The Effect of Variations of Prior on Knowledge Tracing

Matti Nelimarkka
School of Information, UC Berkeley
102 South Hall
Berkeley, California 94720-4600
Helsinki Institute for Information Technology HIIT,
Aalto University
PO Box 15600
Aalto, Finland 00076
matti.nelimarkka@hiit.fi

Madeeha Ghori
Department of Electrical Engineering and
Computer Sciences, UC Berkeley
387 Soda Hall
Berkeley, California 94720-17761
madeeha.ghori@berkeley.edu

ABSTRACT
Knowledge tracing is a method which enables approximation of a student's knowledge state using a Bayesian network for approximation. As the applications of this method increase, it is vital to understand the limits of this approximation. We are interested how well knowledge tracing performs when students' prior knowledge on the topic is extremely high or low. Our results indicate that the estimates become more erroneous when prior knowledge is extremely high ($prior = 0.90$).

Keywords
bayesian knowledge tracing, personalization, prior, parameter estimation

1. INTRODUCTION
The Bayesian Knowledge-Tracing (BKT) algorithm was developed in 1995 in an effort to model students' changing knowledge state during skill acquisition [5]. The idea is to interpret students' knowledge – a hidden variable – based on observed answers to a set of questions. The algorithm tracks the change in this probability distribution over time using a simple Bayes' net. The model is often presented as four parameters: prior, learn, guess and slip (see Figure 1). *Prior* refers to the probability that the student knows the material initially, before acquiring any skills, *learn* indicates that the student did not have the skill initially but acquired it through doing the exercise, *guess* refers to accidentally answering the question correct and *slip* to answering accidentally wrong.

Knowledge tracing is the most prominent method used to model student knowledge acquisition and is used in most intelligent learning systems. These systems have been said to be outperforming humans since 2001 [3] and have been used in the real world to tutor students [4]. For these reasons it is important to fully understand the strengths and limitations of knowledge tracing before applying it more widely in the classroom. As the parameters of the model are now known, there is a need to estimate these parameters from the given data. Previous research has demonstrated that the accuracy of parameter estimation – and therefore knowledge tracing – can be improved by applying different heuristics [17, 13] or methods [16, 18] including personalizing the model for each user [20, 8] or by extending the data used for analysis [15, 6, 1].

Our work starts from a different premise: how robust is the BKT approach to variation in the parameter space? Our special interest is in the *prior* variable, which correlates to a student's knowledge of the topic before answering a question. In any classroom, MOOC or otherwise, some students will come in with a better understanding of the material than others. Therefore it is important to study the effectiveness of knowledge tracing on parameter estimation when prior is extremely high or low.

If knowledge tracing models are inaccurate in modelling students of a certain prior parameter, then smart tutors and other systems designed to help those students learn will be less effective. Especially if the students being modelled inaccurately are those students doing poorly in the class, as the smart tutors exist to help them the most.

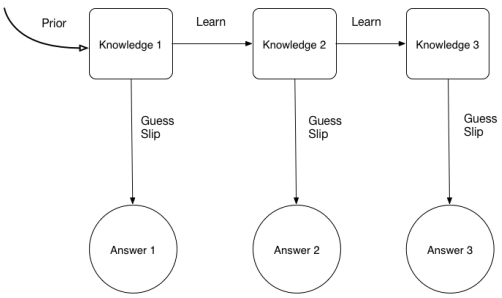

Figure 1: The model of knowledge tracing

2. PREVIOUS WORK

For the purposes of this work, here we shortly summarize three methods previously applied to improve the prediction capabilities of BKT models. However, these methods are insufficient to address the practical problem described above, resulting in a need for our own experiment.

2.1 Individualization

Yudelson et al. [20] experimented with individualization by bringing student-specific parameters into the BKT algorithm on a larger scale. They split the usual skill-specific BKT parameters into two components: one skill-specific and one student-specific. They then built several individualized BKT models and added student-specific parameters in batches, examining the effect each addition had on the model's performance. They found that student-specific prior parameters did not provide a vast improvement. However, student-specific learning provided a significant improvement to the model's prediction accuracy.

Pardos and Heffernan furthered the experiment by developing a method of formulating the individualization within the Bayes' Net framework [11]. Especially interesting in terms of our work is the difference prior values and methods suggested for this individualization. Pardos observes that models taking student spesific priors based on students' prior knowledge clearly outperform traditional knowledge trace approach. This is a contrast Yudelson et al.'s findings [20] but it still underscores the importance of individualization in the BKT algorithm.

