Spectral Bayesian Knowledge Tracing

Mohammad Falakmasir
University of Pittsburgh
210 South Bouquet Street,
Pittsburgh, PA 15213
(412) 624-5755
falakmasir@pitt.edu

Michael Yudelson
Carnegie Learning, Inc.
437 Grant St.
Pittsburgh, PA 15219
(412) 690-2442
myudelson@
carnegielearning.com

Steve Ritter
Carnegie Learning, Inc.
437 Grant St.
Pittsburgh, PA 15219
(412)-690-2442
sritter@
carnegielearning.com

Ken Koedinger
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
(412)-268-7667
koedinger@cmu.edu

ABSTRACT

Bayesian Knowledge Tracing (BKT) has been in wide use for modeling student skill acquisition in Intelligent Tutoring Systems (ITS). BKT tracks and updates student's latent mastery of a skill as a probability distribution of a binary variable. BKT does so by accounting for observed student successes in applying the skill correctly, where success is also treated as a binary variable. While the BKT served the ITS community well, representing both the latent state and the observed performance as binary variables is, nevertheless, a simplification. In addition, BKT as a two-state and two-observation first-order HMM is prone to noise in the data. In this paper, we present work that uses feature compensation and model compensation paradigms in an attempt to conceptualize a more flexible and robust BKT model. Validation of this approach on the KDD Cup 2010 data shows a tangible boost in model accuracy well over the improvements reported in the literature.

Keywords

Cognitive model of student practice, Bayesian Knowledge Tracing.

1. INTRODUCTION

Bayesian Knowledge Tracing (BKT) is one of the most popular student modeling techniques in the field of Intelligent Tutoring Systems (ITS). It has been used for 20 years now, and it has served the educational community well. Among the major weaknesses of BKT are the non-identifiability of the parameters, parameter degeneracy [1], and, in general, susceptibility to the noise in the naturally-occurring data. BKT is, by definition, a first-order Hidden Markov Model (HMM) with a binary latent variable representing student knowledge and a binary observed variable indicating student performance. While representing latent student knowledge as a binary variable with *known* and *unknown* states has been widely accepted by the Intelligent Tutoring Community (ITS), it is, no doubt, a simplification. Accounts of the need for a larger number of latent states can be found in the literature, including but not limited to the work of Aleven et al. [2].

Practical issues occur in other fields where first-order HMMs are used intensively (e.g., speech recognition, handwriting recognition, etc.). In these fields, it is common to adopt various compensation measures including model compensation and feature compensation [3]. In this paper, we are applying both compensation paradigms to create a variant of BKT – Spectral BKT – in an attempt to overcome some of BKT's shortcomings. Spectral BKT uses spectral observations – *n*-grams of the consecutive original unary observations of correct and incorrect skill application. It also relies on an extended set of latent states. While a number of Spectral BKT configurations can be conceived, we constructed and empirically tested a setup with eight spectral observations (3-grams of original observations) and four states. To validate the Spectral BKT approach uses an openly available KDD Cup 2010 data set of the 2008-2009 Carnegie Learning's Cognitive Tutor data. The resulting improvement is well above all reported in the literature.

2. RELATED WORK

2.1 Bayesian Knowledge Tracing

Bayesian Knowledge Tracing (BKT) was introduced by Corbett and Anderson [4] in 1995. The standard BKT model assumes that student knowledge of a particular skill is an unobserved binary latent variable that changes based on the binary correctness of the observed student performance. Standard BKT has 4 parameters. Probability of knowing skill a priori (pInit), probability of learning the skill after each opportunity to apply it (pLearn), probability of making a mistake when applying an already known skill (pSlip), and probability of luckily producing a correct response when the skill is not known (pGuess). The probability of knowledge decay (pForget) is assumed to be zero in standard BKT. In general, a HMM with two states and two observations that has a total of 10 numeric values would be said to have 5 parameters (last value in every row is redundant). However, since forgetting is set to zero, BKT is assumed to have 4 parameters.

