Degree Project : Specification and schedule

Sugandh Sinha, sugandh@kth.se

Preliminary title: Adaptive personalization of e-learning curricula using deep reinforcement learning

CSC Supervisor: Pawel Herman (pherman@kth.se)

Principal: Sana Labs

Supervisor at principal's workplace: Anton Osika (anton@sanalabs.com)

1 Background & Objective

Research since the 19^{th} century has shown that spacing repetition of study material with delays between reviews has a positive impact on the duration over which the material can be recalled [1, 4]. Different ways for performing this spacing have been argued for and investigated - the preferable method for deciding on the spacing scheme is however to empirically learn the optimal spacing scheme.

This could be done in an e-learning environment where an agent/model is in charge of scheduling and presenting learning materials to students. Since every student might possess a different learning capability, the agent/model should preferably be able to infer the learning pattern for each student to be able to present learning material effectively. The term effectively is used to denote how many times and when the material should be presented again, as spaced practice has shown to reduce this number.

Most of the previous work in the knowledge tracing domain has been about predicting whether the question that the agent/model is going to present will be answered correctly or not by the student [6] but without paying much heed to the importance of the order of the questions. In this work we will examine whether presenting the questions in a specific sequence different from random ordering can enhance the effectiveness of students' learning.

As an example, one of the most commonly used methods in flashcards for spaced repetition is Leitner System [5]. The basic idea behind this technique is to make users interact more with items that they are likely to forget and let them spend less time on items that they are able to recall efficiently. This system maintains a set of n decks. When the user sees an item for the first time then that item is placed in deck 1. Afterwards, when the user sees an item placed in deck i and recalls it correctly, the item is moved to the bottom of deck i+1. However, if the user answers incorrectly then it is moved to the bottom of deck i-1. Leitner systems make users spend more time on lower decks so that they can recall items that they are frequently forgetting. Reddy et al. [9] use a Queue based Leitner system to generate a mathematical model for spaced repetition system.

This thesis work finds it inspirations and has its applications in the areas of machine learning, e-learning, educational data mining and online education.

The thesis work will be done at Sana Labs, Stockholm. This work is relevant for Sana Labs as the company uses machine learning to predict student user/student behaviour for optimizing learning experiences. Literature on the subject and present work at Sana Labs, typically involves manually deciding criteria for what content to show, when and to which users. This work will explore opportunities for replacing the manually selected criteria with an end-to-end recommendation system that learns the criteria by itself.

The goal of this project work is to develop a proof-of-concept for a reinforcement learning-based approach to the problem of spacing the review of learning material in a student simulator with emphasis on data efficiency of the learning algorithm.

2 Research Question & Method

2.1 Research question

In previous work Reddy et al. [8] investigated a model-free review scheduling algorithm for spaced repetition systems, which learns its policy using the observations made on students study history without actually learning a student model. This thesis work will be focused on finding out how data efficient a reinforcement learning algorithm can be when compared to previously published heuristics like [5] when it comes to making spaced repetition and making students learn. Being data efficient, here, refers to learning good policies with less student interaction data. The work done by Reddy et al. [8] relies on a number of iterations to compare performance but this thesis work will be concerned with using number of questions as a measure for performance comparison.

The research questions that the project seeks to answer are

- *How does reinforcement learning perform when we replace one of the reward functions which maximizes the likelihood of recalling all items used by Reddy et al. [8] to a realistically observable reward such as samples of exercise outcome where 0 is incorrect and 1 is correct?*
- *How does reinforcement learning perform when we replace the same reward function with an RNN model that predicts the reward?*

Key Challenges in the project are:

- It is really hard to model a human learning model that has all the characteristics of a typical student and as such is well beyond our scope.
- reinforcement learning needs lots of data.
- Formulation of a good reward function for an agent.
- Deciding on how to induce variance in the simulated student population.

2.2 Method

We will have three simulators that will act as students/human learning models namely half-life regression, exponential forgetting curve and generalized power law, each having its own learning and retaining characteristics.

Exponential forgetting curve:
Ebbinghaus's [1]classic study on forgetting learned materials states that when a person learns something new, most of it is forgotten in an exponential rate within the first couple of days and after that the rate of loss gradually becomes weaker.