Related to individualization per user, there have been discussion on using different values per resources. It can be argued that different exercises teach different topics [7, 14]. This can be further used to individualize the model for different topics, an approach which has gained initial support on empirical studies [14].

2.2 Enhancing the data

The second approach to improve these methods is related to enhanching the data used for prediction. In its most simple form, this can be done by adding additional relevant data, such as data from past years, to the analysis [15]. Others have explored the possibility of adding more data to the general domain-related knowledge on the models, and suggest that these indeed improve the estimates [6].

However, the current direction in enhanced data relates to information available on user interaction – especially in MOOC environments where it is possible to access this kind of data. To illustrate, Baker, Corbett, and Aleven [1] explore interactions with the learning system and other non-exercise related data, such as time spent on answering and asking help, to determine the difference between slips and guesses.

We applaud these efforts and acknowledge that data other than just student responses may indeed help to detect both the cases where initial knowledge (prior) is high and when it is low, instead of tweaking the EM algorithm further.

2.3 Improving the methods

There are several heuristics currently used to enhance the BKT algorithm. One such heuristic involves expecting the sum of slip and guess to be less than or equal to 1 [17]. Other work determined that one's starting estimated parameters could affect where the algorithm converged to. In order to improve the accuracy of the convergence, it was suggested that starting parameters be selected from a Dirichlet distrbution derived from the data set [2, 13].

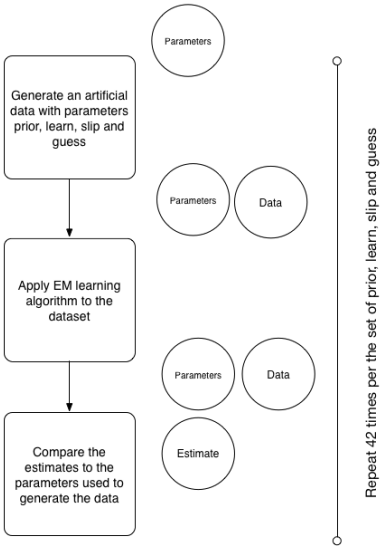

Figure 2: The approach used in this study

There have also been efforts to explore other machine learning methods on educational data. Initial trials born in the KDDCup competition use a medley of random forests and other machine learning algorithms but these methods have proven largely unsuccessful [16, 18].

The knowledge tracing community, while accepting the validity of some of these heuristics [9, 12], has criticized their inability to provide any insight into the student learning model. Individualization, however, has the potential to improve the BKT algorithm while also providing a pedagogical explanation for said improvements.

3. METHODOLOGY

We began by generating datasets with specific known initial parameters in order to simulate groups of students at different knowledge levels. We then ran expectation maximization (EM) on these datasets and allowed knowledge tracing to calculate its own estimated parameters. We then compared these estimated parameters to the original ones used for generation to determine if the accurency of the parameter estimation depends on the initial parameters.

Table 1: Ground Truth Parameter Sets

	prior	learn	guess	slip
Set 1.1 ... 1.6	0.15	0.10	0.10	0.05
Set 2.1 ... 2.6	0.30	0.10	0.10	0.05
Set 3.1 ... 3.6	0.15	0.20	0.10	0.05
	⋮			
Set 48.1 ... 48.6	0.90	0.20	0.20	0.10

3.1 Generating the Data

As our goal was to determine how the prior ground truth affects parameter estimation, we varied the prior used to synthesize the data sets. We used six different priors $(0.15, 0.30, \ldots, 0.75, 0.9)$, and two variations on learn, slip and guess[1] each (see Table 1); total of 48 variations of these parameters. Each of these data sets consists of 10,000 students and 20 observations per student. To increase the variation, we generated 6 datasets per condition. This kind of simulated approach has been previously used to evaluate the success of Bayesian machine learning methods [8].

3.2 Analysis Procedure

For each data set, we estimated the parameters using the *expectation maximization fitting* (EM) algorithm using the fastHMM implementation [10]. The parameter estimation was conducted using a grid search with ten parameters, and the best fitting model was selected using the log likelihood.

