International Journal of Artificial Intelligence in Education 18 (2016) xxx – xxx
Springer

Automated Test Assembly for Handling Learner Cold-Start in Large-Scale Assessments

Jill-Jênn Vie
RIKEN Center for Advanced Intelligence Project
Nihonbashi 1-4-1, Mitsui Building 15F, Chuo-ku
Tokyo 〒103-0027
jill-jenn.vie@riken.jp
http://jilljenn.github.io

Fabrice Popineau, Yolaine Bourda
Laboratoire de recherche informatique – Bât. 650 Ada Lovelace
Université Paris-Sud
91405 Orsay Cedex, France
{fpopineau, bourda}@lri.fr

Éric Bruillard
École normale supérieure de Paris-Saclay – STEF – Bât. Cournot
61 avenue du Président Wilson
94230 Cachan, France
eric.bruillard@stef.ens-cachan.fr

Abstract. In large-scale assessments such as the ones encountered in MOOCs, a lot of usage data is available because of the number of learners involved. Newcomers, that just arrive on a MOOC, have various backgrounds in terms of knowledge, but the platform hardly knows anything about them. Therefore, it is crucial to elicit their knowledge fast, in order to personalize their learning experience. Such a problem has been called learner cold-start. We present in this article an algorithm for sampling a group of initial, diverse questions for a newcomer, based on a method recently used in machine learning: determinantal point processes. We show, using real data, that our method outperforms existing techniques such as uncertainty sampling, and can provide useful feedback to the learner over their strong and weak points.

Keywords. cold-start, test-size reduction, learning analytics, determinantal point processes, multistage testing, cognitive diagnosis.

INTRODUCTION

Learning analytics is a recent field in educational research that aims to use the collected data in learning systems in order to improve learning. Massive online open courses (MOOCs) receive hundreds of thousands of users that have acquired knowledge from different backgrounds. Their performance in the assessments provided by the platform is recorded, therefore it is natural to wonder how to use it to provide personalized assessments for new users. So far adaptive tests have been mainly designed for standardized tests, but recent developments have enabled their use in MOOCs (Rosen et al., 2017).

Whenever a new learner arrives on a MOOC, the platform knows nothing about the knowledge they have acquired in the past: this problem has been coined as *learner cold-start* (Thai-Nghe et al., 2011). Thus, it is important to be able to elicit their knowledge in few questions, in order to prevent boredom or frustration, and possibly recommend to them lessons they need to master before they can follow the course (Lynch and Howlin, 2014). Such learner profiling is a crucial task, and raises the following question: how to sample efficiently questions from an item bank for a newcomer, in order to explore their knowledge at best, and provide to them valuable feedback?

In this paper, we present a new algorithm for selecting the questions to ask to a newcomer. Our strategy is based on a measure of diversity called determinantal point processes that has recently been applied to large-scale machine-learning problems, such as document summarization or clustering of thousands of image search results (Kulesza and Taskar, 2012). We apply it to the automatic generation of low-stake practice worksheets by sampling diverse questions from an item bank. Our method is fast, and only relies on a measure of similarity between questions, therefore it can be applied in any environment that provides tests, such as large residential courses, or serious games.

This article is constructed as follows. First, we expose the background studies related to this work, in student modeling in assessment, cold-start in multistage testing and determinantal point processes. Then, we clarify our research context. We then explain our strategy, InitialD, for tackling the cold-start problem. Finally, we expose our experiments, results and discussion.

BACKGROUND STUDIES

Student Modeling in Assessment

Student models allow to infer latent variables from the learners based on their answers throughout a test, and potentially predict future performance. Several models have been developed that we describe below under three categories: Item Response Theory, Knowledge Tracing and Cognitive Diagnosis.

Item Response Theory (IRT) In assessments such as standardized tests, learners are usually modelled by one static latent variable – Rasch model (Hambleton and Swaminathan, 1985) – or several latent variables – Multidimensional Item Response Theory aka MIRT (Reckase, 2009; Chalmers, 2016). Based on the binary outcomes (right or wrong) of the learners over the questions, an estimate of their latent variables is devised. Therefore, this category of models allow summative assessment, e.g., scoring and ranking examinees. MIRT models are said to be hard to train (Desmarais and Baker, 2012). However,

with the help of Markov chain Monte Carlo methods such as Metropolis-Hastings Robbins-Monro (Cai, 2010a), efficient techniques have been proposed and implemented in ready-for-use tools (Chalmers, 2016) to train MIRT models in a reasonable time. Variants have been designed in order to incorporate several attempts from the learners over a same question (Colvin et al., 2014) or evolution over time (Wilson et al., 2016).

Knowledge Tracing In this family of models, learners are modelled by a *latent state* composed of several binary variables, one per knowledge component (KC) involved in the assessment. One particular advantage of the Bayesian Knowledge Tracing model (BKT) is that the latent state of the learner can change after every question they attempt to solve. Questions are tagged with only one KC. Based on the outcomes of the learner, the KCs they may master are inferred. Some variants include *guess* and *slip* parameters from the learners, i.e., probabilities of answering a certain question correctly while the corresponding skill is not mastered, or answering a question incorrectly while the skill is mastered. At the end of the test, a feedback can be provided to the learner, based on the KCs that seem to be mastered and those that do not. Therefore, this category of models allows formative assessment that benefit learning (Dunlosky et al., 2013) and detection of learners that need further instruction. More recently, Deep Knowledge Tracing (DKT) models have been developed (Piech et al., 2015), based on neural networks, outperforming the simplest BKT models at predicting student performance. Quite surprisingly, a study (Wilson et al., 2016) has even proven that a variant of unidimensional IRT performed better than DKT. Presumably, this is because IRT models measure a latent variable shared across all questions while in BKT, a question will only give information over the KC it involves. Also, IRT models are simpler, therefore less prone to overfitting than DKT models.

