug. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3):354, 2006.
- Corbett, Albert T and Anderson, John R. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4):253–278, 1994.
- Dempster, Frank N. Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1(4):309–330, 1989.
- Ebbinghaus, Hermann. *Memory: A contribution to experimental psychology*. Number 3. University Microfilms, 1913.
- Elmes, Damien. Anki. <http://ankisrs.net>, 2015.
- Gwern, Patrick. Spaced repetition. <http://www.gwern.net/Spaced%20repetition>, 2016.
- Kelly, Frank P. *Reversibility and stochastic networks*. Cambridge University Press, 2011.
- Leitner, Sebastian. *So lernt man lernen*. Herder, 1974.
- Linden, WJ and Hambleton, Ronald K. Handbook of modern item response theory. *New York*, 1997.
- Lindsey, Robert V, Shroyer, Jeffery D, Pashler, Harold, and Mozer, Michael C. Improving students’ long-term knowledge retention through personalized review. *Psychological science*, 25(3):639–647, 2014.
- Novikoff, Timothy P, Kleinberg, Jon M, and Strogatz, Steven H. Education of a model student. *Proceedings of the National Academy of Sciences*, 109(6):1868–1873, 2012.
- Pashler, Harold, Cepeda, Nicholas, Lindsey, Robert V, Vul, Ed, and Mozer, Michael C. Predicting the optimal spacing of study: A multiscale context model of memory. In *Advances in neural information processing systems*, pp. 1321–1329, 2009.
- Pimsleur, Paul. A memory schedule. *Modern Language Journal*, pp. 73–75, 1967.

Reddy, Siddharth, Labutov, Igor, Banerjee, Siddhartha, and Joachims, Thorsten. Unbounded human learning: Optimal scheduling for spaced repetition. *arXiv*, 2016. URL <http://arxiv.org/abs/1602.07032>.

Wozniak, PA and Gorzelanczyk, Edward J. Optimization of repetition spacing in the practice of learning. *Acta neurobiologiae experimentalis*, 54:59–59, 1994.