Using our 288 data sets, we can compare the estimates and ground truths for each parameter and analyze the accuracy of the estimates. We apply the standard methods of root-mean-square error (RMSE) and other visualizations to do our analysis. Using RMSE, we will be able to see if certain ground truths lend themselves to more accurate estimations.

4. RESULTS

First, let us explore the parameter estimation in detail. The avarage RMSE measurement in the data (Table 2) indicate that the prediction quality decreases as the prior increases; there is also increase of variance of the RMSE. This indicates that the predictions with higher priors are first more erronous and second, they converge in a larger area, resulting in variance. To confirm our observations, we conducted a Wilcox-Mann-Whitney test to explore if the computed RMSEs differented in statistically significant manner. As shown in Table 3, both the RMSEs computed from the data sets with priors 0.15 and 0.90 statistically differ significantly from the other datasets ($p < 0.05$). Therefore we conclude that the EM algortihm performs badly when prior is high.

To further understand this phenomena, we explore the estimates per parameter. The errors per parameter are shown in the Figure 3. The mean estimates are rather constantly close by the zero, though a higher prior does affect variance. As ground truth prior increases, the variance of guess and learn increases while the variance of prior decreases. In theory, a lesser variance on the prior prediction should imply a more accurate prior estimate. However, as we saw in Table 2, this is not actually the case. The prior estimate gets less accurate as the value of the ground truth prior increases. In Figure 3 we can see again some of the results we saw in Table 2: the prediction accuracy decreases when prior is 0.6 and continues to decrease as prior increases.

[1]Variations were 0.10 and 0.20 for learn and guess, and 0.05, 0.10 for slip.

Ground truth prior	mean RMSE	var RMSE
0.15	0.056639	0.000594
0.30	0.069073	0.001137
0.45	0.070005	0.000584
0.60	0.074044	0.001874
0.75	0.075946	0.002229
0.90	0.085257	0.004876

Table 2: The mean and variance of the root-mean-square errors per prior

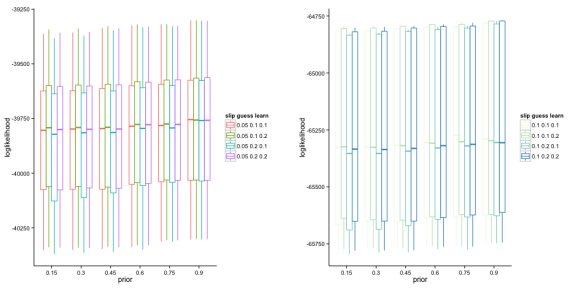

Figure 4: Log likelihoods with different parameters

Figure 4 shows that the log likelihood for each of the parameter combinations we analyzed. We see a slight, but non-significant increase in the log likelihoods, suggesting that the model is performing better – even while our RMSE error indicator demonstrates otherwise. It is also noteworthy to observe that that when slip is 0.10, all log likelihoods range between -65500 and -65250 but when slip is 0.05, all log likelihoods range between -40000 and -35750, indicating that the slip value had a dramatic effect on the model estimation accurancy.

5. IMPLICATIONS

Our findings indicate that there are higher errors in the parameter estimations when prior is high (0.90). This is probably due to the lack of evidence available for the HMM to attribute to the learn and guess parameters. One approach to examine the impact of these errors is to examine the students' subjective experience in different conditions [19]. As our data is syntetic, we can not measure the time consumed by students due to errors, as examined by Youdelson & Koedinger [19]. Instead we explore the difference on the number of questions students' need to answer to achieve mastery learning – for our purposes knowledge above 95 % and assuming that the students answer each question correctly.

Examining the case of high prior knowledge, and when the true learning was 0.1, we observed that majority of students needed to answer over 5 times to achieve mastery (or: from the 168 predicted value sets available, only 24 achieved mastery), and for the high learning (0.2) the situation was not

Table 3: Significant differences between the RMSEs

	0.15	0.30	0.45	0.60	0.75	0.90
0.15	1	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
0.30		1	0.347	0.614	0.967	0.014
0.45			1	0.660	0.125	0.081
0.60				1	0.744	0.035
0.75					1	0.007
0.90						1