Cognitive Diagnosis This family of models is similar to Knowledge Tracing. Learners are modelled by a static latent state composed of K binary variables, where K is the number of KC involved in the assessment. Cognitive diagnostic models allow mapping questions to several KCs involved in their resolution. For example, the DINA model (De La Torre, 2009) requires that every KC involved in a question should be mastered by a learner so they can answer it correctly. DINA also considers slip and guess parameters. Being given the answers of learners, it is possible to infer their mastery or non-mastery of each KC, according to some cognitive diagnostic model (Desmarais and Baker, 2012). At the end of the test, the estimated latent state of the learner can be provided as feedback, in order to name their strong and weak points. Therefore, this category of models allow formative assessment. In order to reduce the complexity of 2^K possible latent states, some variants of these models allow specifying a dependency graph over KCs (Leighton, Gierl, and Hunka, 2004). Lynch and Howlin (2014) have developed such an adaptive test for newcomers in a MOOC that relies on a dependency graph of knowledge components to learn, but they do not consider slip and guess parameters, i.e., their method is not robust to careless errors from the examinees.

Adaptive and Multistage Testing

Multistage testing (MST) (Zheng and H. Chang, 2014) is a framework that allows asking questions in an adaptive, tree-based way (see Figure 1): according to their performance, examinees are routed towards new questions. Therefore, learners follow a path in a tree, of which every node contains a pool of questions, and different learners may receive different questions, tailored to their level. Computerized adaptive testing (CAT) is a special case of multistage testing in which every node contains a single question.

While it is possible to assemble MST tests manually, *automated test assembly* algorithms save human labor and allow specifying statistical constraints (Zheng and H. Chang, 2014; Yan, A. A. v. Davier, and Lewis, 2014). Three approaches are mainly used, either exact methods that solve optimization problems but are usually long to compute, greedy heuristics that do not get optimal subsets, or Monte Carlo methods that rely on random samples.

Such CAT systems have been successfully used in standardized tests such as the GMAT and the GRE, administered to hundreds of thousands of students. Such an idea of selecting the next question based on the previous outcomes has recently been embedded in MOOCs (Rosen et al., 2017), where it has proven successful in eliciting the knowledge of the students with fewer questions. In the literature, some approaches assume that the level of the learners can change between questions, such as (Rosen et al., 2017), some others do not.

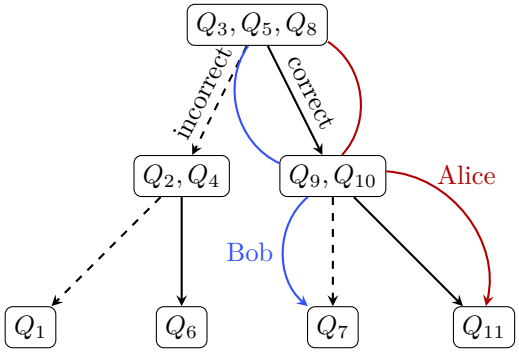

Fig.1. Example of multistage test. Alice and Bob do not receive the same set of questions during the test.

CAT systems rely on a student model that can predict student performance, such as the ones presented above. Based on this model, a CAT system can infer the student parameters, and ask questions accordingly, e.g., ask harder questions if the estimated latent ability of the student is high. Lan et al. (2014) have developed a framework called Sparse Factor Analysis (SPARFA), similar to MIRT, and they have shown that adaptive strategies provided better results for SPARFA than non-adaptive ones for the same number of questions asked. Cheng (2009) have applied adaptive strategies to the DINA model. Such cognitive-diagnostic computerized adaptive tests (CD-CAT) have been successfully applied in Chinese classrooms (H.-H. Chang, 2014; Zheng and H. Chang, 2014). Vie et al. (2016b) have applied adaptive strategies for the General Diagnostic Model (M. Davier, 2005), which is another cognitive di-

agnostic model, and proven that it outperformed adaptive tests based on the DINA model at predicting student performance. For a review of recent models in adaptive assessment, see Vie et al. (2016a).

Cold-Start

As we want to predict student performance, we try to infer the outcome of the learner over new questions, based on their current parameter estimates, according to their previous outcomes. But at the very beginning of the test, we have no data regarding the learner's knowledge. This is the learner cold-start problem, a term usually coined for recommender systems (Anava et al., 2015). In an educational setting, at the best of our knowledge, the best reference to cold-start is Thai-Nghe et al. (2011), where they say that the cold-start problem is not as harmful as in an e-commerce platform where new products and users appear every day. But with the advent of MOOCs, the cold-start problem regains some interest.

CAT systems usually compute an estimate of the student parameters at some point during the test. But when few questions are asked, such an estimate may not exist or may be biased (Chalmers, 2012; Lan et al., 2014), for example if all outcomes provided by the learner are correct (or all incorrect) and if we want to compute a maximum-likelihood estimate. This is why the choice of the very first questions is important. Chalmers (2016) suggests to ask a group of questions before starting the adaptive process, and this is the work we focus on in this paper.

Determinantal Point Processes

Determinantal Point Processes (DPP) rely on the following idea: if the objects we try to sample from can be represented in a way that we can compute a similarity value over any pair of elements (*kernel*), then it is possible to devise an algorithm that will sample efficiently a subset of elements that are diverse, i.e. far from each other. DPPs have been successfully applied to a variety of large-scale scenarios: sampling key headlines from a set of newspaper articles, summarizing image search results (Kulesza and Taskar, 2012), but to the best of our knowledge, not in cold-start scenarios.