Reddy et al. [9] give the probability of recalling an item as

$$P[recall] = exp(-\theta \cdot d/s), \tag{1}$$

where θ is the item difficulty, d is the time elapsed since the material was last reviewed and s is the memory strength.

Half life regression:
As described by Settles and Meeder [7], the memory decays exponentially over time:

$$p = 2^{\Delta/h}, \tag{2}$$

In this equation, p denotes the probability of correctly recalling an item (e.g., a word), which is a function of Δ, the lag time since the item was last practiced, and h, the half-life or measure of strength in the learners long-term memory.
When $\Delta =$ h, the lag time is equal to the half-life, so $p = 2^{-1} = 0.5$, and the student is on the verge of being unable to remember.
Assuming that the half-life should increase exponentially with each repeated exposure. The estimated half life $\hat{h}_\Theta$ is given by

$$\hat{h}_\Theta = 2^{\Theta x}, \tag{3}$$

where x is a feature vector that describes the study history for the student-item pair and the vector Θ contain weights that correspond to each feature variable in x.

Generalized power law:
Wixted and Carpenter [10] state that the probability of recalling decays according to a generalized power law as a function of t.

$$P[recall] = \lambda(1 + \beta.t)^{-\Psi}, \tag{4}$$

where t is the retention interval, λ is a constant representing the degree of initial learning, β is a scaling factor on time $(h > 0)$ and Ψ represents the rate of forgetting.

From each of these student simulators, we will generate interaction histories for students that will act as our data. The initial step in this thesis work would be to set up the student simulators and potentially adding logic to these simulators so that they resemble real students more closely. For example, parameters representing particular attributes of each student, such as memory strength and learning rate, should drawn from a distribution representing the population of the different students that exist.

We will be using Trust Regional Policy Optimization (TRPO) reinforcement learning algorithm that will interact with the student data and update its policy and reward function. TRPO has been shown to be robust and works well with domains having high dimensional inputs [2, 3].

In Reddy et al.[8] experiments, the reward for updating the agents policy is computed by directly accessing the recall likelihood specified by the student model, which is not possible in real situations. Instead of using recall likelihoods for reward, we will investigate how the results vary when we use the sum of observed correct outcomes as the reward.

2.3 Expected Scientific Results

Our hypothesis is that the reinforcement learning agent should be able to give a performance which is, if not better, comparable to heuristics like Leitner, Supermemo, etc.
This hypothesis will be tested on the three student simulators mentioned above in the Method subsection.

3 Evaluation & News Value

3.1 Evaluation

The implementation of the system will be tested in a simulated environment against multiple benchmark(s) for ordering the review questions.
The trained model will be tested on student simulators and will be evaluated using student recall rate based on how many of the questions presented were correctly/incorrectly answered and also how the student recall rate changes over time as the reinforcement learning algorithm learns.

3.2 The work's innovation/news value

The innovative part of the thesis work lies in finding out the data efficiency of the reinforcement learning when applied to the problem of human learning and spaced repetition, which to the best of our knowledge has not been done before. Data efficiency, here, refers to learning good policies with less student interaction data.

4 Pre-Study

The literature study would be mostly focused on two areas -

- Reinforcement learning
- Human Learning, Spaced Repetition and Knowledge tracing

The literature mentioned below is not an exhaustive list but this is where I will get started with the pre-study.

4.1 Reinforcement Learning

- Andrew G. Barto, Richard S. Sutton. Reinforcement Learning : An Introduction. The MIT Press, 2014.
- Siddharth Reddy, Sergey Levine, Anca Dragan. Accelerating Human Learning with Deep Reinforcement Learning. In Conference on Neural Information Processing Systems, 2017.
- Naoki Abe, Edwin Pednault, Haixun Wang, Bianca Zadrozny, Wei Fan, Chid Apte. Empirical Comparison of Various Reinforcement Learning Strategies for Sequential Targeted Marketing. In IEEE Int. Conf. Data Mining 310 (2002).
- Matthew Hausknecht, Peter Stone. Deep Recurrent Q-Learning for Partially Observable MDPs. In Association for the Advancement of Artificial Intelligence, 2015.
- Silver, D. et al. Concurrent Reinforcement Learning from Customer Interactions. In Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. PMLR volume 28.
- Xiujun Li, Lihong Li, Jianfeng Gao, Xiaodong He, Jianshu Chen, Li Deng, and Ji He. Recurrent Reinforcement Learning: A Hybrid Approach. arXiv:1509.03044, 2015.
- R. McAllister and C. Rasmussen. Data-efficient reinforcement learning in continuous-state POMDPs. arXiv preprint arXiv:1602.02523, 2016.