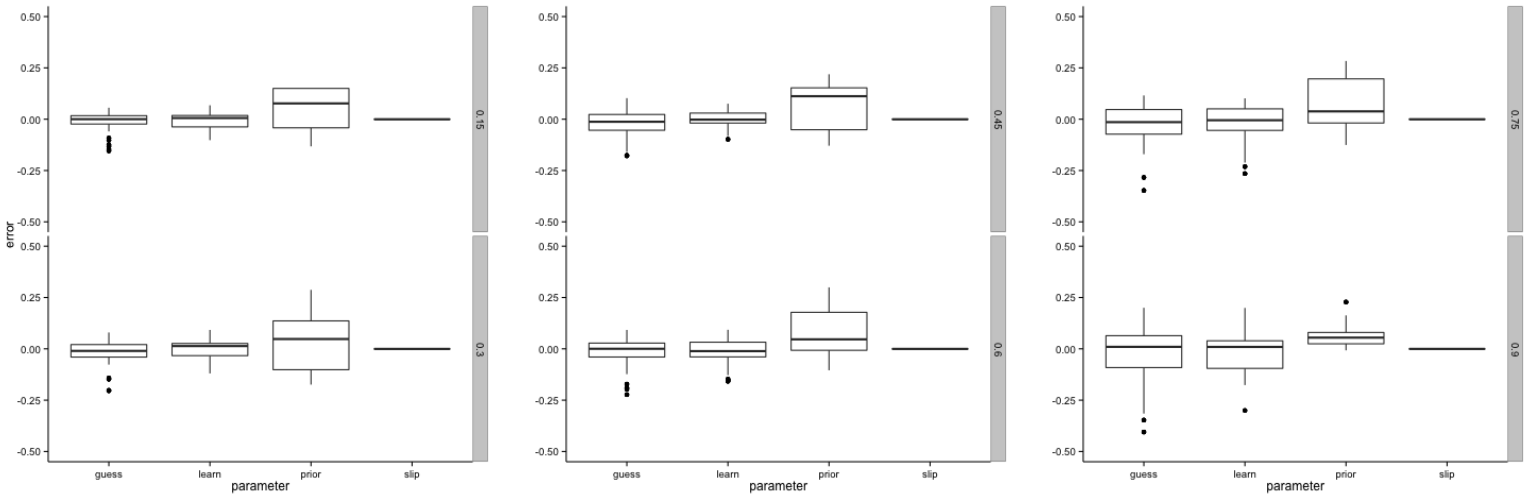

Figure 3: Predicting parameters with different values of prior

significantly better – there 56 values achieved mastery with 5 responses. This indicates that the impact indeed was significant in terms of impact to students learning and highlights the importance of this study.

6. CONCLUSIONS

We started this study with the motivation to explore how well the knowledge tracing method performs when the prior is high or low; this performance has practical implications when applying this approach in a heterogenius classroom where students arrive with highly different knowledge of the domain. We studied this empirically by generating 288 different synthetic datasets and explored the difference between the predicted parameters and the parameters used to generate the dataset.

Our results indicated a slightly increased in the estimation error when prior was 0.90, which we mostly attribute to higher error in learn and guess parameters. This observation was statistically significant and most likely due to the fact that students with higher priors produce less information to be used by the HMM to estimate the guess and learn parameters.

We explored the influence these errors had on the propability of knowledge and observed that these errors significantly reduced the speed students achieved mastery learning. This result therefore implies that more work needs to be done to detect those with high prior knowledge to cater their learning needs.

Acknowledgments

This work was conducted during UC Berkeley School of Information class “INFO290: Machine learning in education“ instructed by Zach Pardos. We thank the support of the course staff and peers on the presentation.

References

[1] RyanS.J.d. Baker, AlbertT. Corbett, and Vincent Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In BeverleyP. Woolf, Esma AÃŕmeur, Roger Nkambou, and Susanne Lajoie, editors, *Intelligent Tutoring Systems*, volume 5091 of *Lecture Notes in Computer Science*, pages 406–415. Springer Berlin Heidelberg, 2008.

[2] Joseph E Beck and Kai-min Chang. Identifiability : A Fundamental Problem of Student Modeling. pages 137–146, 2007. doi: 10.1007/978-3-540-73078-1_17.

[3] Albert Corbett. Cognitive computer tutors: Solving the two-sigma problem. In *User Modeling 2001*, volume 2109 of *Lecture Notes in Computer Science*, pages 137–147. Springer Berlin Heidelberg, 2001.