In this paper, we apply this technique to the sampling of questions for the initial stage of the test. Indeed, questions are usually modelled by parameter vectors, be it the KCs that are required, or some difficulty parameters, or some tag information. A question will measure the knowledge of the learner in the direction of its parameter vectors, and close questions in parameter space will measure similar constructs. If we want to reduce the number of questions asked, we should avoid selecting questions that measure similar knowledge, which is why it is natural to rely on a measure of diversity.

RESEARCH CONTEXT AND NOTATIONS

Data

We assume we have access to a list of questions from a test, called *item bank*. This item bank has a history: learners have answered some of those questions in the past.

We only consider dichotomous answers of the learners, i.e., raw data composed of lines of the following form:

- a `student_id` $i \in \{1, \ldots, m\}$;
- a `question_id` $j \in \{1, \ldots, n\}$;
- whether student i got the question j *right* (1) or *wrong* (0).

The *response pattern* of a student i is a sequence $(r_{i1}, \ldots, r_{im}) \in \{0, 1, \text{NA}\}$ where r_{ij} denotes the correctness of student i's answer over question j. NA means that the learner i did not try to answer question j. Such response patterns enable us to calibrate a student model.

As the assessment we are building should be formative, we assume we have access to a mapping of each question with the different knowledge components (KC) that are involved in its resolution: it is a binary matrix called *q-matrix* of which the (j, k) entry q_{jk} is 1 if the knowledge component k is involved in question j, 0 otherwise. Thanks to this structure, we can name the KCs.

Student Model: General Diagnostic Model

Student models rely on different components:

- features for every question and learner (the learner features will be called *ability features*), under the form of vectors with real values;
- probability model that a certain learner answers a certain question correctly;
- training step in order to extract question and learner features from an history of answers;
- update rule of learner parameters based on learner answers.

The probability model relies solely on the features, i.e., for a fixed student, questions with close features will have a similar probability of being solved correctly. Also, students with similar ability features will produce similar response patterns.

In order to extract question and learner features for our purposes, we choose to calibrate a General Diagnostic Model, because it predicts student performance better than other cognitive diagnosis models (Vie et al., 2016b):

- K is the number of knowledge components of the q-matrix.
- Learners i are modelled by ability vectors $\theta_{\mathbf{i}} = (\theta_{i1}, \ldots, \theta_{iK})$ of size K. When estimated, any component θ_{ik} can be interpreted as the strength of the student over KC k.
- Questions j are modelled by discrimination vectors $\mathbf{d_j} = (d_{j1}, \ldots, d_{jK})$ of size K and an easiness parameter δ_j. When estimated, any component d_{jk} can be interpreted as how much question j discriminates over KC k. Such vectors can be specified by hand, or directly learned from data (Vie et al., 2016b).
- According to the q-matrix $(q_{jk})_{j,k}$, some parameters are fixed at 0: $\forall j, k, (q_{jk} = 0 \Rightarrow d_{jk} = 0)$.

Thus, the probability model that a learner answers correctly a question is given by the formula:

$$Pr(\text{“Student } i \text{ answers question } j \text{ correctly”}) = \Phi\left(\sum_{k=1}^{K} \theta_{ik} q_{jk} d_{jk} + \delta_j\right) = \Phi(\theta_{\mathbf{i}} \cdot \mathbf{d_j} + \delta_j) \quad (1)$$

where $\Phi : x \mapsto 1/(1+e^{-x})$ is the logistic function.

We assume that the level of the learner does not change after they give an answer. This is reasonable because the learner only knows their results at the end of the sequence of questions, not after every answer they make. So for the very first selected pool of questions we can make this assumption, that allows us to prefer IRT models or Cognitive Diagnosis models over Knowledge Tracing models.

In a cold-start scenario, we do not have any data about a newcomer. But we do have discrimination vectors for the questions, calibrated using the response patterns of the student data. These can allow us to sample k questions from the item bank of n questions and ask them to the newcomer. According to their answers, we can infer the learner ability features, and compute their performance over the remaining questions using the response model.

Diversity Model

Let us call V the matrix of question feature vectors, of which the j-th line is $\mathbf{d_j} = (d_{j1}, \ldots, d_{jK})$. As a recall, such features allow us to know whether a question measures more one knowledge component than another. A measure of diversity for a subset $A \subset \{1, \ldots, n\}$ of questions is the volume of the parallelotope formed by their feature vectors (Kulesza and Taskar, 2012):

$$Vol(\{\mathbf{d_j}\}_{j \in A}) = \sqrt{\det V_A V_A^T} \quad (2)$$

where V_A denotes the submatrix of V that only keeps the lines indexed by A. Indeed, if two vectors are correlated, the volume of the corresponding parallelotope will be 0. If k vectors form an orthonormal basis, the volume of the corresponding parallelotope will be 1.

Testing every possible subset would be not feasible, as there are $\binom{n}{k}$ of them, and computing their volume has a complexity $O(k^2K + k^3)$. Furthermore, determining the best subset is NP-hard, therefore intractable in large-scale scenarios (Kulesza and Taskar, 2012).

In order to implement our strategy, we also need a kernel $\mathcal{K}$ that helps compute a similarity value for each pair of questions. It can be represented by a positive definite matrix L such that $L_{ij} = \mathcal{K}(\mathbf{d_i}, \mathbf{d_j})$. For our experiment, we choose the linear kernel $\mathcal{K}(\mathbf{d_i}, \mathbf{d_j}) = \mathbf{d_i} \cdot \mathbf{d_j}$ (Kulesza and Taskar, 2012).