- Burr Settles, Brendan Meeder. A Trainable Spaced Repetition Model for Language Learning. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 18481858, Berlin, Germany, August 7-12, 2016. Association for Computational Linguistics.

4.2 Human Learning, Spaced Repetition and Knowledge Tracing

- Spaced repetition. http://www.gwern.net/Spaced

- Sean H. K. Kang. Spaced Repetition Promotes Efficient and Effective Learning: Policy Implications for Instruction. In Policy Insights from the Behavioral and Brain Sciences 2016, Vol. 3(1) 1219.

- Piech, C. et al. Deep Knowledge Tracing. In Advances in Neural Information Processing Systems, pages 505513, 2015.

- Michael C Mozer and Robert V Lindsey. Predicting and improving memory retention: Psychological theory matters in the big data era, 2016.

- Christopher James Piech. Uncovering Patterns in Student Work: Machine Learning to Understand Human Learning. PhD thesis, Stanford University, 2016.

- S. Reddy, I. Labutov, S. Banerjee, and T. Joachims. Unbounded human learning: Optimal scheduling for spaced repetition. In ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2016.

- PA Wozniak and Edward J Gorzelanczyk. Optimization of repetition spacing in the practice of learning. Acta neurobiologiae experimentalis, 54:5959, 1994.

- Siddharth Reddy, Igor Labutov, Thorsten Joachims, Learning Student and Content Embeddings for Personalized Lesson Sequence Recommendation, Work in Progress, ACM Conference on Learning at Scale (L@S), 2016.

- Siddharth Reddy, Igor Labutov, Thorsten Joachims, Learning Representations of Student Knowledge and Educational Content, ICML Workshop on Machine Learning for Education, 2015.

5 Conditions & Schedule

5.1 Limitations

For the thesis project, we do not plan on testing our reinforcement learning agent in real live environment with students as we expect that it will take a lot of time. Also, the student simulators do not model all the characteristics of a typical students. The parameters of the student simulators will not be derived from the distribution of real student data but will just be set to reasonable values.

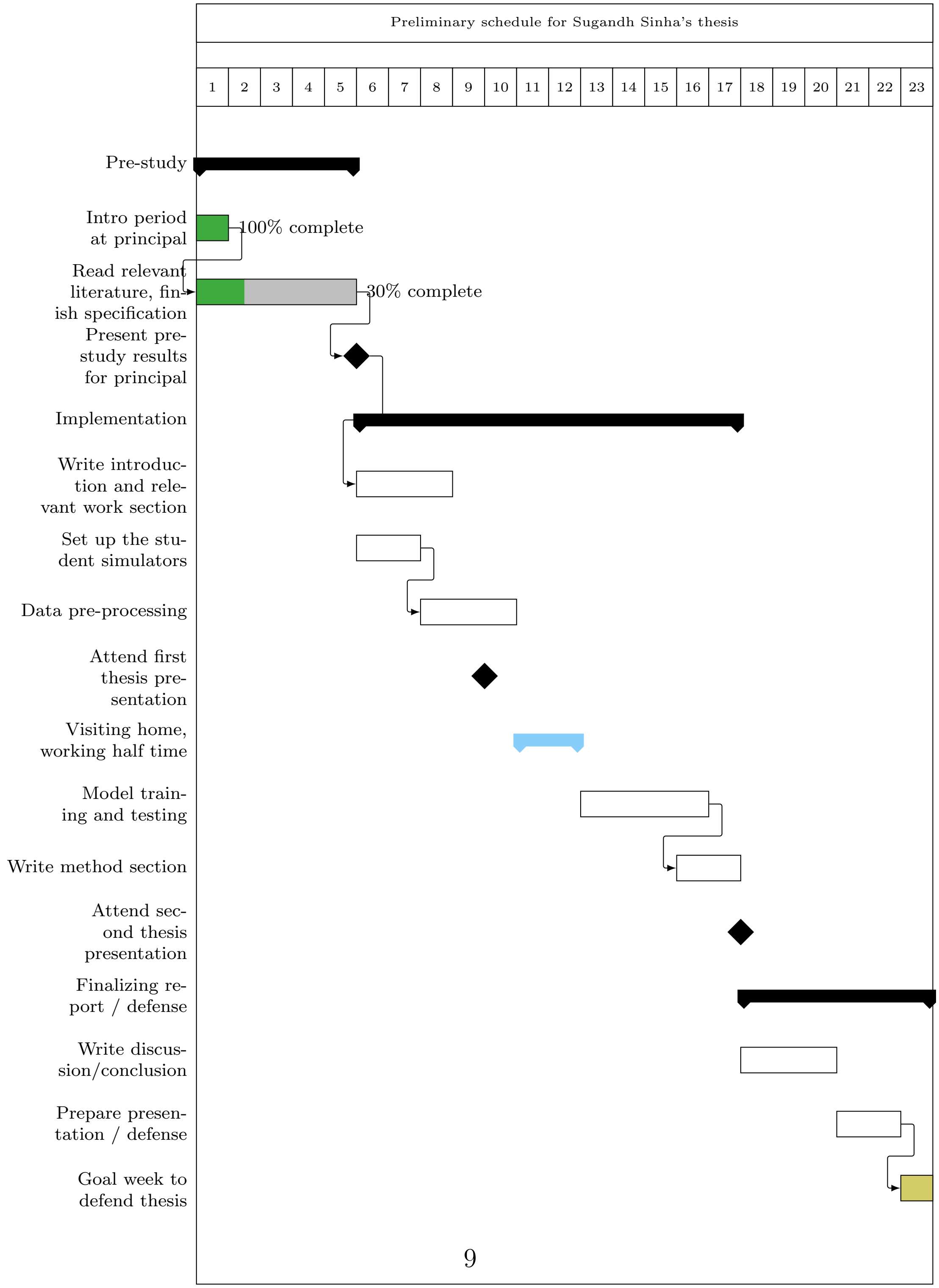
5.2 Collaboration with the principal (external supervisor)