A large volume of work has been published on fitting BKT models and its variations. Wang and Beck [6] introduced two hierarchical factors into BKT to account for and compare class and student level parameter variability. Xu and Mostow [7] blend BKT approach with logistic regression and create an LR-DBN model that is capable of addressing multiple skill coding for a single step (something that BKT technically doesn't, due to conditional independence assumptions). González-Brenes et al. [8] generalized BKT model to address a feature-rich context addressing multiple skills per step, temporal features, and expert knowledge. Another work of Pardos and Heffernan is an extension of BKT call KT-IDEM [9]. It addressed item variance in the data via introducing item difficulty observable nodes.

2.2 Empirical Problems of BKT

A noticeable portion of the work on BKT models is devoted to discussing problems researchers face when fitting them to the data. Baker et al. [1], when talking about the contextual estimation of guess and slip parameters in BKT, stipulate that their model is less prone to the BKT model degeneracy. What is often meant by degeneracy are the cases when probabilities of slipping and guessing assume unjustifiably high values, and this often calls for the use of parameter caps. BKT model degeneracy is the artifact of the known issue in HMM called label switching [10]. The issue is made more convoluted by the fact that forgetting is not allowed to vary in BKT and is set to zero.

Work by Beck and Chang [11] discusses an example of yet another problem of BKT – identifiability. There often exists a range of parameter value sets that result in the same likelihood given the data it's estimated on. Falakmasir and colleagues [12] have encountered the same problem in their previous work on the Spectral Learning approach to fitting BKT models. In that work, the formulation of the best-fitting parameter search problem was transformed into the spectral space, where a global optimum of the objective function is guaranteed to be reached. When translating the spectral solution back to the HMM space, the authors had to define a heuristic to pick the most *plausible* parameters from an infinite set of equally good parameter sets.

2.3 Theoretical Issues with BKT

Arguably, it's the two Markov assumptions and the setup of the BKT that result in its known shortcomings. First, is the Limited Horizon Assumption states that the probability of being in a state at time *t* depends only on the state at time *t*-1. This kind of HMM is called a first-order HMM since it only has a *memory* of one previous time slice. Second Markovian assumption is the Stationary Process Assumption that the conditional distribution over the next state given the current state does not change over time. Given the fact that BKT has only one parameter to capture state transition, student learning rate is forced to remain constant.

Both, the limited *memory*, and the constant learning rate are simplifications and one can easily construct a case for a more flexible representation of skill learning. For example, between the *unknown* state and *known* state there can be states that capture the preliminary stage of learning when the student having just seen one or two problems is mostly guessing. Before transitioning to the known state, the skill could be in the state that often results in slips since student's knowledge is not strong enough. Another likely reason for BKT's limitations is sensitivity to noise. In BKT, Gaussian noise is assumed for the latent (knowing the skill) and the observed variables. However, when dealing with naturally occurring data, the signal to noise ratio might drop considerably. As a result, one might arrive at degenerate model parameters.

There are two main approaches to handling noise in HMM: feature compensation and model compensation. In feature compensation, the noisy traits (for example, observations) are enhanced to remove the effect of the noise. In model compensation, the original models are mapped into a new model that can be learned from the noisy observations. It has been empirically established that feature compensation is simpler and more efficient to implement, but model compensation has the potential for the greater robustness [3].

3. SPECTRAL BKT

In this work, we are attempting to combine feature compensation and model compensation to overcome the shortcomings of the standard BKT that assumes an ideal noise-free environment and is represented by a first-order HMM. We address feature compensation by changing the way we treat the observations. Instead of a single observation, we are considering *n*-grams – sequences of consecutive observations for the skill, where next *n*-gram observation inherits *n*-1 atomic observations from the previous one. In NLP, 3-grams are often successfully used for feature compensation and we have empirically found that 3-grams work sufficiently well while 2-grams do not. From the information-theoretic point of view, the entropy rate of Hidden Markov Processes with two states proved to have at most second order behavior (captured by second-order HMM) [13]. This means that if we consider the data to be generated by a relatively noise-free naturally-occurring process and that the skills are fine-grained enough, we only need to look at 3-grams of the observations in order to find the true model. One may use *n*-grams with *n* greater than 3. However, the computations involved would grow exponentially. Figure 1 shows how the original sequence of observations is encoded into 3-grams.