OUR SAMPLING STRATEGY: INITIALD

The main idea of our strategy is the following: if we want to ask as few questions as possible, we want to minimize redundancy in the learner's answers, and sample questions that measure different knowledge components. In other words, we want to sample vectors that are the least correlated possible, therefore the most diverse possible. For example, in Figure 2, those three questions are embedded in a 2-dimensional space. Questions 1 and 2 measure the second knowledge component more than the first

one, while conversely, question 3 measures KC 2 more than KC 1. If only two questions should be sampled, questions 1 and 3 will be more informative (less correlated) than questions 1 and 2.

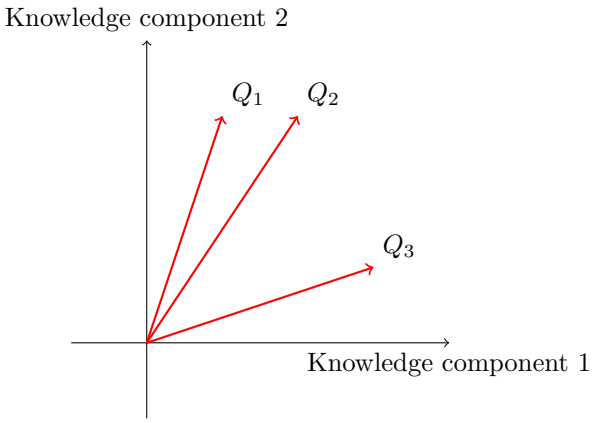

Fig.2. Example of question feature vectors.

Determinantal Point Processes

A probability distribution over subsets of $\{1, \ldots, n\}$ is called a *determinantal point process* if any subset Y drawn from this distribution verifies, for all subset $A \subset \{1, \ldots, n\}$:

$$Pr(A \subset Y) \propto \det L_A \tag{3}$$

where L_A is the squared submatrix of L indexed by elements of A in row and column.

In order words, subsets of diverse questions are more likely to be drawn than subsets of correlated questions. Indeed, if V_A is the submatrix of V that only keeps lines indexed by A, $L_A = V_A V_A^T$ because the kernel is linear, so $Pr(A \subset Y)$ is proportional to $\det L_A = Vol(\{\mathbf{d_j}\}_{j \in A})^2$.

Determinantal point processes have a useful property: k elements among n can be drawn with complexity $O(nk^3)$, after a unique preprocessing step in $O(n^3)$ (Kulesza and Taskar, 2012). This costly step relies in getting the eigenvalues of the kernel matrix, which can be reduced to $O(nK^2)$ in the case of the linear kernel, as the question features span over $K << n$ dimensions.

After the preprocessing step, sampling 10 questions over an item bank of size 10000 takes about 10 million operations, which makes it suitable for large-scale data.

Sampling Strategy

InitialD (for *Initial Determinant*) samples k questions from the item bank according to a determinantal point process, using the question feature vectors. It selects the first bulk of questions, in a test, before any adaptation can be made.

After those questions are asked to the newcomer, a first estimate of the learner ability features can be computed. As the knowledge components are known, such an ability feature can provide a useful feedback for the learner, specifying which KCs need further work.

Sampling several times will give different subsets. This allows balancing over the item bank, and not asking the same bulk of questions to every newcomer, which is interesting for security purposes, and allows calibrating various items from the bank (Zheng and H. Chang, 2014).

VALIDATION

We compared the following strategies for asking the first k questions to a newcomer.

Random Sample k questions from the item bank at random.

Uncertainty Sample the k questions for which the probability that the learner will answer them correctly is closest to 0.5 (which means, questions of average difficulty). This method is known as uncertainty sampling in the context of active learning (Settles, 2010).

InitialD Sample k questions according to a determinantal point process, as described above.

For our experiments, we could use the following datasets.

TIMSS Trends in International Mathematics and Science Study (TIMSS) organizes a standardized test in mathematics. The collected data is freely available on their website for researchers, in SPSS and SAS formats. This dataset comes from the 2003 edition of TIMSS. It is a binary matrix of size 757×23 that stores the results of 757 learners from 8th graders over 23 questions in mathematics. The q-matrix was specified by experts from TIMSS and has 13 knowledge components that are described in Su et al. (2013). It was available in the R package CDM (Robitzsch et al., 2014).

Fraction This dataset contains the results of 536 middle school students over 20 questions about fraction subtraction. Items and the corresponding q-matrix over 8 knowledge components are described in DeCarlo (2010).

Cross validation

Such a benchmark is performed using cross-validation. 80% of the learners from the history are supposed known, while 20% are kept for simulation of a cold-start situation. The cross validation is stratified, i.e., students have close distribution between the train set and the test set.

Model calibration for feature extraction

Using the history of answers, we want to extract the question and learner features that explain learner data best. For this, we minimize the log-loss between the true response patterns and those computed by the General Diagnostic Model:

$$\text{logloss}(y, y^*) = \frac{1}{n}\sum_{k=1}^{n} \log(1 - |y_i - y_i^*|). \tag{4}$$

This optimization problem has been difficult to solve for MIRT models and the General Diagnostic Model but since (Cai, 2010a; Cai, 2010b), a Metropolis-Hastings Robbins-Monro algorithm has been devised, allowing faster calibration. It has been implemented in the R package mirt (Chalmers, 2012), that we used.

Experimental Protocol

The experimental protocol was implemented according to the following algorithm.