[4] Albert Corbett, Megan McLaughlin, and K Christine Scarpinatto. Modeling student knowledge: Cognitive tutors in high school and college. *User modeling and user-adapted interaction*, 10(2-3):81–108, 2000.

[5] Albert T Corbett and John R Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4(4): 253–278, 1994.

[6] Albert T Corbett and Akshat Bhatnagar. Student modeling in the act programming tutor: Adjusting a procedural learning model with declarative knowledge. *COURSES AND LECTURES-INTERNATIONAL CENTRE FOR MECHANICAL SCIENCES*, pages 243–254, 1997.

[7] Tanja KÃďser, Severin Klingler, AlexanderGerhard Schwing, and Markus Gross. Beyond knowledge tracing: Modeling skill topologies with bayesian networks. In Stefan Trausan-Matu, KristyElizabeth Boyer, Martha Crosby, and Kitty Panourgia, editors, *Intelligent Tutoring Systems*, volume 8474 of *Lecture Notes in Computer Science*, pages 188–198. Springer International Publishing, 2014.

[8] Z. A. Pardos and N. T. Heffernan. Navigating the parameter space of Bayesian Knowledge Tracing models Visualizations of the convergence of the Expectation

Maximization algorithm. *In Proceedings of the 3rd International Conference on Educational Data Mining*, 2010.

[9] ZA Pardos and NT Heffernan. Using HMMs and bagged decision trees to leverage rich features of user and skill from an intelligent tutoring system dataset. *Journal of Machine Learning Research W & CP*, 2010. URL `http://people.csail.mit.edu/zp/papers/pardos_JMLR_in_press.pdf`.

[10] Z.A. Pardos, M.J. Johnson, and et al. Scaling cognitive modeling to massive open environments. *TOCHI Special Issue on Learning at Scale*, (in preparation).

[11] ZacharyA. Pardos and Neil T. Heffernan. Modeling individualization in a bayesian networks implementation of knowledge tracing. In Paul Bra, Alfred Kobsa, and David Chin, editors, *User Modeling, Adaptation, and Personalization*, volume 6075 of *Lecture Notes in Computer Science*, pages 255–266. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-13469-2.

[12] Pardos, Zachary A, Sujith M. Gowda, Ryan S.J.d. Baker, and Neil T. Heffernan. The sum is greater than the parts. *ACM SIGKDD Explorations Newsletter*, 13(2):37, May 2012. ISSN 19310145. doi: 10.1145/2207243.2207249. URL `http://dl.acm.org/citation.cfm?id=2207249` `http://dl.acm.org/citation.cfm?doid=2207243.2207249`.

[13] Dovan Rai, Yue Gong, and Joseph E Beck. Using dirichlet priors to improve model parameter plausibility. *International Working Group on Educational Data Mining*, 2009.

[14] Leena Razzaq, Neil T Heffernan, Mingyu Feng, and Zachary A Pardos. Developing Fine-Grained Transfer Models in the ASSISTment System. *Technology, Instruction, Cognition & Learning*, 5(3):1–16, 2007.

[15] Steven Ritter, Thomas K Harris, Tristan Nixon, Daniel Dickison, R Charles Murray, and Brendon Towle. Reducing the knowledge tracing space. *International Working Group on Educational Data Mining*, 2009.

[16] A Toscher and Michael Jahrer. Collaborative filtering applied to educational data mining. *Journal of Machine Learning Research*, 2010.

[17] Brett van De Sande. Properties of the Bayesian Knowledge Tracing Model. *Journal of Educational Data Mining*, 5(2):1–10, 2013.

[18] Hsiang-Fu Yu, Hung-Yi Lo, Hsun-Ping Hsieh, Jing-Kai Lou, Todd G McKenzie, Jung-Wei Chou, Po-Han Chung, Chia-Hua Ho, Chun-Fu Chang, Yin-Hsuan Wei, et al. Feature engineering and classifier ensemble for kdd cup 2010. *JMLR: Workshop and Conference Proceedings*, 1, 2010.

[19] Michael V Yudelson and Kenneth R Koedinger. Estimating the benefits of student model improvements on a substantive scale. In *Proceedings of the 6th International Conference on Educational Data Mining*, 2013.

[20] Michael V Yudelson, Kenneth R Koedinger, and Geoffrey J Gordon. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pages 171–180. Springer, 2013.