The model compensation is addressed by adding two intermediate states between the *unknown* and *known* to the original BKT. Once the new observations are defined, the new model that we will call Spectral BKT (due to the use of spectral observations) can be treated as a first order HMM for the purposes of fitting the parameters.

In Spectral BKT, state 1 is the *known* state and state 4 is the *unknown* state. States 2 and 3 we leave unlabeled at this point. Like in the standard BKT, once the student is in the *known* state we assume no un-learning. Moreover, the probability of going from *the unknown* state directly to *the known* state is zero. Finally, once the knowledge transitions from the *unknown* state, there's no return. Given these assumptions, the sparsity structure changes the number of state transition parameters from 1 in standard BKT to 6 in Spectral BKT. By enforcing the sparsity structure in our transition matrix, we guarantee the forward progressing from unknown to known in each iteration and prevent the EM algorithm from learning degenerate models. We assume no further sparsity in any of the 4 priors and 4*7=28 values of the observation matrix, we have (4-1)+6+(7-2)*4=37 parameters in this particular Spectral BKT conceptualization.

The transformation of the original data for fitting the new Spectral BKT is fairly simple (rf. Figure 1). However, when we talk about model predictions, the Spectral BKT produces probability distributions over 8 3-gram observations and one has to make special arrangements to convert them to 2 (probability of correct and of incorrect) in order to compare it with the standard BKT algorithm fairly. First, we ordered the spectral observations from 000 to 111 linearizing a partial order heuristic (rf. Figure 2a). According to this heuristic a spectral observation 011 is the second best indication of success after observation 111. Spectral observation 101 is third best with, potentially, a careless slip in the middle. Spectral observations 001 and 110 were a judgment call. We have placed 001 before 110, assuming it is an early indicator of learning, and 110 is a premature indicator of learning with a failure in the end.

When mapping 8 values to binary success and failure, we came up with three rules. A *regular* rule splits 8 probabilities exactly in half and sums of the two groups are the new probability of correct and incorrect (third column in Figure 2a). The *regular* rule can also be interpreted as looking at the third bit of each 3-gram A *strict* rule is more stringent about which observation probabilities are counted toward success. A *relaxed* rule is more. Since our Spectral BKT produces a first 8-probability predictions starting with the third original observation (due to the use of 3-grams), we have also devised mapping of the 8 probabilities to produce predictions for the first two observations. These mappings are given in Figure 2b,c and reference the spectral observations from Figure 2a. For example, if the observed data contained observations 0, 1, and 0, and the Spectral BKT prediction of

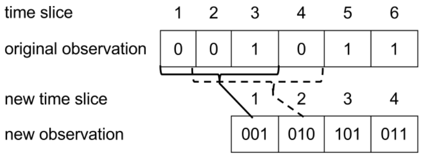
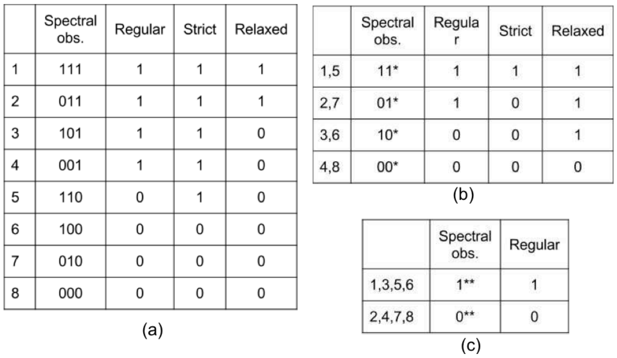