Algorithm 1 Simulating strategies for selecting the k first questions

procedure SIMULATECOLDSTART(strategy S, I_{train}, I_{test}, k questions)
  $(\mathbf{d_j})_j, (\delta_j)_j \leftarrow$ TRAININGSTEP($D[I_{train}]$)
  for every learner s of the test set I_{test} **do**
    $\theta \leftarrow$ PRIORINITIALIZATION()
    $Y \leftarrow$ FIRSTBUNDLE(strategy S, k questions, $(\mathbf{d_j})_j, (\delta_j)_j, \theta$)
    Ask questions Y to the learner s
    Collect success/error values $(r_i)_{i \in Y}$ of their answers
    $\theta \leftarrow$ ESTIMATEPARAMETERS($\{(i, r_i)\}_{i \in Y}, \theta$)
    $p \leftarrow$ PREDICTPERFORMANCE($\theta, (\mathbf{d_j})_j, (\delta_j)_j$)
    $\sigma \leftarrow$ EVALUATEPERFORMANCE($p, D[s], \theta$)
  end for
  Compute the mean performance metrics $\bar{\sigma}$ over the test set
end procedure

TrainingStep: feature extraction

According to the available response patterns of the train set $D[I_{train}]$, question vector features are extracted, as described above.

PriorInitialization: initialization of a newcomer

At the beginning of the test, the newcomer's ability features are set to 0, which means the probability that they answer question j correctly is $\Phi(\delta_j)$, i.e., it depends solely on the facility parameter (or bias).

FirstBundle: sampling the first questions

This is where strategies Random, Uncertainty or InitialD are applied. They return a subset of k questions $Y \subset \{1, \ldots, n\}$.

EstimateParameters: estimate learner ability features based on the collected answers

Given their answers $(r_i)_{i \in Y}$, the ability features θ of the learner are inferred using logistic regression. If the answers are all-right or all-wrong, other algorithms are used, defined in (Chalmers, 2012; Lan et al., 2014; Magis, 2015).

As the q-matrix has been specified by an expert, the k-th component of the ability feature θ can be interpreted as a degree of mastery of the learner over the k-th knowledge component.

PredictPerformance: based on the learner ability estimate, compute performance over the remaining questions

Knowing the learner ability features and the question feature vectors, we can compute the probability that the learner will answer correctly every question by applying the formula at equation 1.

EvaluatePerformance: performance metrics

We compute two performance metrics. The log-loss, as stated in equation 4, and the distance $d(\theta, \theta^*) = ||\theta - \theta^*||$ to the final diagnosis θ^*, that is, the learner ability estimate when all questions have been asked (Lan et al., 2014).

Adding a CAT to the benchmark

We also added a normal computerized adaptive test to the benchmark, called GenMA, developed in (Vie et al., 2016b) and proven better at predicting performance than the Rasch and DINA models on a variety of datasets. We choose at each step the question that maximizes the determinant of Fisher information (the so-called D-rule, see Chalmers (2016)).

RESULTS

Results are given in Figures 3 to 6, where the x-axis represents the number of questions asked for tackling the cold-start problem. On the y-axis, either the log-loss, or the distance to the final diagnosis of each of the 3 strategies described above for cold-start, and the CAT model as a baseline. Therefore, lower is better for all figures. In Tables 1 to 4, the best values are denoted in bold, and the percentage of correct predictions is denoted between parentheses.

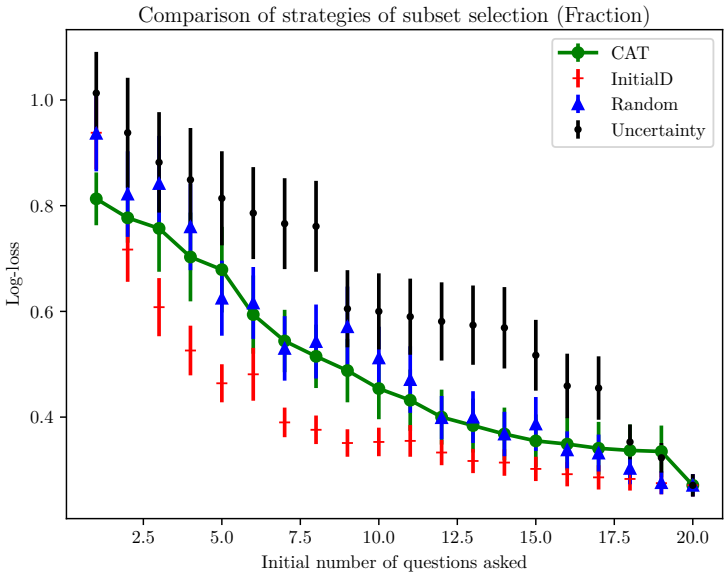

Fig.3. Log-loss of the predictions after a group of questions has been asked for different strategies for the dataset Fraction.

Fraction

In Figure 3, InitialD performs better than the other strategies, and with a narrow confidence interval. 8 questions seem enough to reconstruct correctly 82% of answers, and converge to a minimal log-loss.

In Figure 4, for $k \leq 9$ InitialD converges faster towards the final diagnosis, while for $k \geq 14$, CAT converges faster, showing a benefit in adaptation in later periods of the process.

TIMSS

In Figure 5, InitialD performs better than Random, CAT and Uncertainty. As early as the first question, InitialD clearly has a lower log-loss in response pattern reconstruction. This happens because the question of biggest "volume" has the vector of highest norm, which means, the most discriminant question, while other models will pick a question of average difficulty.

Asking 7 questions over 23 using the strategy InitialD leads in average to the same estimation error than asking 12 questions at random, or asking 19 questions using a traditional adaptive test.

In Figure 6, InitialD converges towards the final diagnosis faster than the other strategies. 12 questions using InitialD seem enough to get a diagnostic that predicts correctly 70% of the outcomes.