The thesis work will be done at Sana Labs, Stockholm and the supervisor at Sana Labs would be involved in all the steps of the thesis work.

5.3 Resources

- Hardware, my laptop and a computer equipped with graphics card.
- Toolkits for reinforcement learning like OpenAI Gym etc., deep learning frameworks like tensorflow, keras, theano, etc.
- Simulated data of student interaction histories.

Schedule


Preliminary schedule for Sugandh Sinha's thesis
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
Pre-study
Intro period at principal
100% complete
Read relevant literature, finish specification
30% complete
Present pre-study results for principal
Implementation
Write introduction and relevant work section
Set up the student simulators
Data pre-processing
Attend first thesis presentation
Visiting home, working half time
Model training and testing
Write method section
Attend second thesis presentation
Finalizing report / defense
Write discussion/conclusion
Prepare presentation / defense
Goal week to defend thesis

References

[1] Hermann Ebbinghaus. *Memory: A contribution to experimental psychology.* Teachers College, Columbia University., 1885. Translated by Henry A. Ruger & Clara E. Bussenius (1913).

[2] Pieter Abbeel Michael Jordan John Schulman, Sergey Levine and Philipp Moritz. Trust region policy optimization. *ICML*, 2015.

[3] Miles Brundage Kai Arulkumaran, Marc Peter Deisenroth and Anil Anthony Bharath. A brief survey of deep reinforcement learning. *IEEE Signal Processing Magazine, Special Issue on Deep Learning for Image Understanding*, 34(6), 2017. arXiv:1406.2040 [quant-ph].

[4] Sean H. K. Kang. Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences*, 3(1) 1219, 2016.

[5] Sebastian Leitner. *So lernt man lernen.* Herder, 1974.

[6] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 505–513. Curran Associates, Inc., 2015.

[7] Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learningac. *ACL(1)*, 2016.

[8] Anca Dragan Siddharth Reddy, Sergey Levine. Accelerating human learning with deep reinforcement learning. *Conference on Neural Information Processing Systems*, 2017. Workshop paper.

[9] Siddhartha Banerjee Thorsten Joachims Siddharth Reddy, Igor Labutov. Unbounded human learning: Optimal scheduling for spaced repetition. *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

[10] J. T. Wixted and S. K. Carpenter. The wickelgren power law and the ebbinghaus savings function. *Psychological Science*, 18(2):33134, feb 2007.