Figure 1. New *n*-gram observations



	Spectral obs.	Regular	Strict	Relaxed
1	111	1	1	1
2	011	1	1	1
3	101	1	1	0
4	001	1	1	0
5	110	0	1	0
6	100	0	0	0
7	010	0	0	0
8	000	0	0	0

(a)

	Spectral obs.	Regular	Strict	Relaxed
1,5	11*	1	1	1
2,7	01*	1	0	1
3,6	10*	0	0	1
4,8	00*	0	0	0

(b)

	Spectral obs.	Regular
1,3,5,6	1**	1
2,4,7,8	0**	0

(c)

Figure 2. Mapping spectral observations from a distribution over 8 probabilities to 2: a) predicting starting with a 3rd original observation when two prior observations are available. b) & c) predicting original observations 2 and 1.

correctness was {0, 0, 0.1, 0.1, 0.1, 0.2, 0.2, 0.3}, then, according to the regular mapping rules, probabilities of correct for the three observations would have been 0.4, 0.3, and 0.2.

4. DATA

To validate our models we used data from KDD Cup 2010 donated by Carnegie Learning, Inc. and available for downloading at http://pslcdatashop.web.cmu.edu/KDDCup. Of the two datasets available we chose Bridge to Algebra. This dataset contains about 20 million transactions belonging to over 6 thousand students working on nearly 150 sections of mathematics curriculum practicing around 1650 skills. The dataset contains information about curriculum context (unit and section the student is in), problem context (problem name and problem step name), cognitive skill labels, timing, as well as correctness of the first attempt to solve the problem step and assistance information (number of hints requested and number of errors). The KDD Cup 2010 is currently the largest freely available collection of learner data. That and the fact that this data was collected by Carnegie Learning's Cognitive Tutor that uses BKT model makes it a good candidate for testing the Spectral BKT. According to the custom of the Carnegie Learning’s Cognitive Tutor, skills were considered unique within each curriculum section even if the skill label repeated across several sections. Also, we have treated an absence of the skill (a null skill) as a special skill.

5. MODEL VALIDATION

For the purposes of training the models, we have transformed the original data with unigram observations into a dataset with 3-gram observations. We ran 10-fold student-based and item-based cross-validations that each produced a set of predictions for the transformed 3-gram data. To fit and cross-validate the models we used the `hmmsclbl` tool – a C/C++ utility specially developed to work with large data sets and successfully used in [5] (available for download at http://github.com/IEDMS/standard-bkt). Standard BKT outputs two predictions per data row – probability of correct application of the skills in question and the probability of incorrect application. Spectral BKT works with 8 spectral observations and its predictions come in the form of probability distributions of 8 values per row of the predicted data. Spectral BKT models predictions were mapped from the 8-values onto the 2-value probability distribution schema in Figure 2. The summary of the cross-validation results for the training dataset is listed in Table 1. Here we list the performance of standard BKT next to the performance of Spectral BKT model. We only list results the *relaxed* 3-gram-to-unigram mapping, since *regular* and *strict* mapping performed worse. We tested several solver algorithms `hmmsclbl` supports, including EM and stochastic gradient descent. EM gave a consistently better performance, but the margin was small: within 1% in accuracy and 0.03 in RMSE.

When running student-stratified cross-validation, we were repeatedly *hiding* the full data belonging to 10% of the students. In item-stratified cross-validation, the transactions belonging to problems that we intended to *hide* could appear in individual students’ data in arbitrary locations. For the purposes of item-stratification, we have marked the data of 10% of the items as unobserved but accounted for the opportunity to apply skills.