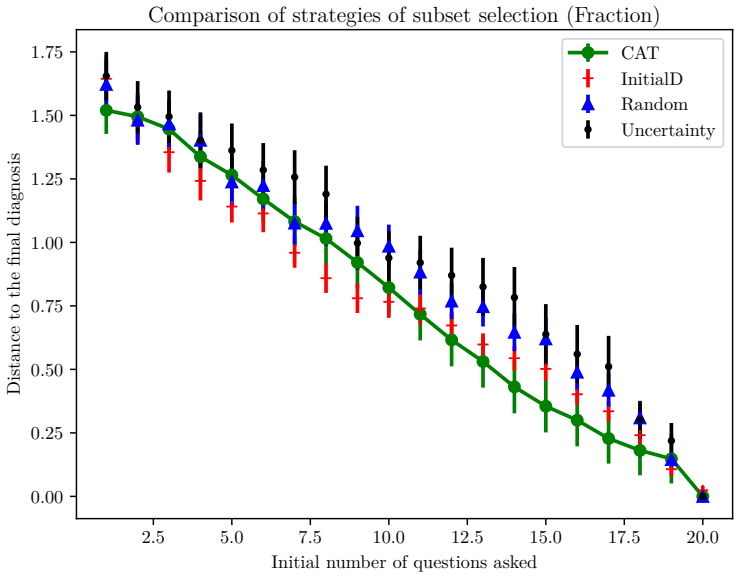

Fig.4. Distance to the final diagnosis after a group of questions has been asked for different strategies for the dataset Fraction.

DISCUSSION

On every dataset we tried, InitialD performed better, and with a narrower confidence interval, than the other strategies. On the TIMSS dataset, the Random strategy performs well compared to a normal adaptive test. It may be because the test is already well balanced, so random questions have high probability to be diverse.

If the number of questions to ask (k), the number of questions in the bank (n) and the number of measured knowledge components (K) are low, it is possible to simulate every subset of k questions over

Table 1
Log-loss values obtained (and precision rates) for the dataset Fraction.

	After 3 questions	After 8 questions	After 15 questions
CAT	0.757 ± 0.082 (67%)	0.515 ± 0.06 (82%)	$\mathbf{0.355 \pm 0.05}$ (88%)
Uncertainty	0.882 ± 0.095 (72%)	0.761 ± 0.086 (76%)	0.517 ± 0.067 (86%)
InitialD	$\mathbf{0.608 \pm 0.055}$ (74%)	$\mathbf{0.376 \pm 0.027}$ (82%)	$\mathbf{0.302 \pm 0.023}$ (86%)
Random	0.842 ± 0.09 (70%)	0.543 ± 0.07 (80%)	$\mathbf{0.387 \pm 0.051}$ (86%)

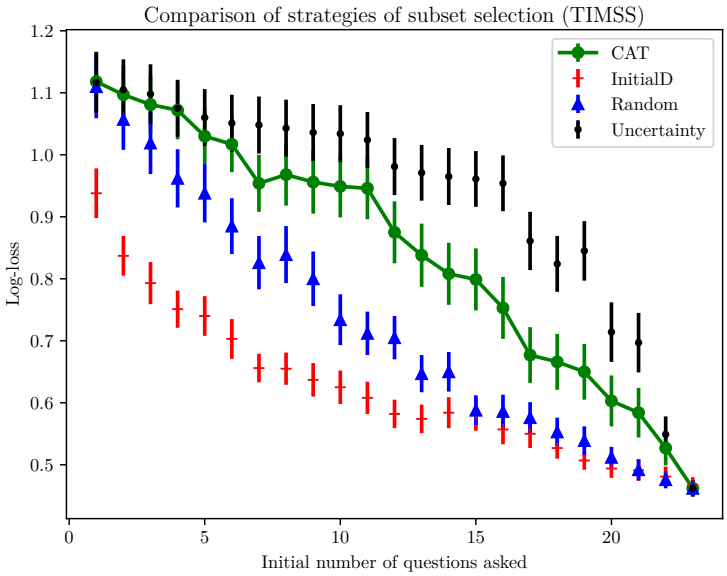

Fig.5. Log-loss of the predictions after a group of questions has been asked for different strategies for the dataset TIMSS.

n. However, in practice, question banks on platforms will be large, so InitialD's complexity, $O(nk^3)$ after a preprocessing step in $O(n^3)$, will be an advantage. In this paper, we tested our method on datasets of up to 23 questions, but the exact determinantal point process sampling algorithm has already been tried on databases of thousands of items (Kulesza and Taskar, 2012). Please note that this work naturally extends to the question cold-start problem: having a new question on the platform, how to identify a group of students to ask it to in order to estimate its discrimination parameters over all knowledge components.

InitialD could be improved by sampling several subsets of questions, and keeping the best of them. Sampling ℓ subsets of k questions has complexity $O(\ell nk^3)$, finding the one achieving the biggest volume

Table 2
Distance to the final diagnosis obtained for the dataset Fraction.

	After 3 questions	After 8 questions	After 15 questions
CAT	1.446 ± 0.094	1.015 ± 0.101	$\mathbf{0.355 \pm 0.103}$
Uncertainty	1.495 ± 0.103	1.19 ± 0.112	0.638 ± 0.119
InitialD	1.355 ± 0.08	$\mathbf{0.859 \pm 0.058}$	0.502 ± 0.047
Random	1.467 ± 0.095	1.075 ± 0.089	0.62 ± 0.083

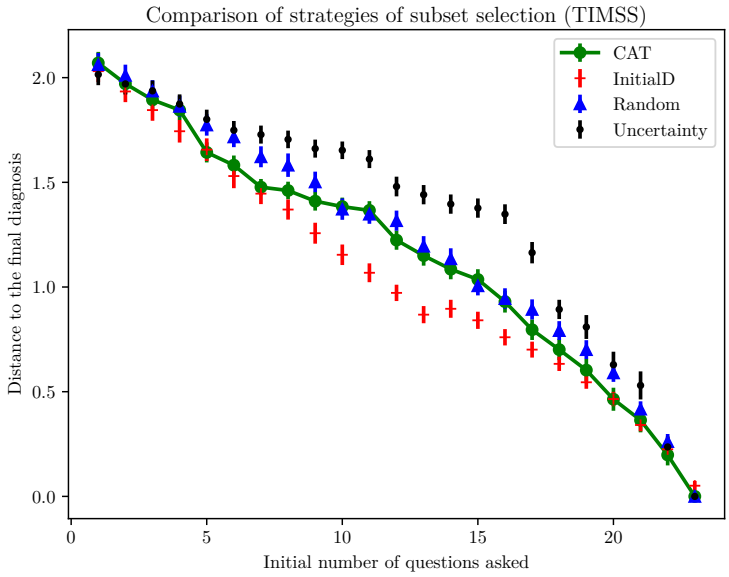

Fig.6. Distance to the final diagnosis after a group of questions has been asked for different strategies for the dataset TIMSS.

has complexity $O(\ell k^3)$. Drawing several subsets increases the chance to determine the best subset to ask first.

CONCLUSION AND FURTHER WORK

We showed, using real data, that our strategy InitialD, based on determinantal point processes, performed better than other strategies for cold-start at predicting student performance. As it is fast, this method can be applied to the generation of several diverse worksheets from the item bank of an educational platform:

Table 3
Log-loss values obtained (and precision rates) for the dataset TIMSS.

	After 3 questions	After 12 questions	After 20 questions
CAT	1.081 ± 0.047 (62%)	0.875 ± 0.05 (66%)	0.603 ± 0.041 (75%)
Uncertainty	1.098 ± 0.048 (58%)	0.981 ± 0.046 (68%)	0.714 ± 0.048 (72%)
InitialD	$\mathbf{0.793 \pm 0.034}$ (61%)	$\mathbf{0.582 \pm 0.023}$ (70%)	$\mathbf{0.494 \pm 0.015}$ (74%)
Random	1.019 ± 0.05 (58%)	0.705 ± 0.035 (68%)	$\mathbf{0.512 \pm 0.017}$ (74%)

a learner can request a worksheet of k questions, attempt to solve them, receive their ability features as feedback (strong and weak points), then ask for another worksheet. Items already administered to some student can be removed from the item bank, in their view, so that the same learner does not get the same exercise in two consecutive worksheets.

As further work, we would like to check if sampling according to a determinantal point processes is still useful in later stages of a multistage test, after a first learner ability estimate has been computed.

InitialD solely relies on pairwise similarities between questions: it can be used in conjunction with other response models, using other feature extraction techniques that allow better vector representations of the questions. For example, one could use various information at hand such as a bag-of-words representation of the problem statement, or extra tags specified by a teacher, in order to improve the embedding of items. Such extra information will improve the selection of questions, with the same algorithm InitialD. In this paper, we used a linear kernel for the predictions and for the student model, but nonlinear kernels could be used, performing better but at the cost of interpretation.

For interpretation of KCs, a q-matrix is useful. Koedinger, McLaughlin, and Stamper (2012) have shown that it is possible to combine q-matrices by crowdsourcing in order to improve student models. We would like to see if it also applies to the General Diagnostic Model, and if observing the discrimination parameters can help us determine better q-matrices.

Our strategy allows a fast and narrow ability estimate of the learner knowledge, that can be revealed to them in order to help them progress. If the lessons in the course are also mapped to KCs, recommendations of lessons could be made based on this initial evaluation, for example. We aim to apply InitialD to automatic generation of worksheets at the beginning of a MOOC, in order to provide low-stakes formative assessments, and evaluate them on real students.

ACKNOWLEDGEMENTS

This work is supported by the Paris-Saclay Institut de la Société Numérique funded by the IDEX Paris-Saclay, ANR-11-IDEX-0003-02.

Table 4
Distance to the final diagnosis obtained for the dataset TIMSS.

	After 3 questions	After 12 questions	After 20 questions
CAT	1.894 ± 0.05	1.224 ± 0.046	$\mathbf{0.464 \pm 0.055}$
Uncertainty	1.937 ± 0.049	1.48 ± 0.047	0.629 ± 0.062
InitialD	1.845 ± 0.051	$\mathbf{0.972 \pm 0.039}$	$\mathbf{0.465 \pm 0.034}$
Random	1.936 ± 0.052	1.317 ± 0.048	0.59 ± 0.043

References

Anava, Oren, Shahar Golan, Nadav Golbandi, Zohar Karnin, Ronny Lempel, Oleg Rokhlenko, and Oren Somekh (2015). "Budget-Constrained Item Cold-Start Handling in Collaborative Filtering Recommenders via Optimal Design". In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 45–54 (cit. on p. 5).

Cai, Li (2010a). "High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm". In: *Psychometrika* 75.1, pp. 33–57 (cit. on pp. 3, 10).

— (2010b). "Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis". In: *Journal of Educational and Behavioral Statistics* 35.3, pp. 307–335 (cit. on p. 10).

Chalmers, R. Philip (2012). "mirt: A multidimensional item response theory package for the R environment". In: *Journal of Statistical Software* 48.6, pp. 1–29 (cit. on pp. 5, 10, 11).

— (2016). "Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications". In: *Journal of Statistical Software* 71.1, pp. 1–38 (cit. on pp. 2, 3, 5, 11).

Chang, Hua-Hua (2014). "Psychometrics Behind Computerized Adaptive Testing". In: *Psychometrika*, pp. 1–20 (cit. on p. 4).

Cheng, Ying (2009). "When cognitive diagnosis meets computerized adaptive testing: CD-CAT". In: *Psychometrika* 74.4, pp. 619–632 (cit. on p. 4).

Colvin, Kimberly F, John Champaign, Alwina Liu, Qian Zhou, Colin Fredericks, and David E Pritchard (2014). "Learning in an introductory physics MOOC: All cohorts learn equally, including an on-campus class". In: *The International Review of Research in Open and Distributed Learning* 15.4 (cit. on p. 3).