Standard BKT model has 4 parameters per skill. Spectral BKT model, as per our conceptualization of the transition matrix, has 37. The number of parameters being an order of magnitude higher, the AIC and BIC metrics that penalize for that go up 3% and 9% (item-stratified cross-validation). In the case of student-stratified cross-validation, both AIC and BIC are decreaseв by 21% and 13%. Accuracy and RMSE in case of Spectral BKT improve a lot. To the best of our knowledge, the overall accuracy of BKT or its variations was never reported to be above 90% on the dataset we used and Spectral BKT hits an impressive 92%.

Recall that we had to *back-predict* the predictions of Spectral BKT for student skill opportunities one and two due to the use of 3-gram observations. For this purpose, in Table 1 we list the additional accuracy and the RMSE values for student skill opportunity 1 alone (7% of the data), opportunity 2 alone (6% of the data), and opportunity 3 and further (87% of the data). To no surprise, the first opportunity prediction of Spectral BKT is slightly worse than the one of standard BKT by a margin in the

Table 1. Comparison of cross-validation results for standard BKT and Spectral BKT

					All opportunities		**Opportunity 1**		**Opportunity 2**		**Opportunity 3+**	
Model	**Par/skill**	**CV**	**AIC**	**BIC**	**Acc.***	**RMSE**	**Acc.**	**RMSE**	**Acc.**	**RMSE**	**Acc.**	**RMSE**
BKT	4	item	15380089	15478083	0.8609	0.3293	0.7469	0.4112	0.8099	0.3731	0.8740	0.3181
Spectral BKT	37	item	15853433	16758456	0.9208	0.2472	0.7472	0.4146	0.8915	0.2940	0.9337	0.2289
BKT	4	student	13947080	14045074	0.8659	0.3153	0.7435	0.4108	0.8126	0.3665	0.8799	0.3020
Spectral BKT	37	student	11553442	12458465	0.9196	0.2405	0.7469	0.4130	0.8897	0.2937	0.9325	0.2210

* For a reference, the majority class accuracy of predicting correct response for every row is 0.8569.

third digit of both accuracy and RMSE, both around 74% and 0.41 respectively. On the second opportunity prediction, Spectral BKT has a decisive edge of almost 9% and 0.08 in RMSE. On the third opportunity and further, Spectral BKT has a comfortable advantage of around 5% in accuracy and 0.09 in RMSE.

6. DISCUSSION

The performance of the Spectral BKT demonstrated a tangible improvement over standard BKT and only with an incremental change in the underlying computations. We attribute the boost in predictive performance to the several factors. First, feature compensation via considering 3-grams of original observations allows for a more stable estimate of the learning process. In a sequence of responses {0,1,0,1,1}, the third value of 0 would be treated a potential slip by the standard BKT. At the same time, Spectral BKT would consider it, as a part of the first triple {0,1,0} to be the *noisy guessing*, and then, in the second triple {1,0,1}, as part of the *noisy slipping*. Finally, in the third triple {0,1,1}, 0 would be considered to be a part of noise-free learning pattern. The fact that there are more than 2 states allows Spectral BKT to represent an intermediate configuration of student learning in addition to just known or unknown. As a result, Spectral BKT is able to deal with the noise in the observations better.

The interpretation of a new conceptualization of the process of learning remains an open question. Having agreed on that state 1 is the known state and state 4 is the unknown state, we could offer several hypotheses of what the remaining middle states are. The first hypothesis relates to the linear view of the stages of mastering the skill. When a student just started learning and only seen a few problems, their knowledge is overly specific, and they would end up guessing and failing a lot. We can call this state 3 – *too-specific*. Once the student sees more problems and starts to generalize the knowledge, they would still occasionally slip due to over-generalization. We can call this state 2 – *too-general*.

Our second hypothesis is related to a publication by Aleven and colleagues [2]. In this work, authors study the metacognitive behavior of students by administering two types of tutors. First, the cognitive tutor that implements a mastery learning approach. Second, the meta-cognitive tutor used a previously created model of effective and ineffective help-seeking behavior in order to study the effect of different meta-cognitive traits on learning. Authors conclude that the use of the standard BKT model with two states might be limiting the capability of the meta-cognitive tutor to offer effective help due to lack of intermediate states between the *known* state and *unknown* state that might give us a better insight into student behavior. In the light of the work by Aleven et al., the progression of the states could be reflecting an interaction of binary latent capturing skill mastery (*known*, *unknown*) with the binary latent capturing effective use of meta-cognitive strategies (2 mastery states * 2 metacognitive states = 4 overall states). To address this hypothesis, one might consider step durations (available in the original dataset) or design and run a focused investigation like the one in [2].

In our work, we used 3-grams of original binary observations, giving us 8 new spectral observations and we also used 4 states. This particular setup can be changed in the search of a better Spectral BKT model. Increasing the number of states could be potentially beneficial. However, one must be careful, for as the number of states grows, the chance to observe relevant patterns of binary observations drops and the Spectral BKT might be under-defined and this could have problems with performing on unseen data whether the patterns missing from the training set are present. When there are fewer states that there are spectral observations, the states serve the aggregation role. We empirically tried configurations of Spectral BKT with 2 states and 4 bigram spectral observations that did not result in an improvement over standard BKT, as well as a configuration with 8 states and 16 4-gram spectral observations that did not result in a tangible improvement over the configuration we discussed in this paper.

7. ACKNOWLEDGMENTS

This work is partially supported by the Andrew Mellon Pre-doctoral Fellowship and extends a project initiated during the Pittsburgh Science of Learning Center's Summer School at Carnegie Mellon University. We extend our special thanks to Dr. Geoffrey J. Gordon from Machine Learning Department at Carnegie Mellon University for insightful comments.

8. REFERENCES

[1] Baker, R.S.J.d., Corbett, A.T., Aleven, V. (2008) Improving Contextual Models of Guessing and Slipping with a Truncated Training Set. EDM 2008, pp.67-76.

[2] Aleven, V., McLaren, B., Roll, I., & Koedinger, K. (2006). Toward meta-cognitive tutoring: A model of help seeking with a Cognitive Tutor. *International Journal of Artificial Intelligence and Education*, 16, 101-128.

[3] Gales, M. & Young, S. (2007) The Application of Hidden Markov Models in Speech Recognition. *Foundations and Trends in Signal Processing*, 1(3): 195-304.

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

[5] Yudelson, M., Koedinger, K., & Gordon, G. (2013) Individualized Bayesian Knowledge Tracing Models. AIED 2013, pp. 171-180.

[6] Wang, Y. & Beck, J. (2013) Class vs. Student in a Bayesian Network Student Model. AIED 2013, pp. 151-160.

[7] Xu, Y. & Mostow, J. (2012). Comparison of methods to trace multiple subskills: Is LR-DBN best? EDM 2012, pp.41-48.

[8] González-Brenes, J.P., Huang, Y., & Brusilovsky, P. (2014). General Features in Knowledge Tracing: Applications to Multiple Subskills, Temporal Item Response Theory, and Expert Knowledge. EDM 2014, pp.84-91.

[9] Pardos, Z. & Heffernan, N. (2011) KT-IDEM: Introducing Item Difficulty to the Knowledge Tracing Model. UMAP 2011, pp. 243-254.

[10] Redner,R. & Walker,H. (1984) Mixture densities, maximum likelihood and the EM algorithm. SIAM Rev., 26, 195-239.

[11] Beck, J. & Chang, K.-m. 2007. Identifiability: A fundamental problem of student modeling. UM 2007, 137–146.

[12] Falakmassir, M.H., Pardos, Z.A., Gordon, G.J., & Brusilovsky, P.L. (2013) A Spectral Learning Approach to Knowledge Tracing. EDM 2013, pp. 28-34.

[13] Zuk, O., Kanter, I., Domany, E., & Aizenman, M. (2006) Taylor series expansions for the entropy rate of hidden Markov processes. ICC 2006, pp.1598-1604.