Davier, Matthias (2005). "A general diagnostic model applied to language testing data". In: *ETS Research Report Series* 2005.2, pp. i–35 (cit. on p. 4).

De La Torre, Jimmy (2009). "DINA model and parameter estimation: A didactic". In: *Journal of Educational and Behavioral Statistics* 34.1, pp. 115–130 (cit. on p. 3).

DeCarlo, Lawrence T. (2010). "On the analysis of fraction subtraction data: The DINA model, classification, latent class sizes, and the Q-matrix". In: *Applied Psychological Measurement* (cit. on p. 9).

Desmarais, Michel C. and Ryan S. J. D. Baker (2012). "A review of recent advances in learner and skill modeling in intelligent learning environments". In: *User Modeling and User-Adapted Interaction* 22.1-2, pp. 9–38 (cit. on pp. 2, 3).

Dunlosky, John, Katherine A. Rawson, Elizabeth J. Marsh, Mitchell J. Nathan, and Daniel T. Willingham (2013). "Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology". In: *Psychological Science in the Public Interest* 14.1, pp. 4–58 (cit. on p. 3).

Hambleton, Ronald K. and Hariharan Swaminathan (1985). *Item response theory: Principles and applications*. Vol. 7. Springer Science & Business Media (cit. on p. 2).

Koedinger, Kenneth R., Elizabeth A. McLaughlin, and John C. Stamper (2012). "Automated Student Model Improvement." In: *International Educational Data Mining Society* (cit. on p. 16).

Kulesza, Alex and Ben Taskar (2012). "Determinantal point processes for machine learning". In: *arXiv preprint arXiv:1207.6083* (cit. on pp. 2, 5, 7, 8, 14).

Lan, Andrew S., Andrew E. Waters, Christoph Studer, and Richard G. Baraniuk (2014). "Sparse factor analysis for learning and content analytics". In: *The Journal of Machine Learning Research* 15.1, pp. 1959–2008 (cit. on pp. 4, 5, 11).

Leighton, Jacqueline P., Mark J. Gierl, and Stephen M. Hunka (2004). "The Attribute Hierarchy Method for Cognitive Assessment: A Variation on Tatsuoka's Rule-Space Approach". In: *Journal of Educational Measurement* 41.3, pp. 205–237 (cit. on p. 3).

Lynch, Danny and Colm P. Howlin (2014). *Real world usage of an adaptive testing algorithm to uncover latent knowledge* (cit. on pp. 2, 3).

Magis, David (2015). "Empirical comparison of scoring rules at early stages of CAT". In: (cit. on p. 11).

Piech, Chris, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein (2015). "Deep knowledge tracing". In: *Advances in Neural Information Processing Systems*, pp. 505–513 (cit. on p. 3).

Reckase, Mark (2009). *Multidimensional item response theory*. Vol. 150. Springer (cit. on p. 2).

Robitzsch, A., T. Kiefer, A. C. George, and A. Ünlü (2014). "CDM: Cognitive diagnosis modeling". In: *R Package version* 3 (cit. on p. 9).

Rosen, Yigal, Ilia Rushkin, Andrew Ang, Colin Federicks, Dustin Tingley, and Mary Jean Blink (2017). "Designing Adaptive Assessments in MOOCs". In: *Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale*. L@S '17. Cambridge, Massachusetts, USA: ACM, pp. 233–236. ISBN: 978-1-4503-4450-0. DOI: 10.1145/3051457.3053993. URL: http://doi.acm.org/10.1145/3051457.3053993 (cit. on pp. 2, 4).

Settles, Burr (2010). "Active learning literature survey". In: *University of Wisconsin, Madison* 52.55-66, p. 11 (cit. on p. 9).

Su, Yu-Law, K. M. Choi, W. C. Lee, T. Choi, and M. McAninch (2013). "Hierarchical cognitive diagnostic analysis for TIMSS] 2003 mathematics". In: *Centre for Advanced Studies in Measurement and Assessment* 35, pp. 1–71 (cit. on p. 9).

Thai-Nghe, Nguyen, Lucas Drumond, Tomáš Horváth, Lars Schmidt-Thieme, et al. (2011). "Multi-relational factorization models for predicting student performance". In: *Proc. of the KDD Workshop on Knowledge Discovery in Educational Data*. Citeseer (cit. on pp. 2, 5).

Vie, Jill-Jênn, Fabrice Popineau, Yolaine Bourda, and Éric Bruillard (2016a). "A review of recent advances in adaptive assessment". In: *Learning analytics: Fundaments, applications, and trends: A view of the current state of the art*. Springer, in press (cit. on p. 5).

— (2016b). "Adaptive Testing Using a General Diagnostic Model". In: *European Conference on Technology Enhanced Learning*. Springer, pp. 331–339 (cit. on pp. 4, 6, 11).

Wilson, Kevin H, Xiaolu Xiong, Mohammad Khajah, Robert V Lindsey, Siyuan Zhao, Yan Karklin, Eric G Van Inwegen, Bojian Han, Chaitanya Ekanadham, Joseph E Beck, et al. (2016). "Estimating student proficiency: Deep learning is not the panacea". In: *In Neural Information Processing Systems, Workshop on Machine Learning for Education* (cit. on p. 3).

Yan, Duanli, Alina A. von Davier, and Charles Lewis (2014). *Computerized Multistage Testing*. CRC Press (cit. on p. 4).

Zheng, Yi and HH Chang (2014). “Multistage testing, on-the-fly multistage testing, and beyond”. In: *Advancing methodologies to support both summative and formative assessments*, pp. 21–39 (cit. on pp. 4, 9).